Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques

Jean-François Girres

▶ **To cite this version:**

Jean-François Girres. Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Autre [cs.OH]. Université Paris-Est, 2012. Français. NNT : 2012PEST1080 . tel-00809273

HAL Id: tel-00809273
https://theses.hal.science/tel-00809273

Submitted on 8 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-EST

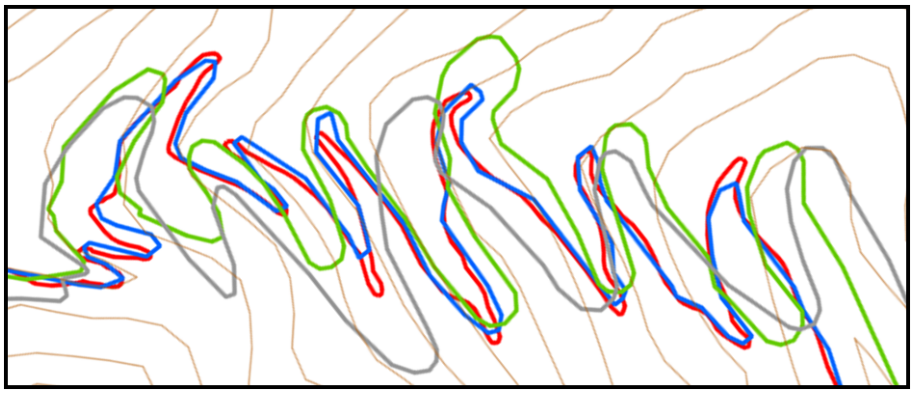
Thèse

pour obtenir le grade de docteur de l'Université Paris-Est
Spécialité : Sciences et Technologies de l'Information Géographique

Jean-François Girres

Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques

Application aux mesures de longueur et de surface



Soutenue publiquement le 4 Décembre 2012

Composition du jury :

Pr. Nicholas Chrisman, Université Laval	Examinateur
Dr. Eric Desjardin, Université de Reims	Examinateur
Pr. Thomas Devogèle, Université de Tours	Rapporteur
Dr. Anne Ruas, IFSTTAR	Directrice de thèse
Pr. Sophie de Ruffray, Université de Rouen	Rapporteur
Dr. Christiane Weber, CNRS	Examinatrice

Université Paris-Est – École doctorale MSTIC

Cette thèse a été réalisée au sein du Laboratoire COGIT de l'Institut National de l'Information Géographique et Forestière (IGN), sous la direction d'Anne Ruas.

Institut National de l'Information Géographique et Forestière
Service de la Recherche, Laboratoire COGIT
73, Avenue de Paris
94165 Saint-Mandé Cedex
Tél : 01 43 98 80 00

Résumé

De nombreuses applications SIG reposent sur des mesures de longueur ou de surface calculées à partir de la géométrie des objets d'une base de données géographiques (comme des calculs d'itinéraires routiers ou des cartes de densité de population par exemple). Cependant, aucune information relative à l'imprécision de ces mesures n'est aujourd'hui communiquée à l'utilisateur. En effet, la majorité des indicateurs de précision géométrique proposés porte sur les erreurs de positionnement des objets, mais pas sur les erreurs de mesure, pourtant très fréquentes. Dans ce contexte, ce travail de thèse cherche à mettre au point des méthodes d'estimation de l'imprécision des mesures géométriques de longueur et de surface, afin de renseigner un utilisateur dans une logique d'aide à la décision.
Pour répondre à cet objectif, nous proposons un modèle permettant d'estimer les impacts de règles de représentation (projection cartographique, non-prise en compte du terrain, approximation polygonale des courbes) et de processus de production (erreur de pointé et généralisation cartographique) sur les mesures géométriques de longueur et de surface, en fonction des caractéristiques des données vectorielles évaluées et du terrain que ces données décrivent. Des méthodes d'acquisition des connaissances sur les données évaluées sont également proposées afin de faciliter le paramétrage du modèle par l'utilisateur. La combinaison des impacts pour produire une estimation globale de l'imprécision de mesure demeure un problème complexe et nous proposons quelques premières approches pour approximer au mieux cette erreur cumulée. Le modèle proposé est implémenté au sein du prototype EstIM (Estimation de l'Imprécision des Mesures).

Mots clefs : SIG, qualité des données géographiques, mesures géométriques, imprécision, modèle d'estimation, aide à la décision.

Abstract

Many GIS applications are based on length and area measurements computed from the geometry of the objects of a geographic database (such as route planning or maps of population density, for example). However, no information concerning the imprecision of these measurements is now communicated to the final user. Indeed, most of the indicators on geometric quality focus on positioning errors, but not on measurement errors, which are very frequent. In this context, this thesis seeks to develop methods for estimating the imprecision of geometric measurements of length and area, in order to inform a user for decision support.
To achieve this objective, we propose a model to estimate the impacts of representation rules (cartographic projection, terrain, polygonal approximation of curves) and production processes (digitizing error, cartographic generalisation) on geometric measurements of length and area, according to the characteristics and the spatial context of the evaluated objects. Methods for acquiring knowledge about the evaluated data are also proposed to facilitate the parameterization of the model by the user. The combination of impacts to produce a global estimation of the imprecision of measurement is a complex problem, and we propose approaches to approximate the cumulated error bounds. The proposed model is implemented in the EstIM prototype (Estimation of the Imprecision of Measurements).

Key words : GIS, spatial data quality, geometric measurements, imprecision, estimation model, decision support.

Remerciements

Je tiens tout d'abord à remercier les membres du jury pour avoir accepté d'évaluer ce travail de thèse.

Je remercie tout particulièrement Anne Ruas pour m'avoir accueilli au sein du laboratoire COGIT, et pour avoir accepté de diriger ce travail de thèse. Merci pour tes conseils avisés, ta disponibilité et ton soutien.

Je tiens à remercier l'ensemble des membres du laboratoire COGIT, avec qui travailler au quotidien fut un véritable plaisir tout au long de cette aventure doctorale. Je les remercie pour leur sympathie, leurs encouragements, et pour tous les échanges partagés qui ont contribué à enrichir ce travail si personnel.

J'associe à ces remerciements les nombreuses personnes de l'IGN qui m'ont ouvert leur porte pour répondre à mes questions, ainsi que le personnel de l'ENSG qui s'est mobilisé pour m'aider à organiser la soutenance.

Enfin, je remercie Estelle, ma famille et mes amis pour leur soutien permanent.

Table des matières

Introduction **13**

1 Etat de l'art sur l'imprécision géométrique des données géographiques **17**
Introduction . . . 17
1.1 Les causes d'imperfections des données géographiques . . . 18
1.1.1 L'information géographique, une simplification du monde réel . . . 18
1.1.1.1 L'information géographique vectorielle . . . 18
1.1.1.2 Abstraction et modélisation du monde réel . . . 19
1.1.1.3 Contenu et niveau de détail des bases de données . . . 22
1.1.1.4 Impact de la représentation à différentes échelles . . . 23
1.1.2 Les erreurs dans les bases de données géographiques . . . 26
1.1.2.1 La notion d'erreur de mesure . . . 26
1.1.2.2 Les sources d'erreurs dans les données géographiques . . . 28
1.1.2.3 Erreurs d'acquisition et processus de production . . . 30
1.1.2.4 Impact de l'indépendance des bases de données . . . 38
1.2 Imperfection, prise de décision et qualité des données géographiques . . . 41
1.2.1 Imperfection des données géographiques et prise de décision . . . 41
1.2.1.1 Un manque de consensus terminologique . . . 41
1.2.1.2 Les notions d'imperfection . . . 43
1.2.1.3 Incertitude et prise de décision . . . 44
1.2.2 La qualité des données géographiques . . . 46
1.2.2.1 La notion de qualité . . . 46
1.2.2.2 Eléments d'évaluation de la qualité . . . 47
1.2.2.3 Normalisation de l'évaluation de la qualité . . . 48
1.2.2.4 Communication de l'information qualité . . . 50
1.3 Evaluation de l'imprécision géométrique des données géographiques . . . 53
1.3.1 Mesures d'évaluation de l'imprécision géométrique . . . 53
1.3.1.1 Mesures sur les primitives géométriques . . . 53
1.3.1.2 Mesures proposées dans le cadre règlementaire . . . 58
1.3.2 Les modèles d'évaluation de l'imprécision géométrique . . . 61
1.3.2.1 Modèles relatifs au positionnement . . . 61
1.3.2.2 Modèles relatifs aux mesures géométriques . . . 68
Conclusion . . . 71

2 Proposition : un modèle d'estimation de l'imprécision des mesures géométriques **73**
Introduction . . . 73
2.1 Besoins et moyens actuels d'évaluation de l'imprécision géométrique . . . 74

2.1.1 Métadonnées, qualité géométrique et besoins des utilisateurs 74
2.1.1.1 Métadonnées et qualité externe 74
2.1.1.2 Imprécision géométrique et besoins utilisateurs 76
2.1.2 Limites des modèles d'évaluation existants 79
2.1.2.1 Limites relatives à l'impact du processus de production . . 79
2.1.2.2 Limites impliquées par le processus d'abstraction 81
2.1.3 Positionnement de la thèse 84
2.2 Spécifications générales du modèle d'estimation 86
2.2.1 Objectifs du modèle d'estimation 86
2.2.1.1 Bases de données à évaluer 86
2.2.1.2 Identification des impacts à évaluer 89
2.2.1.3 Communication des résultats 90
2.2.2 Fonctionnement du modèle 93
2.2.2.1 Principe d'évaluation 93
2.2.2.2 Interface homme-machine 95
2.2.3 Modélisation par approche orientée-objet 97
Conclusion 99

3 Méthodes d'estimation de l'impact des règles de représentation 101
Introduction 101
3.1 Projection cartographique 102
3.1.1 La représentation de la Terre et ses impacts 102
3.1.1.1 De la Terre à la carte : principes généraux 102
3.1.1.2 Impact sur les mesures géométriques 108
3.1.2 Méthodes pour le calcul de mesures géométriques sur ellipsoïde . . . 114
3.1.2.1 Création des grilles d'altération 114
3.1.2.2 Mesures sur ellipsoïde par enrichissement de la géométrie . 116
3.1.3 Application du modèle d'estimation 119
3.2 Prise en compte du terrain 121
3.2.1 Modélisation de l'information altimétrique et mesures géométriques . 121
3.2.1.1 Des mesures géométriques souvent limitées en 2D 121
3.2.1.2 Proposition pour des mesures géométriques en 2D5 123
3.2.2 Méthodes d'enrichissement géométrique pour des mesures en 2D5 . . 125
3.2.2.1 Primitives linéaires 126
3.2.2.2 Primitives surfaciques 127
3.2.2.3 Limites des méthodes proposées 130
3.2.3 Application du modèle d'estimation 132
3.2.3.1 Impact de la résolution spatiale 133
3.2.3.2 Comparaison de l'impact entre MNT et MNE 135
3.3 Approximation polygonale des courbes 137
3.3.1 L'approximation polygonale et son impact 137
3.3.1.1 L'approximation polygonale des courbes : principe général . 137
3.3.1.2 Degré d'approximation polygonale 139
3.3.1.3 Solutions envisagées pour estimer son impact 140
3.3.2 Proposition de méthodes de modélisation réaliste des courbes 142
3.3.2.1 Méthodes d'interpolation polynomiale 142
3.3.2.2 Préservation du réalisme des courbes 145
3.3.2.3 Détection des objets sinueux 148

3.3.3 Application de la méthode de reconstruction . . . 154
Conclusion . . . 159

4 Méthodes d'estimation de l'impact des processus de production 161
Introduction . . . 161
4.1 Erreur de pointé . . . 162
4.1.1 Le bruitage de la géométrie des objets vectoriels . . . 162
4.1.1.1 Principe général . . . 162
4.1.1.2 Problèmes impliqués par le bruitage . . . 165
4.1.2 Paramétrage de l'amplitude du bruit . . . 168
4.1.2.1 Estimation de l'imprécision de pointé . . . 168
4.1.2.2 Estimation de l'échelle de saisie . . . 169
4.1.3 Contraintes appliquées pour préserver le réalisme du bruitage . . . 174
4.1.3.1 Respect de la forme initiale des objets . . . 174
4.1.3.2 Corrélation des erreurs entre points successifs . . . 177
4.1.3.3 Combinaison des contraintes pour la simulation d'erreurs . 180
4.1.4 Expérimentation pour la validation du modèle de simulation . . . 182
4.1.4.1 Présentation de l'expérimentation . . . 182
4.1.4.2 Résultats de l'expérimentation . . . 183
4.1.4.3 Calibration du modèle de simulation . . . 188
4.1.5 Application du modèle d'estimation . . . 191
4.1.5.1 Application sur des géométries linéaires . . . 191
4.1.5.2 Application sur des géométries surfaciques . . . 193
4.2 Généralisation cartographique . . . 199
4.2.1 Les impacts de la généralisation cartographique . . . 199
4.2.1.1 Opérations de généralisation . . . 199
4.2.1.2 Impact sur les mesures géométriques . . . 201
4.2.1.3 Proposition de méthode d'estimation . . . 204
4.2.2 Acquisition de connaissances pour le paramétrage du modèle . . . 207
4.2.2.1 Estimation de l'échelle de représentation des données sources 207
4.2.2.2 Délimitation du contexte géographique des objets . . . 219
4.2.3 Impact de la généralisation sur les mesures de longueur . . . 228
4.2.3.1 Sensibilité de l'erreur relative à la longueur . . . 228
4.2.3.2 Elaboration du modèle statistique . . . 235
4.2.3.3 Validation du modèle . . . 240
4.2.4 Impact de la généralisation sur les mesures de surface . . . 247
4.2.4.1 Méthodologie de comparaison . . . 248
4.2.4.2 Elaboration du modèle statistique . . . 250
Conclusion . . . 262

5 Conception du modèle d'estimation de l'imprécision des mesures géométriques 263
Introduction . . . 263
5.1 Combinaison des impacts sur les mesures géométriques . . . 264
5.1.1 Méthodes de combinaison des impacts . . . 264
5.1.1.1 La combinaison des impacts : une démarche complexe . . . 264
5.1.1.2 Proposition de méthodes de combinaison . . . 266
5.1.2 Application des méthodes de combinaison . . . 269

5.1.2.1 Présentation de l'expérimentation . . . 269
5.1.2.2 Estimation et combinaison des impacts . . . 273
5.1.2.3 Comparaison des méthodes de combinaison des impacts . . 280
5.1.3 Perspectives pour une estimation globale . . . 284
5.1.3.1 Apprentissage automatique supervisé . . . 284
5.1.3.2 Composition de modèles . . . 287
5.2 Principe de modélisation . . . 292
5.2.1 Modélisation des géométries . . . 292
5.2.2 Description des composantes du modèle d'estimation . . . 295
5.2.2.1 Règles de représentation . . . 295
5.2.2.2 Processus de production . . . 298
5.2.2.3 Combinaison des impacts . . . 302
5.3 Développement du prototype EstIM . . . 305
5.3.1 Fonctionnement général du modèle . . . 305
5.3.1.1 Sélection des objets à évaluer . . . 305
5.3.1.2 Sélection et paramétrage des impacts . . . 306
5.3.1.3 Combinaison des impacts . . . 309
5.3.2 Interface homme-machine . . . 310
5.3.2.1 Le logiciel SIG OpenJUMP . . . 310
5.3.2.2 Les interfaces du modèle EstIM . . . 311
5.3.3 Mise en œuvre du modèle EstIM . . . 317
5.3.3.1 Présentation de l'expérimentation . . . 317
5.3.3.2 Estimation de l'imprécision des mesures . . . 318
Conclusion . . . 320

Conclusion Générale **321**

Annexes **327**

Bibliographie **357**

Table des figures **367**

Liste des tableaux **370**

Publications **371**

"Tous les modèles sont faux, certains sont utiles."

George Edward Pelham Box, 1987

Introduction

Contexte général

Les dernières décennies ont été le théâtre de changements majeurs dans le domaine des nouvelles technologies de l'information et de la communication. Ces bouleversements portent tant sur le traitement et la diffusion de l'information, que sur ses usages. Le développement des moyens de communication offre aujourd'hui la possibilité de diffuser des masses colossales d'informations, que les avancées en termes de capacités de traitement rendent toujours plus adaptées aux attentes des utilisateurs. Les utilisateurs quant à eux, élargissent au quotidien leur panel d'usages, par le biais d'applications toujours plus adaptées à leurs besoins.

Dans ce contexte, le traitement, la diffusion et les usages de l'information géographique ont également fortement évolué. En effet, la dernière décennie a vu l'émergence de nombreux outils de consultation et de mise à disposition (gratuite ou non) de données géographiques sur le Web. Les outils simples de visualisation se sont peu à peu enrichis de fonctionnalités avancées (calculs d'itinéraires, requêtes complexes...), participant à une démocratisation des usages de l'information géographique numérique. Dans le même temps, le développement de nombreux logiciels SIG libres, associé à la mise à disposition d'outils d'acquisition à moindre coût (comme les récepteurs GPS) ont profondément transformé le paradigme classique producteur-utilisateur, à tel point que l'utilisateur est aujourd'hui devenu producteur d'information, comme l'atteste le succès des projets collaboratifs (Goodchild, 2007).

Sous cette prolifération de données de sources hétérogènes, produites et pensées pour des usages spécifiques, il est fort probable que les utilisateurs (en particulier les utilisateurs non-experts) prennent parfois le risque d'utiliser ces données de manière non-adéquate (Beard, 1989). En effet, l'information géographique, sous sa représentation numérique, peut laisser supposer une certaine impression d'exactitude, dont l'utilisateur n'a pas forcément le recul nécessaire pour en évaluer la teneur (Chrisman, 1991; Devillers, 2004). Pourtant, l'information géographique, en particulier sous sa représentation vectorielle, est par définition entachée d'imperfections, d'origines diverses. Dans ce contexte, l'évaluation et la communication de la qualité des données géographiques constituent un enjeu majeur pour l'utilisateur.

Problématique

La communication d'une information sur la qualité des données géographiques a pour principal objectif de réduire le risque pris par un utilisateur dans le cadre d'un processus décisionnel sollicitant l'information géographique. Des moyens d'évaluation et de commu-

nication de la qualité ont donc été développés dans ce sens ces dernières années. Pourtant, si les éléments d'évaluation de la qualité proposés aujourd'hui offrent des informations intéressantes à l'utilisateur, ils ne permettent pas de communiquer une information utile à tous les usages envisagés. C'est particulièrement le cas concernant la géométrie des objets, où la grande majorité des indicateurs proposés se focalise sur une information relative à l'exactitude de positionnement. Nous savons pourtant que de nombreux usages mobilisent d'autres informations extraites de la géométrie des objets vectoriels, en particulier les mesures géométriques de longueur ou de surface (pour des calculs d'itinéraires, ou la recherche d'un terrain à bâtir par exemple), et que celles-ci sont également entachées d'erreurs.

L'un des exemples les plus significatifs pour illustrer les erreurs de mesures géométriques fait référence au Plan Cadastral Informatisé au format vectoriel (PCI Vecteur). Dans ce plan cadastral, la surface géométrique diffère de la surface fiscale de référence (relevée au planimètre ou par un géomètre-expert) qui est stockée sous forme d'attribut de la base.
A travers cet exemple, nous voyons bien que le calcul de mesures géométriques est entaché d'erreurs, même sur des objets saisis à grande échelle, comme les planches cadastrales. La question de l'imprécision des mesures géométriques doit par conséquent se poser sur des objets saisis à moyenne échelle. Malgré cela, aucune information sur l'imprécision de ces mesures n'est aujourd'hui proposée à l'utilisateur.
Pour l'estimer, une solution réside dans la comparaison des mesures calculées sur les objets d'une base de données avec leurs homologues extraites d'une base de données de référence. Cependant, l'utilisateur ne dispose généralement pas de données de référence (qu'il utiliserait si c'était le cas), rendant impossible une telle estimation.

Dans ce contexte, ce travail de thèse cherche donc à savoir comment estimer l'imprécision de mesures de longueur et de surface calculées à partir de la géométrie des objets d'une base de données géographiques, sans données de référence, c'est-à-dire dans la position d'un utilisateur. Répondre à une telle question suscite de nombreuses interrogations sous-jacentes, relatives à l'identification des sources d'erreurs de mesure, ainsi qu'aux méthodes permettant de les estimer.

Objectifs

L'objectif de ce travail de thèse porte sur l'élaboration d'un modèle d'estimation de l'imprécision des mesures géométriques de longueur et de surface. Le modèle ainsi proposé doit permettre à un utilisateur d'estimer l'imprécision de la mesure géométrique calculée, en formulant une réponse de la manière suivante : "La mesure calculée est x mètres, cependant nous estimons que celle-ci est probablement comprise entre (x+a) mètres et (x+b) mètres".

Pour permettre une telle évaluation, une réflexion générale doit être menée sur l'identification des composantes de l'erreur de mesure géométrique. Nous savons en effet que différents processus - d'abstraction du monde réel et de production des données - sont impliqués, et dont les impacts s'accumulent pour générer l'erreur de mesure. Nous proposerons donc tout d'abord de définir un cadre conceptuel des composantes de l'erreur de mesure.

Pour estimer les impacts respectifs de chacune des composantes de l'erreur de mesure, des méthodes d'estimation spécifiques doivent être développées, et validées. En l'absence de

données de référence, les méthodes proposées doivent également s'appuyer sur un processus d'acquisition des connaissances, permettant d'enrichir et de qualifier les données évaluées.

Enfin, l'estimation d'une imprécision de mesure à destination de l'utilisateur suppose une réflexion sur la combinaison des impacts des différentes composantes de l'erreur finale. Des méthodes de combinaison des impacts devront donc être envisagées et expérimentées afin de produire une estimation globale la plus réaliste possible.

Etant destiné à des utilisateurs non-experts, le modèle ainsi proposé devra se matérialiser sous forme d'un outil intuitif, implémenté au sein d'une interface SIG classique.

Plan

Le chapitre 1 propose un état de l'art sur l'évaluation de l'imprécision géométrique des données géographiques vectorielles. Nous montrerons que les données géographiques sont par définition entachées d'imperfections d'origines diverses, et que celles-ci peuvent avoir un impact sur les prises de décision d'un utilisateur. Nous reviendrons sur le concept de qualité des données géographiques, et en particulier sur les moyens permettant d'estimer la qualité géométrique. Nous verrons que les éléments proposés aujourd'hui portent davantage sur l'erreur de positionnement que sur l'erreur de mesure géométrique.

Le chapitre 2 permettra de définir les objectifs de ce travail de thèse. Après avoir établi un bilan des moyens et des besoins actuels, nous montrerons qu'estimer l'imprécision des mesures géométriques est une tâche complexe. Nous verrons qu'une telle estimation suppose de prendre en compte des processus hétérogènes (règles de représentation et processus de production), dont les moyens d'évaluation proposés actuellement ne permettent pas de modéliser les impacts respectifs.

Nous présenterons au chapitre 3 les méthodes mises en œuvre afin d'estimer les impacts des règles de représentation des données géographiques sur les mesures géométriques. Nous y étudierons trois impacts : la projection cartographique, la non-prise en compte du terrain, et l'approximation polygonale des courbes.

Dans le chapitre 4, nous proposerons des méthodes d'estimation des impacts des processus de production des données sur les mesures géométriques. Les impacts de l'erreur de pointé de l'opérateur, et de la généralisation cartographique y seront étudiés. Nous proposerons également des méthodes d'acquisition des connaissances sur les données évaluées, permettant de faciliter leur évaluation.

Le chapitre 5 traitera enfin de la conception du modèle d'estimation de l'imprécision des mesures géométriques. Nous proposerons d'expérimenter différentes méthodes de combinaison des impacts, afin de produire une estimation globale de l'imprécision de mesure à l'utilisateur final. Nous y présenterons les principes de modélisation, ainsi que le développement et le fonctionnement du prototype EstIM (Estimation de l'Imprécision des Mesures).

Chapitre 1

Etat de l'art sur l'imprécision géométrique des données géographiques

Introduction

Ce premier chapitre a pour objectif de présenter la notion d'imprécision géométrique des données géographiques vectorielles et les grands principes régissant son évaluation.

Pour mieux comprendre les enjeux liés à cette évaluation, une présentation de l'information géographique numérique, plus spécifiquement sous sa représentation en mode vectoriel, est tout d'abord proposée. Nous verrons que les bases de données géographiques vectorielles, en tant qu'abstraction du monde réel, sont inévitablement entachées d'imperfections. Ces imperfections se manifestent notamment en termes d'imprécision géométrique, qui est principalement conditionnée par les processus de production mobilisés pour saisir la géométrie des objets.

Etant donné qu'aujourd'hui de nombreux systèmes d'aide à la décision mobilisent l'usage de l'information géographique, il est nécessaire d'évaluer la qualité des données géographiques, afin de diminuer l'incertitude liée au processus de prise de décision. Dans ce contexte, nous introduirons le concept de qualité des données géographiques, et présenterons les éléments permettant son évaluation et sa communication.

Nous terminerons ce chapitre en présentant un état de l'art des mesures et modèles d'évaluation de l'imprécision géométrique dévéloppés ces dernières années, tant en recherche que dans un cadre règlementaire.

1.1 Les causes d'imperfections des données géographiques

L'information géographique numérique est aujourd'hui omniprésente dans notre société. Afin d'aborder les questions relatives à l'évaluation de sa qualité, nous allons rappeler dans cette section quelques notions fondamentales sur la représentation du monde réel sous forme de bases de données géographiques. Nous verrons que celle-ci est inévitablement entachée d'imperfections, conditionnées notamment par le niveau de détail recherché.

1.1.1 L'information géographique, une simplification du monde réel

Cette section présente comment sont représentés des phénomènes du monde réel sous forme d'information géographique numérique, et plus particulièrement sous forme de bases de données géographiques vectorielles. Nous verrons que ces données forment une représentation abstraite et simplifiée de la réalité, qui dépend du niveau de détail recherché.

1.1.1.1 L'information géographique vectorielle

Depuis les années 1990 et l'avènement d'Internet et des Technologies de l'Information et de la Communication (TIC), notre société est souvent qualifiée de "société de l'information". A ce titre, il est courant d'entendre dire que 80% de l'information que nous traitons possède une dimension géographique (Franklin, 1992).

L'information géographique L'information géographique est définie comme *la représentation d'un objet ou un phénomène réel, localisé dans l'espace, à un moment donné* (Quodverte, 1994). Deux types d'information géographique peuvent être distingués :
- les **informations de référence**, comme par exemple les bâtiments, les routes ou les voies ferrées, généralement qualifiées de données topographiques. C'est sur ce type d'informations que porte ce travail de thèse.
- les **informations thématiques**, liées à des domaines d'études particuliers, comme par exemple l'évolution spatiale d'une épidémie ou la variation du prix du foncier dans une ville, qui viennent enrichir l'information de référence.

Les données vectorielles Dans le cadre de ce travail de thèse, nous ne travaillerons que sur les données géographiques représentées en mode vecteur, par comparaison avec les données représentées en mode grille. Ainsi, le terme "base de données géographiques" ne concerne ici que les données géographiques vectorielles.

Dans une base de données géographiques vectorielles, le monde réel est modélisé sous forme de "classes d'objets" de même nature. On peut donc par exemple rencontrer des classes d'objets "bâtiment", "route"... Chaque objet d'une même classe est composé :
- d'une **géométrie**, représentée par une primitive graphique qui consiste en une suite structurée de coordonnées de points. Ces primitives graphiques peuvent être de type point, ligne ou polygone.
- d'**attributs**, qui décrivent la nature de l'objet et ses caractéristiques par des informations sémantiques. Par exemple, un objet de la classe "route" pourra prendre comme attribut "nature" les valeurs : "chemin" ou "sentier".

Dans le cadre de ce travail de thèse, nous nous focaliserons uniquement sur la dimension géométrique des objets vectoriels. Cependant, pour le bien de certaines expérimentations, les données attributaires devront parfois être utilisées.

Calculs de longueur et de surface La représentation du monde réel sous forme de base de données vectorielles permet de réaliser aisément des mesures de longueur ou de surface à partir de la géométrie des objets. Dans le cadre de cette thèse, nous les appelerons **mesures géométriques** de longueur ou de surface.
Ainsi, dans un espace en deux dimensions, la longueur L d'une polyligne est calculée par l'abscisse curviligne (cf. équation 1.1) :

$$L = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{1.1}$$

Tandis que l'aire A d'un polygone est calculée à partir de l'équation 1.2 :

$$A = 0.5 \sum_{i=1}^{n} x_i (y_{i+1} - y_{i-1}) \tag{1.2}$$

De ces mesures de longueur et de surface calculées à partir de la géométrie des objets vectoriels sont dérivées de nombreuses applications, alimentant bien souvent des sytèmes d'aide à la décision. Nous verrons par la suite que les imperfections géométriques des bases de données géographiques vectorielles peuvent amener à réaliser des mesures de longueur et de surface erronées, pouvant impacter des processus de prise de décision.

1.1.1.2 Abstraction et modélisation du monde réel

La création d'une base de données géographiques suppose une modélisation préalable, qui n'est pas une tâche triviale. Cette opération suppose en effet de réduire la complexité du monde réel, appelé également **univers** (David et Fasquel, 1997), de manière à en extraire les phénomènes à représenter sous une forme simplifiée et commode. Cette phase, communément appelée **abstraction** du monde réel, vise à réduire une réalité parfois très complexe (comme par exemple les crues d'une rivière) sous forme d'un modèle qui ne conservera que les éléments pertinents pour son analyse, sans être trop compliqué à saisir.

Spécifications L'ensemble des règles d'abstraction du monde réel dans les différentes classes d'objets sont décrites dans un document formel, appelé **spécifications**. L'Organisation de Standardisation Internationale (ISO) définit de manière générale les spécifications comme *le document qui prescrit les exigences auxquelles le produit ou le service doit se conformer* (ISO8402, 1994). Ainsi, les spécifications sont conçues de manière à ce que le produit ou service satisfasse les besoins des utilisateurs.

Pour les bases de données géographiques, on distingue généralement deux types de spécifications : les spécifications de contenu et les spécifications de saisie. Les **spécifications de contenu** décrivent quelle information doit être présente dans la base de données géographiques et sous quelle forme. Les **spécifications de saisie** sont utilisées par les opérateurs des chaînes de production pour savoir comment produire l'information désirée, en termes de techniques et de méthodes (Vauglin, 1997). Les spécifications de contenu explicitent donc les règles de sélection et de représentation réalisées sur l'univers afin d'obtenir une base de données idéale. Cette base de données idéale est appelée univers nominal, ou encore terrain nominal.

Terrain nominal Le terrain nominal se définit comme l'image de l'univers vu à travers le filtre constitué par les spécifications de contenu de la base de données géographiques (cf. figure 1.1). Il représente le contenu virtuellement parfait de la base de données géographiques. Ce concept prendra toute son importance dans ce travail de thèse, car c'est vers ce terrain nominal que cherche à tendre tout producteur de bases de données géographiques.

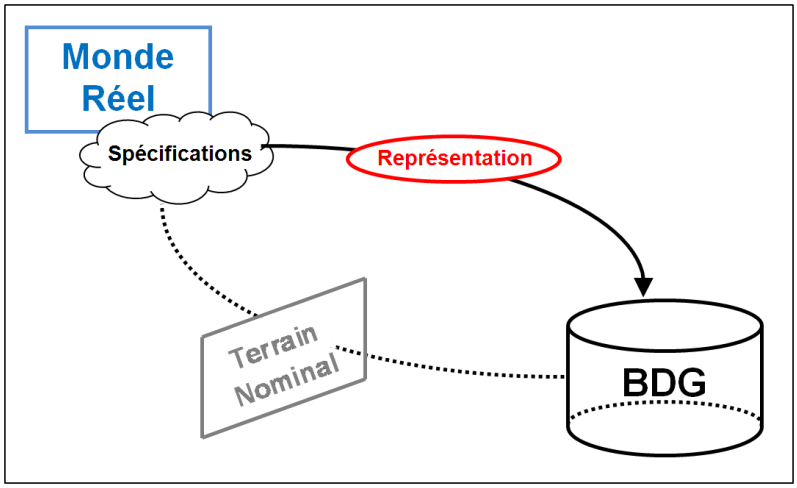
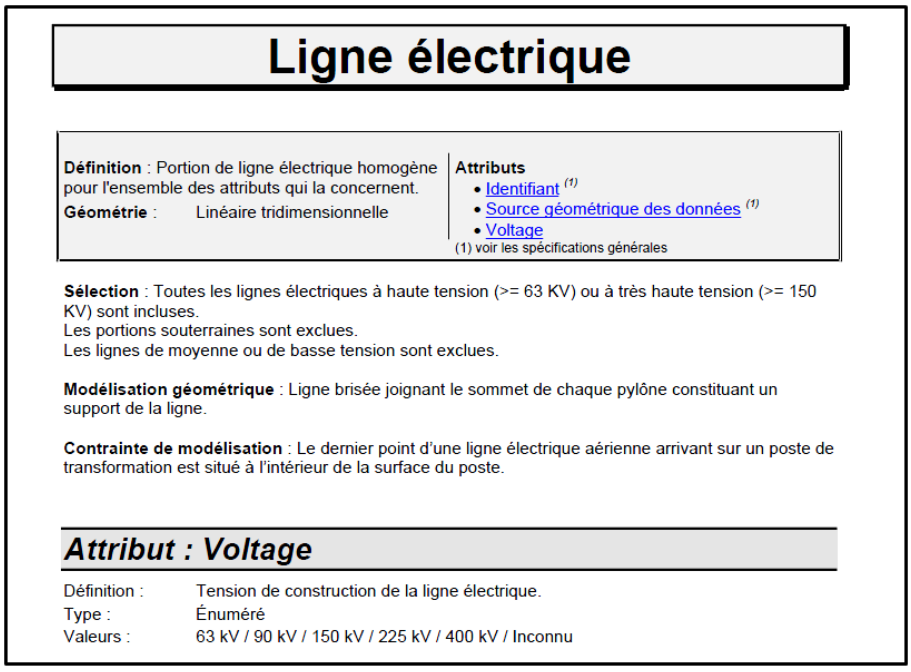


Figure 1.1 – Une base de données géographiques vue à travers le filtre des spécifications

Pour les bases de données géographiques, les spécifications se matérialisent sous forme de volumineux documents textuels. Ainsi, à l'Institut Géographique National (IGN), les spécifications des différentes bases de données présentent toutes une structure relativement similaire (Gesbert *et al.*, 2004), comme illustré en figure 1.2.



Ligne électrique

Définition : Portion de ligne électrique homogène pour l'ensemble des attributs qui la concernent.
Géométrie : Linéaire tridimensionnelle

Attributs
- Identifiant (1)
- Source géométrique des données (1)
- Voltage

(1) voir les spécifications générales

Sélection : Toutes les lignes électriques à haute tension (>= 63 KV) ou à très haute tension (>= 150 KV) sont incluses.
Les portions souterraines sont exclues.
Les lignes de moyenne ou de basse tension sont exclues.

Modélisation géométrique : Ligne brisée joignant le sommet de chaque pylône constituant un support de la ligne.

Contrainte de modélisation : Le dernier point d'une ligne électrique aérienne arrivant sur un poste de transformation est situé à l'intérieur de la surface du poste.

Attribut : Voltage

Définition : Tension de construction de la ligne électrique.
Type : Énuméré
Valeurs : 63 kV / 90 kV / 150 kV / 225 kV / 400 kV / Inconnu

Figure 1.2 – Exemple de fiche de spécifications extraite de la BDTOPO (IGN, 2011c)

Dans les spécifications des bases de données géographiques de l'IGN, chaque classe d'objets constitue un paragraphe découpé en plusieurs parties . Les critères de sélection des objets de la classe y sont clairement explicités, ainsi que la définition des attributs et leurs valeurs. La modélisation géométrique de l'objet et les contraintes de saisie y sont également définies.

Par exemple, dans la BDTOPO (IGN, 2011c), les "tronçons de chemins" sont modélisés par un linéaire tridimensionnel, saisi à l'axe de la route, au sol. Cet exemple illustre bien le fait que les données géographiques n'offrent qu'une représentation simplifiée de la réalité perçue à travers le filtre des spécifications, car dans la réalité un chemin ne se limite pas à une simple ligne tracée à partir de son axe virtuel.

Schéma de données L'ensemble des classes d'objets d'une base de données géographiques sont organisées au sein d'un schéma conceptuel, communément appelé schéma de données, composé - suivant l'approche objet - de classes, d'objets et de relations. Ce schéma conceptuel traduit la catégorisation faite du monde réel (Ruas et Mustière, 2005). Il indique les différentes classes de la base (liées entre elles par des relations) et leurs attributs (cf. figure 1.3). La spécificité des bases des données géographiques fait que les classes disposent d'un attribut géométrique.

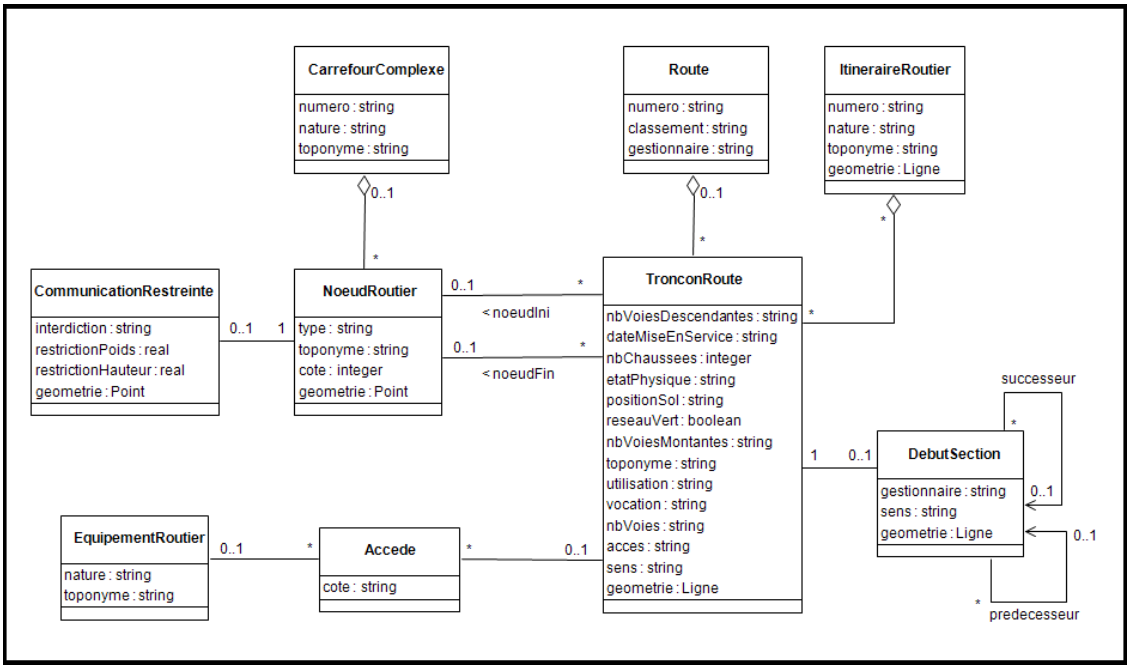


Figure 1.3 – Extrait du schéma de données de la BDCARTO pour le thème routier

Nous venons de voir que la représentation du monde réel au sein d'une base de données géographiques n'est pas triviale, et qu'elle impose une simplification de la réalité. Cette représentation suppose une modélisation géométrique et sémantique adaptée au phénomène à représenter et aux usages prévus. La prise en compte de ces contraintes permet de définir le degré de simplification du monde réel, qui se manifeste en termes de contenu et de niveau de détail de la base de données géographiques.

1.1.1.3 Contenu et niveau de détail des bases de données

Le degré de simplification du monde réel recherché pour la création d'une base de données géographiques conditionne directement son contenu et son niveau de détail. Il est nécessaire de comprendre comment se manifeste ce niveau de détail sur les objets géographiques puisqu'il se répercute directement sur la géométrie des objets, et par voie de conséquence, sur les mesures de longueur et de surface calculées en aval.

Niveaux de détail Toute base de données géographiques dispose d'un niveau de détail et d'un contenu qui lui est propre. En effet, d'après Ruas (2002), "*une base de données correspond à une échelle de raisonnement ou d'analyse [...] et peut être représentée cartographiquement dans une gamme d'échelles mathématiques compatible avec l'ordre de grandeur de l'ensemble des objets correspondants à cette analyse*". Cette notion d'échelle de raisonnement ou d'analyse conditionne donc le niveau de détail de la base de données. Celui-ci concerne à la fois la géométrie des objets et les informations attributaires associées. On parle ainsi de niveaux de détail géométrique et sémantique.

Le **niveau de détail sémantique** traduit la finesse de représentation du monde réel. Il est visible au niveau du schéma de données, non seulement à travers la définition des classes présentes (et donc aussi celles qui sont absentes) mais également par le nombre et la précision des attributs qui les caractérisent. Par exemple, l'attribut "largeur" d'une classe route pourra être saisi au centimètre près, au mètre près, voire par intervalles. Le niveau de détail sémantique définit également les règles de sélection des objets à représenter dans la base de données par rapport au schéma de données.

Le **niveau de détail géométrique** traduit quant à lui la finesse de la description géométrique des objets. Il est principalement lié au type d'implantation choisie pour représenter la géométrie des objets (ponctuelle, linéaire, surfacique), mais concerne également la précision géométrique et la granularité des objets vectoriels.

Résolution et granularité Pour décrire le niveau de détail géométrique d'une base de données, la notion de **résolution** est souvent utilisée. La résolution correspond généralement à la plus petite unité mesurable ou représentable. Cette définition est commodément applicable aux données représentées en mode maillé, puisque dans ce cas la résolution (spatiale) correspond à la taille du pixel (cf. figure 1.4).

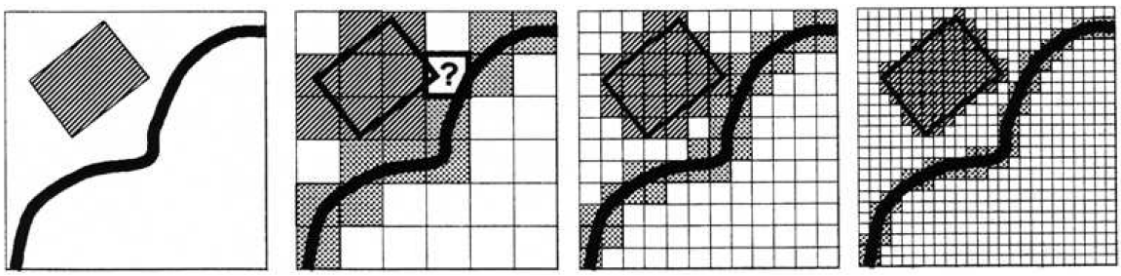


Figure 1.4 – La résolution spatiale de données maillées : représentation d'une route et d'un bâtiment, d'après Ruas (2002)

Cependant pour les données représentées en mode vecteur, le terme de résolution est plus ambigu (Ruas, 2002). En mode vecteur, il est préférable d'utiliser la notion de **granularité**, comme préconisé par Ruas (2002), qui décrit la taille des plus petites formes géométriques présentes dans la base. On peut par exemple approximer la taille des plus petites formes d'un objet par la distance entre les points successifs composant sa géométrie. Pour décrire le niveau de détail géométrique d'une base de données géographiques, on associe également des informations de **précision géométrique** (sur lesquelles nous reviendrons plus tard), qui qualifient l'écart entre la position de l'objet représenté et la position réelle.

La notion de granularité est particulièrement intéressante dans le cadre de ce travail de thèse. Comme illustré dans la figure 1.5, extraite de Ruas (2002), la saisie d'une même courbe modélisée par deux successions régulières de points de granularités différentes, renverra deux polylignes de longueurs différentes, même si la précision géométrique des points constituant leur géométrie est identique. On voit bien dans cet exemple que la granularité des objets vectoriels impacte les mesures réalisées à partir de leur géométrie.

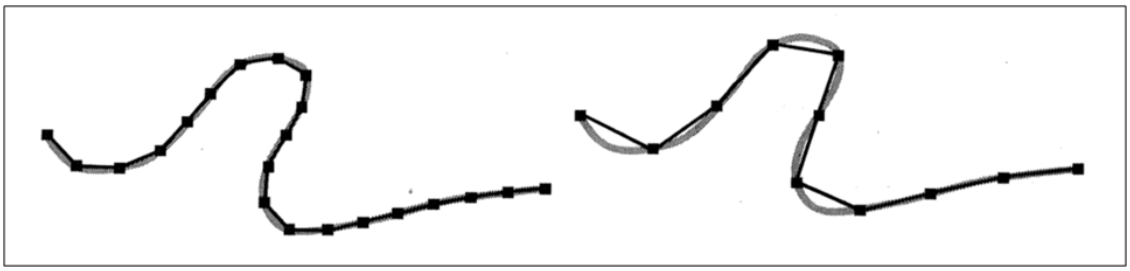


Figure 1.5 – Une même courbe décrite par deux successions de points de granularités différentes, d'après Ruas (2002)

Il est important de noter que la granularité des objets n'est généralement pas constante au sein d'une même base de données. Par exemple, pour modéliser la géométrie d'une route par une polyligne, la granularité est généralement plus grande quand la sinuosité augmente, afin de représenter plus fidèlement les virages, que lorsque celle-ci diminue.

1.1.1.4 Impact de la représentation à différentes échelles

Selon les échelles de raisonnement et d'analyse des phénomènes du monde réel considérés, les niveaux de détail des bases de données géographiques s'avèreront ainsi plus ou moins adaptés. En effet, on ne peut pas utiliser des données produites 1 :5.000 pour réaliser des cartes au 1 :1.000.000, du fait de la surcharge d'informations qui rendrait la carte illisible. Il devient à ce moment nécessaire de pouvoir représenter un même territoire à différentes échelles d'abstraction.

A l'Institut Géographique National, comme chez les autres producteurs de données géographiques institutionnels ou privés, plusieurs bases de données sont produites avec des niveaux de détail géométrique différents. C'est par exemple le cas entre la BDTOPO, Base de Données Topographique (IGN, 2011c) et la BDCARTO, Base de Données Cartographiques (IGN, 2011b), comme illustré en Figure 1.6.

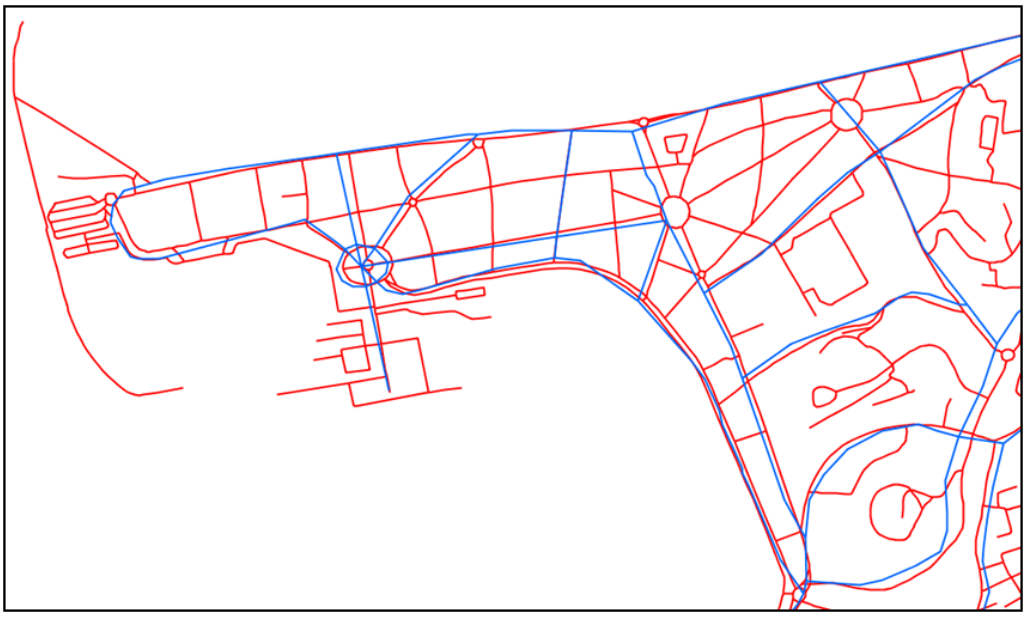


Figure 1.6 – Un même réseau routier saisi selon les spécifications de la BDCARTO (en bleu) et de la BDTOPO (en rouge)

Les spécifications de saisie respectives de la BDTOPO et de la BDCARTO impliquent une granularité ainsi qu'une précision géométrique différentes, mais également un niveau de détail sémantique différent, qui se traduit en termes de complétude. En effet, la figure 1.6 montre bien que tous les tronçons de route ne sont pas présents au sein de la BDCARTO, par comparaison avec la BDTOPO, mais uniquement les principaux. Nous voyons bien à travers cet exemple que le niveau de détail recherché influe sur :

– la modélisation géométrique,
– la sélection des objets représentés.

Au delà de la surcharge d'informations sur une carte, la production de bases de données à plusieurs niveaux de détail s'explique également pour des raisons de volumétrie de la donnée. En effet, la manipulation de bases de données géographiques comme la BDTOPO sur de grandes étendues, suppose des ressources informatiques importantes, tant en termes de stockage, que de mémoire pour réaliser des calculs. La table 1.1 détaille la taille du fichier, le nombre d'objets et de sommets, ainsi que la longueur totale du réseau routier du département des Pyrénées-Atlantiques, représenté par la BDTOPO et la BDCARTO.

	BDTOPO	BDCARTO	Ratio
Taille du fichier .shp (en Ko)	72523	4460	16,3
Nombre d'objets	194956	19498	10
Nombre de sommets	2509402	217131	11,6
Longueur totale (en km)	36011	14295	2,5

Table 1.1 – Volumetrie du réseau routier du département des Pyrénées-Atlantiques représenté par la BDTOPO et la BDCARTO

Le niveau de détail conditionne donc le degré de généralisation de l'information géographique. La **généralisation** se définit comme un processus de simplification des données pour les adapter à une échelle caractéristique. Par définition, en tant qu'abstraction du monde réel, toute base de données géographiques est généralisée. Cependant, et peut-être par abus de langage, on entend davantage par généralisation l'ensemble des processus permettant de réduire les niveaux de détail sémantique et géométrique d'une base de données pour qu'elle s'adapte à une plus petite échelle. La généralisation est dans ce cas vue comme un processus de transformation d'une base de données (ou d'une carte) pour en produire une nouvelle. Cette vision de la généralisation sera définie en section 1.1.2.3.

Comme plusieurs bases de données peuvent être produites pour représenter une même réalité, à des niveaux de détail qui leur sont propres, cela implique des processus de production adaptés au niveau de détail de la base de données considérée. En effet, les processus de production utilisés pour saisir des objets géographiques au 1 :10.000 ne sont pas les mêmes que ceux utilisés pour saisir des objets au 1 :1.000.000. De plus, chaque processus de production génère inévitablement des erreurs, qu'il est nécessaire de comprendre, puisque ces erreurs affectent la géométrie des objets vectoriels, et par conséquent les mesures géométriques de longueur ou de surface. Ainsi, nous aborderons dans la section suivante la notion d'erreur dans les bases de données géographiques, et nous intéresserons plus particulièrement aux erreurs impliquées par les processus de production mis en œuvre pour saisir la géométrie des objets géographiques.

Synthèse Section Nous avons vu dans cette section que les bases de données géographiques vectorielles permettent une représentation simplifiée et commode des phénomènes du monde réel. Elles sont par conséquent inévitablement entachées d'imperfections. Le degré de simplification du monde réel varie notamment en fonction du niveau de détail recherché. Celui-ci conditionne le degré de généralisation de la géométrie des objets, et par conséquent impacte les mesures de longueur et de surface qui seront calculées en aval.

1.1.2 Les erreurs dans les bases de données géographiques

Cette section a pour objectif de présenter la notion d'erreur dans les bases de données géographiques, en particulier au niveau de la géométrie des objets. Pour cela, nous reviendrons dans un premier temps sur la notion d'erreur de mesure, avant de nous intéresser davantage aux sources et aux types d'erreurs affectant les bases de données géographiques. Nous terminerons cette section en nous focalisant sur les erreurs impliquées par le processus de production des données géographiques.

1.1.2.1 La notion d'erreur de mesure

Le concept de mesure La langue française désigne le terme de **mesure** comme étant à la fois l'opération de mesurer et son résultat, ce qui peut conduire à une certaine ambigüité dans son interprétation. Cependant en métrologie, pour bien distinguer ces deux notions, à l'instar des anglophones, le terme de "mesure" est réservé au résultat (comme l'anglais *measure*), et l'opération est elle appelée "mesurage" (comme l'anglais *measurement*). Le **mesurage** peut ainsi être défini comme l'*opération ayant pour but de déterminer la valeur d'une grandeur*.

L'erreur de mesure Il se peut qu'une mesure soit entachée d'erreur. L'**erreur** se définit comme *la différence entre la mesure X_i et la valeur vraie de la mesure X_R* (Heuvelink, 1998), comme exprimé dans l'équation 1.3.

$$d_X = X_i - X_R \tag{1.3}$$

où X_i représente une mesure et X_R la valeur réelle

Type d'erreurs Lorsque l'on répète un grand nombre de fois le mesurage d'une même grandeur, on obtient une distribution des mesures, ce qui permet de distinguer deux types d'erreurs :

- l'**erreur accidentelle**, définie comme une *erreur qui varie d'une façon imprévisible en valeur absolue et en signe lorsqu'on effectue un grand nombre de mesurages de la même valeur d'une grandeur dans des conditions pratiquement identiques* (AFNOR, 1955). Ce type d'erreur revêt donc un caractère aléatoire. Elle est généralement modélisée par une loi de probabilité, comme une loi normale par exemple.
- l'**erreur systématique**, définie quant à elle comme une *erreur qui, lors de plusieurs mesurages effectués dans les mêmes conditions, de la même valeur d'une certaine grandeur, reste constante en valeur absolue et en signe* (AFNOR, 1955). Les causes de ces erreurs peuvent être connues ou inconnues, mais peuvent en général être facilement compensées.

De ces deux types d'erreurs (accidentelles et systématiques), il est important de bien distinguer la notion de **faute**, qui est une erreur grossière, généralement d'origine humaine, ou encore d'un problème d'instrumentation. Les notions d'exactitude et de précision (qui sont souvent confondues) font directement référence aux deux types d'erreurs définis précédemment.

La notion d'exactitude L'exactitude permet de *mesurer la fluctuation d'une série de mesures autour d'une valeur réelle*. Elle est modélisée statistiquement par l'erreur moyenne quadratique (cf. équation 1.4). L'exactitude est conditionnée par la connaissance des erreurs systématiques, qui permettent de déterminer un biais de mesure. Le **biais** est défini comme la différence entre l'espérance de la série de mesure et la valeur réelle.

$$EMQ_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_X^2} \tag{1.4}$$

où d_X représente l'erreur de mesure

La notion de précision La précision, quant à elle, permet de *mesurer la fluctuation de cette série de valeurs autour de son espérance*. Elle fait donc davantage référence aux erreurs accidentelles, et est généralement modélisée statistiquement par l'écart-type (cf. équation 1.5) des erreurs de mesure.

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_X - E_X)^2} \tag{1.5}$$

où d_X représente l'erreur de mesure et E_X l'espérance des erreurs de mesure

En représentant la répartition statistique d'une série de mesures par une distribution gaussienne (même si ce n'est pas forcément le cas dans la réalité), la figure 1.7 permet d'illustrer les relations entre les notions de précision et d'erreurs accidentelles, et celles de biais et d'erreurs systématiques.

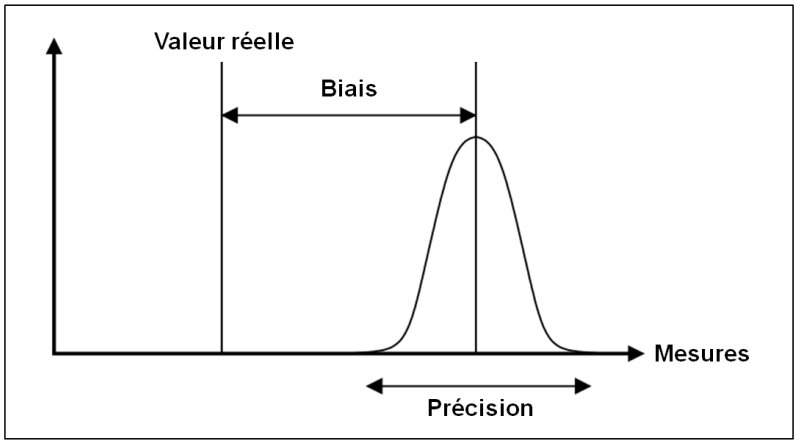


Figure 1.7 – Les notions de précision et de biais représentées par une distribution statistique

De manière plus schématique, les notions d'exactitude, de précision et de biais sont illustrées dans la figure 1.8, où la valeur réelle est représentée par le centre de la cible et la série de mesures par les points noirs.

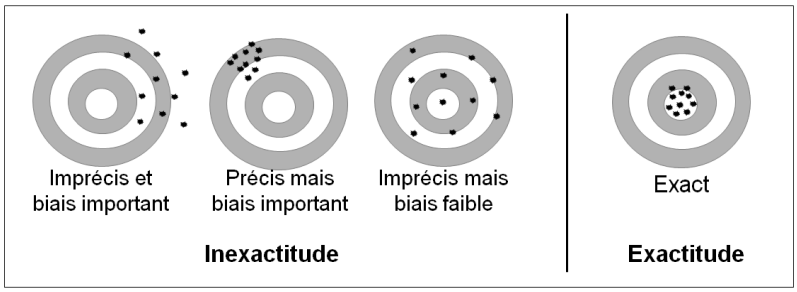


Figure 1.8 – Les notions d'exactitude, de précision et de biais représentées par une série de mesures sur une cible

Nous pouvons remarquer sur la figure 1.8 que la notion d'exactitude est conditionnée à la fois par la précision de la série de mesures, mais aussi par son faible biais. Cependant, la notion d'exactitude est une notion bien relative, en particulier si l'on fait intervenir la notion d'échelle. En effet, si une série de mesure peut s'avérer "exacte" au 1 :100.000, ce ne sera pas forcément la cas 1 :25.000, comme illustré dans la figure 1.9.

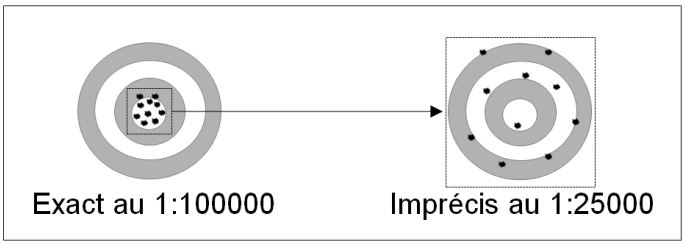


Figure 1.9 – Impact de la notion d'échelle sur l'exactitude d'une série de mesures

Dans le contexte des données géographiques, il est nécessaire de qualifier et de modéliser les erreurs qui impactent la géométrie des objets, afin de prévenir les mauvaises utilisations, en particulier celles liées aux mesures géométriques de longueur et de surface.

1.1.2.2 Les sources d'erreurs dans les données géographiques

Les données géographiques sont donc entachées d'erreurs, mais ces erreurs peuvent être d'origines diverses. L'un des premiers auteurs à avoir proposé une typologie des sources d'erreurs est Burrough (1986). Il propose de diviser les sources d'erreurs dans les données géographiques en trois catégories :
- les sources d'erreurs évidentes (*Obvious sources of errors*)
- les erreurs résultant de variations naturelles ou de mesurages originaux (*Errors resulting from natural variations or from original measurements*)
- les erreurs survenant lors d'opérations (*Errors arising through processing*)

Cette première typologie ne concerne que les erreurs potentielles liées à une seule base de données géographiques. Cependant, dans la pratique, les utilisateurs sont bien souvent amenés à utiliser des bases de données d'origines diverses, ayant chacunes leurs propres sources d'erreurs, ce qui peut générer une **propagation d'erreurs**, en particulier lors de traitements topologiques.

Ainsi, d'autres auteurs ont proposé des classifications des erreurs dans les données géographiques en faisant référence à cette notion de propagation d'erreur. Par exemple, Beard (1989) propose de classifier les erreurs en trois catégories, en distinguant les erreurs reliées :
– à l'acquisition et la compilation des données (*source error*)
– au traitement (*process error*)
– à l'utilisation des données (*use error*)

Hunter et Beard (1992) étendent cette classification en ajoutant aux sources les types d'erreurs. Dans cette approche, traduite par Devillers (2005), les trois catégories de sources d'erreurs proposées par Beard (1989) génèrent des erreurs primaires (erreur de position et erreur attributaire) et des erreurs secondaires (cohérence logique et exhaustivité), le tout formant l'erreur globale attachée au produit final (cf. figure 1.10).

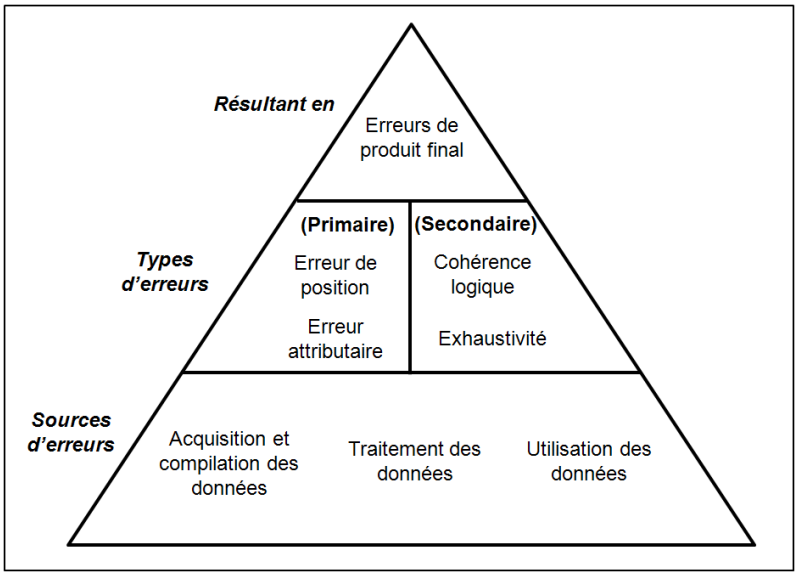


Figure 1.10 – Classification des erreurs proposée par Hunter et Beard (1992), traduite par Devillers (2005)

Enfin, Collins et Smith (1994) proposent une classification des erreurs selon les phases de traitement de l'information géographique. Ils différencient les erreurs liées :
– à la collecte des données
– à la saisie des données
– au stockage des données
– à la manipulation des données
– à la sortie des données
– à l'utilisation des résultats

A travers ces différentes classifications des erreurs, nous observons que l'ensemble de la chaîne de traitement de l'information géographique en est entachée. Cependant, Chrisman (1995) considère qu'il faut pouvoir "vivre avec" ces erreurs, et en maitriser les conséquences même dans les manipulations les plus courantes. Afin de mieux maitriser l'impact de ces erreurs, nous allons nous intéresser plus spécifiquement aux erreurs impliquées par les processus de production des données géographiques, qui font référence aux erreurs d'acquisition des données dans la classification de Beard (1989).

1.1.2.3 Erreurs d'acquisition et processus de production

Le monde réel peut être représenté par différentes bases de données géographiques ayant un niveau de détail qui leur est propre. Ainsi, de nombreux producteurs de données maintiennent en parallèle des bases de données, représentant la même portion du monde réel, à différents niveaux de détail. Le niveau de détail recherché conditionne étroitement les processus de production mobilisés pour la saisie de la géométrie des objets vectoriels. Nous allons voir que chaque processus de production génère des erreurs d'acquisition, plus ou moins grandes, qu'il est nécessaire d'estimer, voire de compenser.

Notion de processus de production A travers le terme "processus de production", nous entendons l'ensemble des processus permettant de produire la géométrie des objets d'une base de données géographiques vectorielles, que ce soit à partir d'une abstraction du monde réel, ou bien de bases de données existantes (cf. figure 1.11). Deux grandes catégories de processus de production d'une base de données géographiques peuvent être différenciées : les moyens d'acquisition primaires et secondaires.

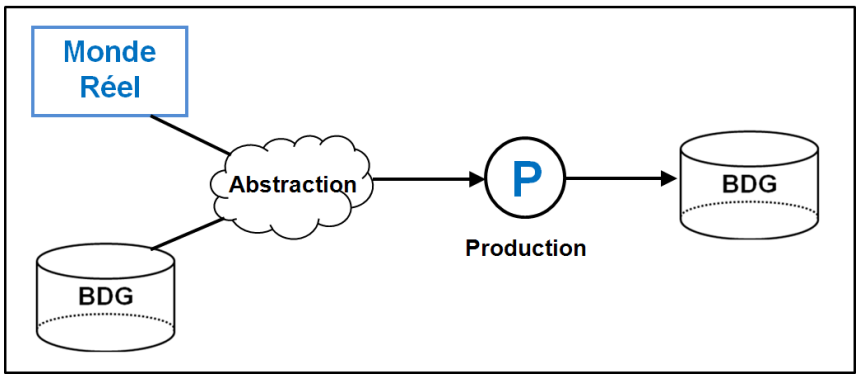


Figure 1.11 – Le concept de processus de production d'une base de données géographiques

Les moyens d'acquisition primaires reposent sur la mesure d'angles et/ou de distances pour saisir les coordonnées utiles à la construction des géométries des objets vectoriels. Le calcul des coordonnées de points inconnus est effectué à partir de points connus au sein d'un système de coordonnées de référence. Sans s'attarder sur les méthodes traditionnelles de saisie d'objets géographiques (comme le relevé à la planchette), nous présenterons comme moyens d'acquisition primaires les relevés topographiques, les relevés par technologies GNSS, ainsi que les relevés LIDAR.

Les moyens d'acquisition secondaires d'objets géographiques vectoriels supposent quant à eux l'acquisition préalable de sources de saisie. Ces sources de saisie peuvent être des photographies aériennes, des images satellites, ou encore des plans ou cartes scannées. Les objets sont ensuite produits par le biais d'outils de restitution informatiques, que nous appelons ici moyens d'acquisition secondaires. On peut également considérer que les processus de généralisation de bases de données sont des moyens d'acquisition secondaires, puisqu'ils supposent au préalable l'existence d'une base de données source.

Les principaux types de processus de production d'objets vectoriels utilisés aujourd'hui sont décrits dans cette section, sans toutefois les différencier en termes de moyens d'acquisition primaires ou secondaires.

a. Les relevés topographiques en général

Les relevés topographiques sont un moyen d'acquisition direct, réalisés sur le terrain à l'aide d'instruments de relevé topographique, comme le théodolite ou le tachéomètre (qui est un couplage entre un théodolite et un distance-mètre) pour ne citer qu'eux.

Principe général L'opération de saisie de la géométrie des objets consiste à déterminer les coordonnées de la position de l'instrument, à partir de calculs d'angles et de distances, en "visant" des points connus dans un système de référence. Les calculs d'angles et de distances sont réalisés de proche en proche, par transmission du gisement. Le **gisement** se définit comme l'angle, dans le plan horizontal, entre un vecteur (défini par deux points connus en coordonnées) et le nord cartographique. Il est compté dans le sens horaire. Le cheminement de proche en proche à partir de points connus permet de calculer les coordonnées de chaque point. Les techniques de relevé topographique sont généralement utilisées pour saisir des objets géographiques à grande échelle, comme des plans cadastraux par exemple.

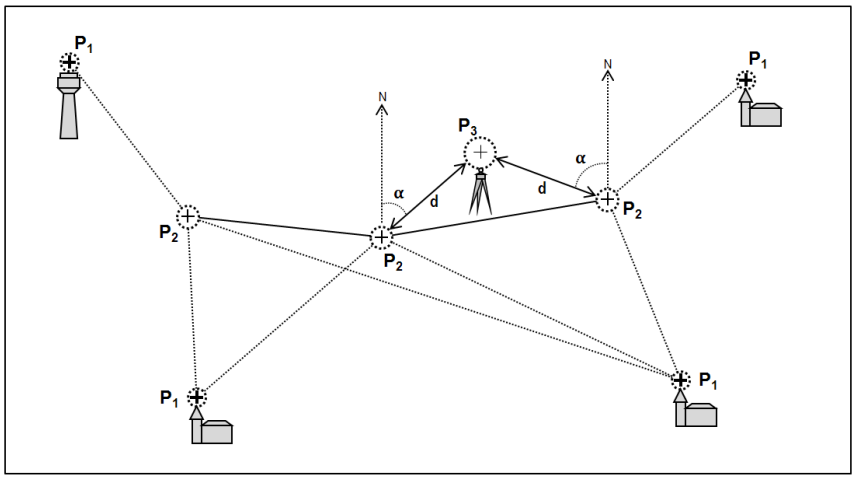


Figure 1.12 – Calcul des coordonnées d'un point P_3, à partir d'un réseau de référence de points P_1, et d'une densification locale de points P_2. La taille du cercle pointillé autour de chaque point illustre l'imprécision de localisation.

Erreurs associées Même si ces moyens de saisie peuvent être considérés comme très précis, les coordonnées calculées sont néanmoins inévitablement entachées d'erreurs, même minimes. En effet, les points connus du réseau portent tous une imprécision de localisation. De plus, chaque opération de calcul d'angle ou de distance est également génératrice d'erreur (cf. figure 1.12).
La saisie par le biais de visées successives a également pour effet de propager ces erreurs sur l'ensemble du cheminement. Ces erreurs peuvent être systématiques, dues par exemple aux imprécisions de l'instrument (qualité des composants, défauts de réglages...), ou encore accidentelles, comme par exemple l'erreur de pointé de l'opérateur, ou encore du fait d'éléments non-maitrisables (température, hygrométrie...). Ces erreurs sont généralement corrigées en partie par différentes méthodes de compensation (proportionnelles, pondérées, par moindres carrés) qui auront pour effet de réduire localement les erreurs en les reportant sur l'ensemble des autres points saisis (Henrye, 2005).

b. Les relevés GNSS

On appelle GNSS (*Global Navigation Satellite System*) ou Systèmes Global de Positionnement par Satellite, l'ensemble des systèmes de positionnement basés sur des signaux émis par des satellites en orbite autour de la Terre et fournissant une couverture mondiale. Le système opérationnel le plus connu, le système GPS (*Global Positionning System*), constituait jusqu'en 2007 le seul réseau GNSS opérationnel. Il a été rejoint par le système russe GLONASS, et deux autres projets sont actuellement en cours de développement, le projet européen GALLILEO, ainsi que le projet chinois BEIDOU (Bosser, 2011).

Principe général Le fonctionnement des GNSS repose sur la mesure du temps de propagation du signal émis par un satellite émetteur jusqu'à sa réception par un récepteur. Le satellite transmet un signal contenant sa position exacte à l'instant de l'émission. Le signal est synchronisé par une horloge atomique embarquée. Le récepteur compare ensuite l'instant d'arrivée du signal par rapport à son horloge propre, ce qui permet de mesurer la distance au satellite. Cette mesure de distance permet de placer le récepteur sur une sphère centrée autour du satellite émetteur. En utilisant au moins trois émetteurs, ces sphères s'intersectent en un seul point, correspondant à la position du récepteur (cf. figure 1.13). Ainsi, afin de saisir la géométrie des objets vectoriels, le récepteur est positionné sur le point dont on cherche à connaître les coordonnées.

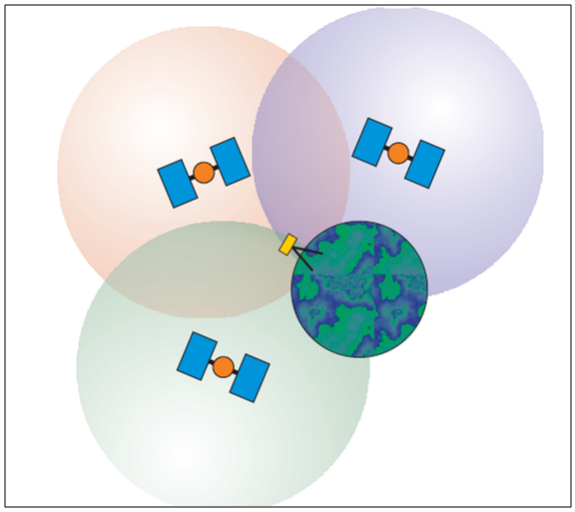


Figure 1.13 – Calcul des coordonnées d'un récepteur par trilatération entre trois satellites, d'après Bosser (2011).

Erreurs associées L'utilisation du GNSS pour le calcul des coordonnées d'objets géographiques est une opération des plus courantes aujourd'hui. Nous savons cependant que les positions calculées par GNSS sont entachées d'erreurs. Ces erreurs peuvent être de plusieurs origines. Elles peuvent être liées au satellite (erreurs d'orbite ou d'horloge), à la propagation des ondes (dans l'ionosphère et la troposphère) pour le calcul de la distance entre les satellites et le récepteur, à l'environnement du récepteur (effets de masque, de multi-trajets) ou encore au récepteur lui-même (qualité des composants).

En outre, la précision de la localisation de chaque point dépendra grandement du type de positionnement utilisé.

- Le **positionnement absolu**, le plus courant, nécessite au minimum quatre satellites. Il dispose d'une exactitude de positionnement de l'ordre de 5 à 20 mètres.
- Le **positionnement différentiel**, beaucoup plus précis que le positionnement absolu, nécessite l'installation de deux stations au sol à partir de points connus du réseau de référence, qui observent simultanément. Les observations sont ensuite comparées pour déterminer la position relative du récepteur. Disposer de telles stations permet de corriger les imprécisions de positionnement des satellites ainsi que les calculs de distance.

Source		Impact de l'erreur	
		Pos. standard	Pos. précis
Satellite	Horloge	3 m	5 mm
	Orbite	2 m	
	Bruit de code	3 m	-
Propagation	Ionosphère	3 m	1 mm
	Troposphère	1 m	5 - 25 mm
Station	Trajet multiple	1 m	2 - 15 mm
	Bruit de mesure	1 m	2 mm
	Centre de phase	1 m	2 mm
	Horloge	0,5 m	1 mm
Erreur totale typique		10 - 15 m	15 - 20 mm

Table 1.2 – Impact des sources d'erreurs sur la précision d'un positionnement absolu et différentiel, d'après Bosser (2011)

D'un point de vue pratique, le temps d'observation pour la saisie d'une position doit également être pris en compte. En effet, plus le stationnement est long, plus la position calculée s'avèrera exacte. C'est pour cette raison que les mesures de points fixes sont généralement plus précises que les points mobiles. Egalement, des post-traitements peuvent être appliqués aux acquisitions brutes, en utilisant des orbites et des corrections d'horloges satellites précises. L'estimation de l'impact des différentes sources d'erreurs sur la précision d'un positionnement absolu ou différentiel (avec stationnement long et post-traitement) est proposée dans la table 1.2.

c. Les relevés Lidar

Le Lidar (Light Detection and Ranging) est une technologie de mesure optique basée sur l'analyse des propriétés d'une lumière laser renvoyée vers son émetteur.

Principe général Le Lidar permet de calculer la distance à des points par retour d'ondes, à une cadence très élevée. Il peut être porté par un avion (afin de construire un modèle numérique d'élévation) ou par des véhicules terrestres. Le résultat d'une saisie Lidar se présente comme un nuage de points en trois dimensions non-structuré (cf. figure 1.14). Bien que l'on puisse reconnaître des objets géographiques en regardant ce nuage de points, il est cependant nécessaire de reconstruire les plans ou volumes par post-traitement informatique afin de produire la géométrie d'objets vectoriels.

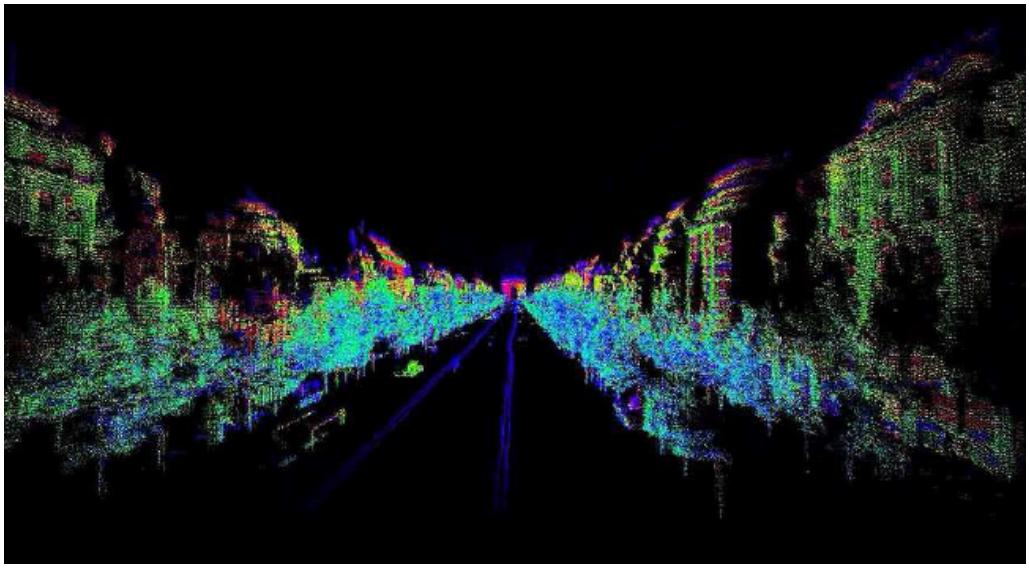


Figure 1.14 – Exemple d'un nuage de points acquis par relevé Lidar sur les Champs Elysées (source IGN)

Erreurs associées Les technologies par relevés Lidar intègrent un système de positionnement GNSS. En effet, pour calculer les coordonnées d'un point à partir de l'émetteur d'où est réalisée la visée laser, il est nécessaire de connaître précisément les coordonnées de l'émetteur. Comme évoqué précédemment, plus l'objet est mobile, plus l'imprécision de son positionnement, et donc de la distance des points saisis, est importante.

Les relevés Lidar sont également entachés d'erreurs d'acquisition qui peuvent être de différentes origines. Celles-ci peuvent être liées à l'instrument, au calcul de trajectographie, à la géométrie du terrain, ou encore à la reconstitution des objets (Lohani, 2010). Comme pour les relevés topographiques et les relevés GNSS, ces erreurs peuvent être compensées par le biais de différents modèles.

d. La restitution photogrammétrique

Les relevés photogrammétriques permettent de reconstituer une vision stéréoscopique du relief en utilisant la parallaxe obtenue entre les images acquises selon les différents points de vue. Cette technique de restitution nécessite l'acquisition préalable d'images (aériennes ou satellites) prises avec des angles de vue différents, ainsi que de points d'appui, qui sont des points facilement identifiables dans les images et dont on connait les coordonnées au sein du système de référence.

Principe général Pour pouvoir réaliser des relevés photogrammétriques, il est nécessaire de reconstruire rigoureusement le modèle géométrique de la prise de vue, à l'aide des données du véhicule support (avion, véhicule terrestre...), de l'angle de vue des images, et des points d'appui. Les points d'appui sont définis dans le cadre de la phase dite de "stéréopréparation".

Une fois les images mises en correspondance, elles permettent, par le biais d'un opérateur ou de processus automatiques ou semi-automatiques, de reconstruire la géométrie des objets en trois dimensions. De nombreuses classes d'objets de la BDTOPO de l'IGN (routes, bâtiments) sont par exemple restituées par relevés photogrammétriques (cf. figure 1.15). L'avantage de ce type de restitution est qu'il permet la saisie des objets en trois dimensions sur de grandes étendues, une fois l'ensemble des images acquises mosaïquées.

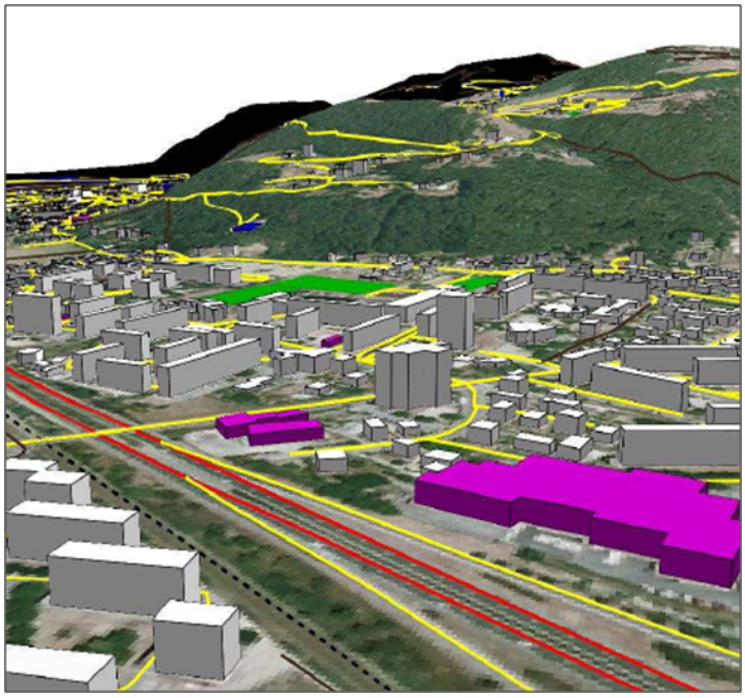


Figure 1.15 – Classes d'objets de bâtiments et de routes de la BDTOPO issues de relevés photogrammétriques

Erreurs associées La qualité de la restitution des objets dépendra de la résolution des images utilisées, ou encore des angles de vue utilisés. Par ailleurs, la qualité de la restitution est assujettie à de nombreuses erreurs potentielles, liées à la reconstitution de la géométrie des images, à des défauts du capteur, ou encore à la restitution humaine de l'opérateur (erreur de pointé).

e. La traitement d'images de télédétection

Le centre canadien de télédétection définit la télédétection comme *l'ensemble des techniques qui permettent, par l'acquisition d'images, d'obtenir de l'information sur la surface de la Terre (y compris l'atmosphère et les océans), sans contact direct avec celle-ci.*
La télédétection englobe tout le processus qui consiste à capter et enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et analyser l'information qu'il représente, pour ensuite mettre en application cette information. Selon cette définition, les images de télédétection peuvent être obtenues par le biais de différents supports, qui peuvent être des satellites, des avions, des drones, ou encore des ballons.

Principe général Sans nous attarder sur les différents types de capteurs (actifs ou passifs) et les caractéristiques de chacun d'entre eux, le traitement d'images de télédétection offre la possibilité de produire des objets vectoriels, à partir de classifications d'images suivant une approche "orientée-objet". Cette approche se différencie des approches classiques de classification d'images, dites "pixel à pixel" où chaque pixel est classé indépendamment selon des critères radiométriques. La démarche orientée-objet ne traite pas le pixel de manière isolée, mais dans son contexte en regroupant des pixels au sein d'objets interprétés en se basant sur leurs valeurs spectrales, leur taille, leur forme, leur contexte. Ainsi, cette approche permet de segmenter une image en polygones classés, afin de restituer automatiquement la géométrie des objets détectés (cf. figure 1.16).

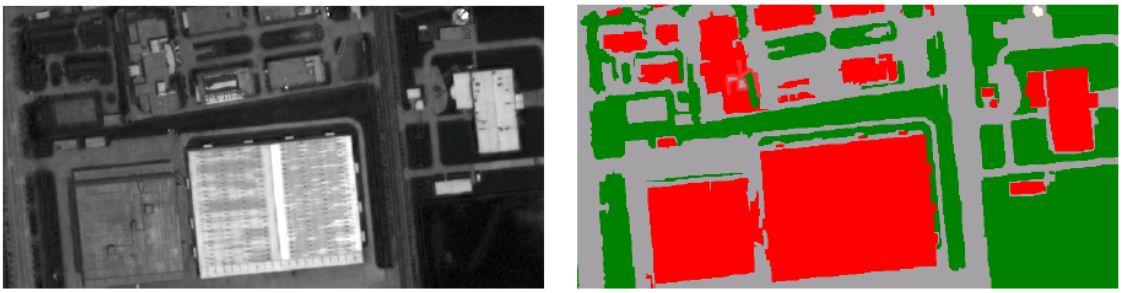


Figure 1.16 – Extraction de la géométrie des bâtiments (en rouge) à partir de la classification orientée-objet d'une image satellite Quickbird, d'après Liu *et al.* (2005)

Erreurs associées Les approches en traitement d'images par classification orientée-objet connaissent un fort intérêt ces dernières années, lié en particulier à l'apparition de logiciels commerciaux (Blaschke *et al.*, 2008). La géométrie des objets produits n'en demeure pas moins impactée d'erreurs ayant des origines classiques en télédétection satellitaire. Ces erreurs peuvent être liées à la géométrie des images, à la résolution (spectrale, spatiale ou radiométrique) des images, ou encore aux algorithmes de classification utilisés.

f. La numérisation manuelle

Lorsque les besoins de l'utilisateur ne l'exigent pas, ou bien que les moyens financiers ne le permettent pas, les spécifications peuvent ne pas imposer de relevés en trois dimensions. Dans ce cas des photographies aériennes ou satellites seront exploitées, sans processus de restitution photogrammétrique. Il peut également arriver, pour diverses raisons, qu'aucune autre source de saisie ne soit disponible mis à part des plans ou des cartes papiers, qu'il faudra numériser préalablement. A titre d'exemple, la constitution de bases de données historiques implique l'intégration de données issues de cartes scannées.

Principe général Les diverses sources de saisie (images, cartes...) sont géoréférencées, à l'aide de points d'appui facilement identifiables et connus dans un système de référence. La numérisation de la géométrie des objets vectoriels est ensuite réalisée à l'aide de logiciels de DAO ou de logiciels SIG disposant de fonctionnalités de saisie.

Ces méthodes de saisie sont extrêmement répandues aujourd'hui, tant les moyens techniques à mettre en œuvre sont facilement disponibles (en particulier par les nombreux logiciels SIG libres disposant de fonctionnalités de géoréférencement et de saisie) et ne nécessitent pas de compétences techniques avancées. Ce type de restitution facilement mobilisable offre l'avantage de pouvoir saisir des données géographiques à grande échelle, sans forcément mettre en œuvre des travaux de complètement de terrain (la saisie étant limitée par ce qui est interprétable sur la carte ou l'image). Les objets saisis par le biais de ces méthodes ne sont généralement restitués qu'en deux dimensions, et bien souvent au format spaghetti (c'est-à-dire en mode non-topologique).

Erreurs associées Les objets saisis par ce type de méthodes sont impactés par les erreurs liées à la qualité des corrections géométriques de l'image. En effet, la précision des points d'appui impacte directement la qualité du géoréférencement. Egalement, si l'image source n'est pas une image orthorectifiée, c'est à dire corrigée des effets du relief, les objets saisis seront affectés d'erreurs de positionnement. Dans le cas où l'image source est une carte scannée, la dégradation du support peut également être génératrice d'erreurs. De même, si la donnée source est une carte généralisée, les effets de la généralisation cartographique se propageront à la géométrie des objets saisis. Enfin, les erreurs accidentelles liées à la restitution de l'opérateur (erreur de pointé) affectent la géométrie des objets saisis.

g. La généralisation automatique

La généralisation est un processus de simplification des données pour les adapter à une nouvelle échelle caractéristique. C'est un processus de diminution du niveau de détail qui peut être comparé au résumé de texte (Ruas, 2002). Ainsi, pour une base de données géographiques, la généralisation permet une diminution des niveaux de détail :
- sémantique : des opérations sont appliquées sur le schéma de données, par le biais par exemple d'agrégations de classes ou de changements de règles de sélection,
- géométrique : des opérations sont appliquées sur les géométries des objets pour par exemple simplifier leur forme ou aggréger plusieurs objets.

La figure 1.17 illustre le résultat d'un processus de généralisation de base de données géographiques, avec changement des niveaux de détail sémantique (la classe bâtiment a été remplacée par la classe surface bâtie et seules les routes de type "nationale" ont été conservées) et géométrique (la granularité des différentes classes a baissé).

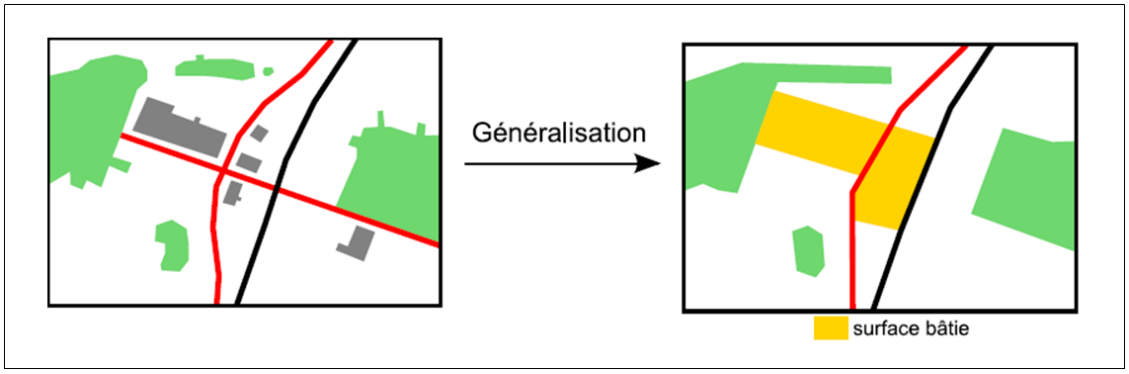


Figure 1.17 – Impact de la généralisation de bases de données, d'après Gaffuri (2008)

La généralisation cartographique est un cas particulier de généralisation. Elle intègre en fait deux types de généralisation :
- La **généralisation de modèle** (ou généralisation conceptuelle) qui implique un changement dans la façon de modéliser le monde réel, en faisant intervenir des modifications sur le schéma de données.
- La **généralisation graphique**, qui prend en compte les règles de lisibilité en effectuant des opérations de simplification, de dilatation, de déplacement et de suppression.

Ainsi, pour produire une carte à moyenne échelle à partir de données à grande échelle, des opérations de généralisation cartographique sont effectuées, puisque le passage d'une carte à grande à échelle à moyenne échelle ne se limite pas à une simple homothétie. Pour que la carte soit lisible et interprétable, il est nécessaire d'appliquer des opérations de généralisation cartographique.

Erreurs associées Une attention toute particulière sera portée sur l'impact de la généralisation (cartographique et de bases de données) dans le cadre de ce travail de thèse. En effet, la généralisation graphique notamment, impacte directement la géométrie des objets généralisés, et par conséquent les mesures de longueur ou de surface. Ceci est particulièrement le cas lorsque la géométrie des objets vectoriels est saisie à partir de cartes. Différents auteurs ont abordé ce sujet, comme Blakemore (1984), João (1998), ou Bard (2004).

1.1.2.4 Impact de l'indépendance des bases de données

Comme nous l'avons vu précédemment, un même territoire peut donc être représenté à des niveaux de détail variables, et selon des règles d'abstraction du monde réel différentes. Si de nombreux travaux de recherche proposent des modèles permettant de généraliser automatiquement des bases de données (Ruas, 1999; Mustière, 2001; Duchêne, 2002), il n'en demeure pas moins que beaucoup de ces bases de données sont produites le plus souvent indépendamment les unes des autres, avec des processus de production différents.

A l'IGN par exemple, la BDTOPO, qui fait partie du Référentiel à Grande Echelle (RGE), est historiquement saisie par le biais de relevés photogrammétriques et de relevés GNSS, tandis que la BDCARTO est historiquement issue de la digitalisation de cartes au 1 :50.000 et de complètements spécifiques. En termes d'usage, la BDTOPO permet la production des cartes de base à des échelles allant du 1 :5.000 au 1 :50.000 et des analyses au niveau local, tandis que la BDCARTO est produite pour effectuer des analyses aux échelles départementales ou régionales, ainsi que la production des cartes au 1 :100.000.

L'automatisation de la généralisation laisse pourtant penser qu'il pourrait être possible de dériver d'une base de données à grande échelle l'ensemble des bases de données produites à plus petite échelle par un processus de généralisation automatique (Brassel et Weibel, 1988). Cette solution idéale n'est pourtant pas mise en œuvre pour diverses raisons.

Mise à jour des données La non-dérivation des bases de données s'explique principalement pour des raisons opérationnelles. En effet, mettre à jour une base de données à grande échelle peut être une tâche extrêmement chronophage. Cela suppose que la dérivation des mises à jour aux plus petites échelles soit conditionnée par la mise à jour de la base à plus grande échelle, qui est la plus gourmande en temps de travail.

Utilisation des données Même si des travaux de recherche interne à l'IGN montrent qu'il est possible de dériver la BDCARTO à partir de la BDTOPO (Touya *et al.*, 2006), la question des usages doit également être prise en compte. En effet, les utilisateurs ayant déjà développé des applications sur les versions antérieures de la BDCARTO, risquent de faire face à de nombreuses différences en utilisant une version générée à partir de la BDTOPO. Par exemple, la géométrie des objets de la BDCARTO, numérisés à partir de cartes au 1 :50.000, est affectée par les effets de la généralisation cartographique, ce qui ne serait pas le cas en dérivant des objets de la BDCARTO à partir de la BDTOPO.

C'est face à ces principales difficultés que les instituts géographiques nationaux, et plus largement les producteurs de bases de données géographiques, ont fait le choix de maintenir à jour parallèlement différentes bases de données à différents niveaux de détail, et en mobilisant des processus de production différents. Cette situation a en particulier pour impact d'affecter chaque base de données d'erreurs de natures hétérogènes selon les processus mis en œuvre. Cette gestion parallèle de plusieurs bases de données oblige également les producteurs à mettre en place des méthodes d'intégration de données, que ce soit au niveau des géométries ou des schémas de données (Sheeren, 2005; Devogèle, 1997), afin de pouvoir mettre en correspondance les objets des différentes bases de données et ainsi faciliter les opérations de mise à jour.

Nous venons donc de voir que des bases de données représentant le même monde réel, à des échelles de raisonnement qui leur sont propres, ne sollicitent pas les mêmes processus de production. Elles sont par conséquent affectées d'erreurs propres au processus de production mobilisé pour saisir la géométrie des objets. Bien que des possibilités de dérivation automatique de bases de données soient envisageables, différentes raisons empêchent la mise en œuvre de ces processus automatisés. L'une de ces raisons fait référence à l'utilisation des données géographiques, qui alimentent bien souvent des systèmes d'aide à la décision.

Synthèse Section Nous avons vu dans cette section que les bases de données géographiques sont inévitablement entachées d'erreurs, de différents types, et qu'il est nécessaire de comprendre leur origine pour maîtriser leur impact. Nous nous sommes principalement focalisé sur les erreurs d'acquisition, qui varient en fonction du processus de production utilisé pour saisir la géométrie des objets. Ces erreurs d'acquisition ont notamment pour impact de se répercuter sur les mesures géométriques de longueur ou de surface.

Conclusion de la section

Nous avons vu dans cette section que les bases de données géographiques sont par définition entachées d'imperfections. Ces imperfections, qui concernent notamment la géométrie des objets vectoriels, sont impliquées par deux processus successifs de natures bien différentes (cf. figure 1.18) :

- le processus d'abstraction du monde réel, qui par le biais de spécifications, opère une simplification des entités du monde réel sous forme d'objets géographiques ($\Delta 1$).
- le processus de production des objets géographiques, qui, selon les moyens et méthodes mobilisés, impacte la géométrie des objets vectoriels d'erreurs plus ou moins importantes ($\Delta 2$).

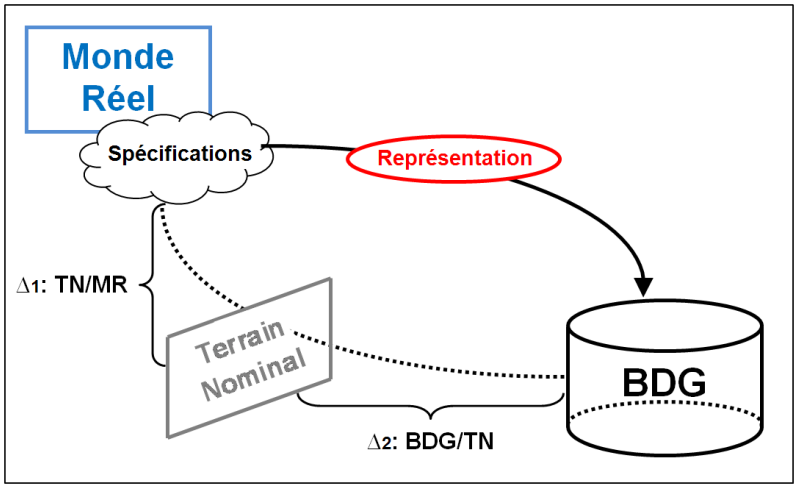


Figure 1.18 – Les deux natures d'imperfections affectant les bases de données géographiques

Ces deux processus générateurs d'imperfections se répercutent sur la géométrie des objets vectoriels, et par voie de conséquence sur les mesures géométriques de longueur et de surface. Malgré cette différenciation théorique, il ne semble pas possible de différencier les impacts de ces deux processus, puisque dans la pratique ils s'accumulent.

Nous allons voir dans la section suivante que ces imperfections affectent les processus décisionnels sollicitant l'information géographique, et que pour réduire l'incertitude associée à une prise de décision, l'évaluation de la qualité des données géographiques doit être envisagée, et communiquée à l'utilisateur final.

1.2 Imperfection, prise de décision et qualité des données géographiques

Depuis le début de cette thèse, nous affirmons que les bases de données géographiques sont par définition entachées d'imperfections, voire d'erreurs. La notion d'imperfection est en effet suffisamment générale pour qualifier les données géographiques. Cependant, nous allons voir dans cette section que l'imperfection sous-entend un ensemble de notions spécifiques, en particulier celle d'imprécision. De manière générale, l'imperfection des données géographiques affecte tout processus de prise de décision les sollicitant. Ainsi, pour réduire l'incertitude liée au processus de prise de décision, il est nécessaire d'évaluer la qualité des données géographiques. Nous introduirons ce concept de qualité des données géographiques, ainsi que les éléments permettant de l'évaluer.

1.2.1 Imperfection des données géographiques et prise de décision

Derrière la notion d'imperfection des données géographiques se cachent de nombreuses notions beaucoup plus spécifiques, notamment celles d'incertitude et d'imprécision, qui ont été abordées par de nombreux auteurs traitant de la qualité des données géographiques. Malgré les nombreuses contributions dans le domaine, aucun consensus terminologique ne s'est véritablement dégagé quant à la définition de ces notions. Ainsi cette section propose de poser définitivement les termes utilisés dans le cadre de cette thèse, qui sont tous des sous-ensembles de la notion d'imperfection, comme proposée par Bouchon-Meunier (1995). Ces définitions seront mises en perspectives avec le processus de prise de décision.

1.2.1.1 Un manque de consensus terminologique

Comme souligné par Devillers *et al.* (2010), il n'existe pas véritablement de consensus au sein de la communauté scientifique de l'information géographique, quant à la terminologie utilisée dans le domaine de la qualité des données géographiques. En effet, de nombreux concepts, comme ceux d'incertitude, d'erreur, de précision, ou d'exactitude sont utilisés, mais pas toujours de manière adéquate.

Tous les auteurs peuvent s'accorder sur le fait que les données géographiques présentent des imperfections (Goodchild, 1995; Hunter, 1998). Nous l'avons vu, au-delà du processus d'abstraction du monde réel, les différents processus de production de la base de données sont également générateurs d'erreurs.

Dans le domaine de l'information géographique, de nombreuses taxonomies des imperfections des données géographiques ont été proposées, et de nombreux termes ont été employés. Les concepts d'incertitude (Fisher, 2003; Fisher *et al.*, 2005), d'imprécision (Worboys, 1998a,b), ou encore de vague (Schneider, 1999) ont été utilisés par les auteurs dans des sens différents, en fonction de leur communauté d'appartenance, des domaines d'applications ou de leurs points de vue (Olteanu, 2008). Ainsi, aucune définition standard n'a véritablement été adoptée, même si Frank (2007) a proposé une ontologie de la qualité des données géographiques.

Dans ce contexte, seules deux définitions sont unanimement reconnues par les différentes communautés, à savoir les concepts d'exactitude et de précision, que nous avons présenté précédemment.

Notion d'incertitude Dans de nombreuses recherches dans le domaine de la qualité des données géographiques, la notion d'incertitude est utilisée comme un concept général (Zhang et Goodchild, 2002). Pour illustrer ce propos, parmi l'ensemble des terminologies proposées, celle proposée par Fisher (2003) et Fisher *et al.* (2005) est particulièrement reconnue au sein de la communauté de l'information géographique.

Fisher (2003) et Fisher *et al.* (2005)) proposent une taxonomie (cf. figure 1.19), qui se présente comme un sous-ensemble de la taxonomie de l'ignorance de Smithson (1989). Les auteurs considèrent ici l'incertitude comme un concept général englobant tous les autres. Comme nous le verrons dans la section suivante, cette vision du concept d'incertitude est en fait similaire au concept d'imperfection dans l'approche de Bouchon-Meunier (1995), que nous adopterons par la suite.

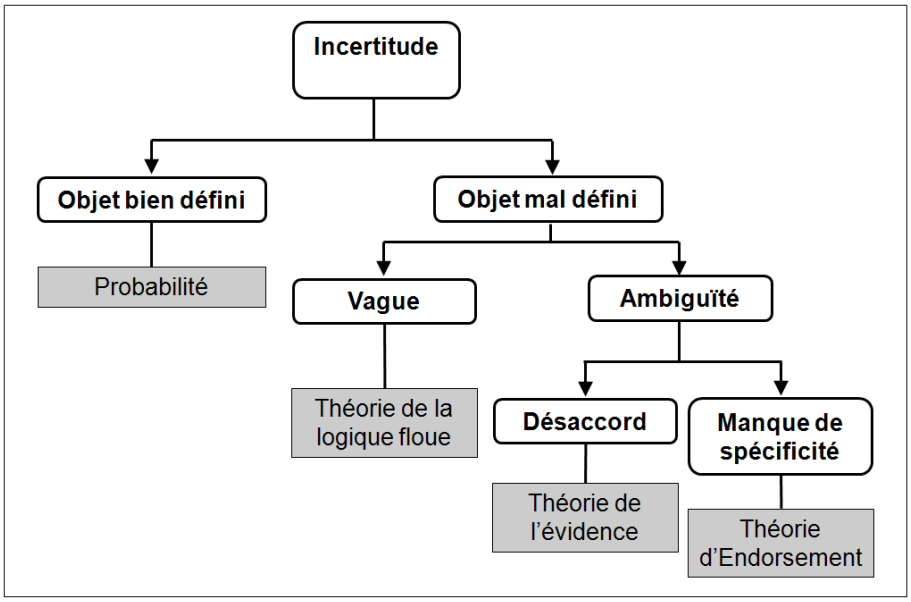


Figure 1.19 – Taxonomie de l'incertitude de Fisher (2003) et Fisher *et al.* (2005) et des modélisations associées

Fisher (2003) et Fisher *et al.* (2005) suggèrent ici que tout processus utilisant les données géographiques prend en compte cette notion d'incertitude des données géographiques, et que cette incertitude est principalement liée au processus d'abstraction du monde réel. Ainsi, les auteurs différencient les objets à la géométrie bien définie (par exemple un bâtiment) des objets à la géométrie mal définie (par exemple une vallée ou encore une forêt). De cette différentiation, les auteurs distinguent trois formes d'incertitude (l'erreur, le vague et l'ambigu) qu'ils proposent de gérer à l'aide de théories mathématiques traitant du domaine de l'incertain.

Dans cette définition, l'erreur concernant les objets bien définis peut être modélisée par des approches statistiques (probabilités). Les objets mal définis, quant à eux, peuvent avoir une géométrie aux limites vagues (comme par exemple les limites d'une forêt) et peuvent être modélisés à l'aide de la théorie des ensembles flous (Zadeh, 1965). Ces objets mal définis peuvent également être de nature ambiguë, faisant intervenir d'autres théories mathématiques.

L'intérêt de l'approche de Fisher (2003) et Fisher *et al.* (2005) dans le cadre de ce travail de thèse, tient dans le principe d'associer le concept d'erreur à des objets bien définis, ainsi que les probabilités comme méthode d'évaluation. En effet, nous nous focalisons dans le cadre de ce travail sur des données de référence, pour lesquelles la modélisation géométrique est bien définie, comme des routes ou des bâtiments par exemple.

Cependant, comme évoqué précédemment, nous n'utiliserons pas la notion d'incertitude telle qu'elle est proposée dans cette approche, nous privilégierons les notions d'imperfection proposées par Bouchon-Meunier (1995).

1.2.1.2 Les notions d'imperfection

Les données géographiques constituent une abstraction du monde réel, donc une simplification, mais qui sont en plus entachées d'erreurs d'origines diverses. En ce sens, les données géographiques sont par définition imparfaites (Goodchild, 1995).

La notion d'imperfection a fait l'objet de nombreux travaux dans le domaine de l'Intelligence Artificielle (Zadeh, 1965; Bouchon-Meunier, 1995). Cette notion d'imperfection fait en fait appel à trois concepts : l'imprécision, l'incertitude et l'incomplétude (Olteanu, 2008) (cf. figure 1.20).

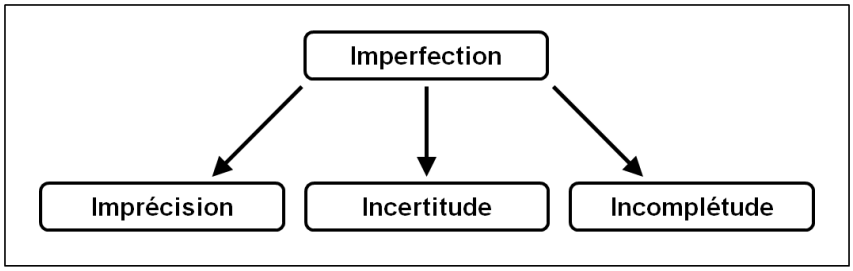


Figure 1.20 – Les grands types d'imperfections, d'après Olteanu (2008)

L'**imprécision** concerne la difficulté d'exprimer précisément un état de la réalité par une proposition. Par exemple, "la distance entre ces deux villes est d'environ 150 km". La notion d'imprécision est ici formalisée par le terme "environ".

L'**incertitude** concerne elle un doute sur les connaissances. Elle formalise le doute ou la prudence de l'observateur quant à une connaissance. Par exemple, "je crois que la distance entre ces deux villes est de 148 km". La notion d'incertitude est ici formalisée par l'expression "je crois".

Enfin, l'**incomplétude** traite de l'absence de connaissances ou de connaissances partielles dans la donnée. Par exemple, si le champ "largeur" d'un objet route n'est pas rempli, ou encore si une route n'est pas présente dans la base de données.

Comme nous le voyons bien à travers ces définitions, ce travail de thèse se focalise sur la notion d'imprécision, à travers sa quantification sur des mesures de longueur et de surface calculées à partir de la géométrie d'objets géographiques.

Les Systèmes d'Information Géographiques (SIG) nous permettent aisément de calculer des longueurs ou des surfaces à partir de la géométrie des objets vectoriels, mais cependant aucune expression de l'imprécision de la mesure n'est proposée. Par exemple, à l'affirmation "la distance entre ces deux villes est d'environ 150 km", l'objectif de ce travail de thèse est de pouvoir estimer le terme "environ", en formalisant une expression du type "la distance entre ces deux villes est de 148 km, plus ou moins 2 km (à 99%)".

Les notions d'incertitude et d'incomplétude, si elles ne sont pas au cœur de ce travail de thèse, seront abordées par rapport aux prises de décision impliquées par des informations imprécises.

1.2.1.3 Incertitude et prise de décision

Les données géographiques sont utilisées comme support d'aide à la décision dans de nombreux domaines (Longley *et al.*, 2005). Comme l'indique Devillers (2004) ces données contiennent toujours un niveau d'incertitude, qui les rendent plus ou moins adaptées à l'utilisation qui en sera faite. Ainsi, il arrive qu'un utilisateur prenne des risques à utiliser des données non-adéquates dans certains processus de décision, pour lesquels les SIG facilitent pourtant grandement la tâche. L'utilisateur peut donc se retrouver dans une situation d'incertitude quant à la décision qu'il doit prendre en utilisant ses données.

Ainsi, Goodchild (1995) suggère que les recherches en qualité des données géographiques ne se concentrent pas uniquement sur la description et la communication de la qualité des données géographiques à l'utilisateur final, mais également sur l'impact qu'ont ces informations sur les processus de décision. Cette idée rejoint le concept de "use error" proposé par Beard (1989) quant aux mauvaises utilisations pouvant être faites des données géographiques pour certains usages.

Lorsqu'un utilisateur doit faire face à une situation d'incertitude lors d'un processus de prise de décision à l'aide de données géographiques, Bédard (1987) considère qu'il est en mesure de choisir entre : 1- ne rien faire, 2- essayer de diminuer le niveau d'incertitude, 3- prendre la décision en acceptant les conséquences possibles, en "absorbant" l'incertitude.
Pour réduire l'incertitude, Epstein *et al.* (1998) proposent quant à eux : 1- d'acquérir plus d'information, 2- d'améliorer la qualité de l'information.
Egalement, Hunter (1998) propose une stratégie globale pour gérer l'incertitude dans les SIG en intégrant les concepts d'absorption et de réduction d'incertitude. Il propose notamment qu'une comparaison soit réalisée entre les besoins des utilisateurs et les caractéristiques de la base de données.
Enfin, Agumya et Hunter (1996) proposent une approche permettant de déterminer le niveau d'acceptabilité de l'incertitude, en analysant les risques potentiels pouvant être associés à une prise de décision basée sur ces données.

Si différentes stratégies de gestion de l'incertitude associées à un processus de prise de décision ont été proposées dans ces contributions, les auteurs s'accordent à dire qu'il est nécessaire de communiquer préalablement un ensemble d'informations sur la qualité des données géographiques utilisées. Dans ce contexte, la section suivante a pour objectif de présenter comment est évaluée et communiquée la qualité des données géographiques.

Synthèse Section Nous avons vu dans cette section qu'il n'existe pas véritablement de consensus terminologique autour du concept d'imperfection des données géographiques. Ainsi, dans le cadre de ce travail de thèse, la terminologie proposée par Bouchon-Meunier (1995) sur les différents types d'imperfections sera utilisée, la notion d'imprécision que l'auteur propose étant la plus appropriée dans le cadre de ce travail. Nous cherchons en effet dans le cadre de cette thèse à estimer l'imprécision des mesures géométriques de longueur et de surface, afin d'aider un utilisateur à réduire l'incertitude d'un processus de prise de décision sollicitant ces mesures.

1.2.2 La qualité des données géographiques

Pour réduire l'incertitude d'un processus de prise de décision sollicitant des données géographiques, il est nécessaire de communiquer à l'utilisateur des informations relatives à la qualité de données utilisées. Aprés avoir présenté la notion de qualité des données géographiques, cette section va nous permettre de voir comment celle-ci est aujourd'hui évaluée et communiquée à l'utilisateur final.

1.2.2.1 La notion de qualité

La qualité, d'une manière générale, est définie par l'ISO comme *l'ensemble des propriétés et caractéristiques d'un produit ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites* (ISO8402, 1994). Dans le domaine des données géographiques, deux concepts de la qualité sont différenciés : la qualité interne et la qualité externe (cf. figure 1.21).

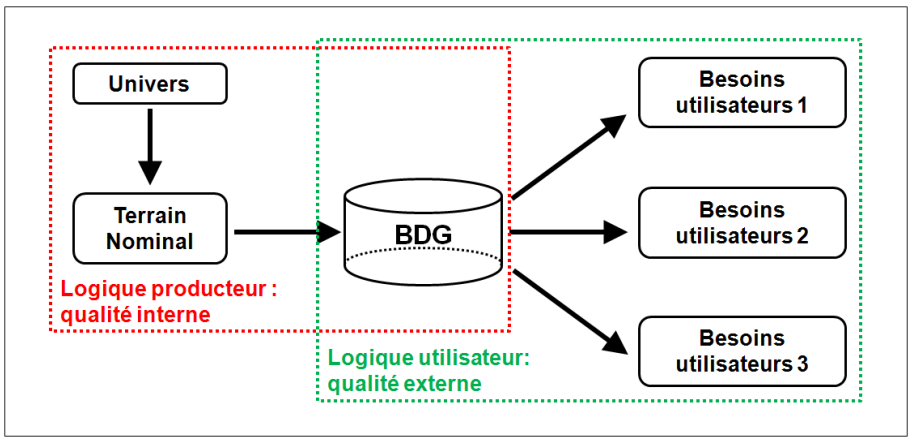


Figure 1.21 – Les deux notions de la qualité des données géographiques

La **qualité interne**, est définie comme *l'ensemble des propriétés et caractéristiques d'un produit ou service qui lui confère l'aptitude à satisfaire aux spécifications de contenu de ce produit ou service* (Vauglin, 1997). La qualité interne correspond donc au niveau de similitude existant entre les données produites et les données "parfaites" qui auraient dues être produites, c'est à dire le terrain nominal (Devillers, 2005). On peut ainsi considérer que le concept de qualité interne s'oriente plus comme une approche de la qualité orientée producteur.

La **qualité externe**, quant à elle, correspond *au niveau d'adéquation entre un produit et les besoins de l'utilisateur* (Devillers, 2005). Ce concept de la qualité correspond à l'expression anglaise "*fitness for use*" (Chrisman, 1984). C'est une approche de la qualité davantage orientée utilisateur, qui est cependant très difficile à évaluer puisqu'elle suppose que la qualité varie selon les attentes des utilisateurs, qu'il faudrait savoir formaliser.

Bien que la définition de la qualité d'un produit proposée par ISO8402 (1994) s'inscrit dans une optique qualité externe de la notion de qualité des données géographiques, nous verrons par la suite que la grande majorité des éléments d'évaluation de la qualité proposés s'orientent davantage dans une optique qualité interne.

1.2.2.2 Eléments d'évaluation de la qualité

Pour évaluer la qualité des données géographiques, différents éléments sont proposés dans la littérature. Ces éléments d'évaluation varient cependant en fonction des auteurs. Par exemple, Vauglin (1997) et David et Fasquel (1997) proposent les six éléments suivants :

- **Généalogie** : Description de l'histoire des données. Elle fournit une description des sources et méthodes d'acquisition des données, des opérations appliquées sur ces données et des organismes responsables.
- **Cohérence logique** : Degré de cohérence interne des données selon les règles de modélisation et les règles inhérentes aux spécifications du jeu de données.
- **Précision géométrique** : Estimation de la fluctuation des écarts entre les positions nominales (positions dans le terrain nominal) et les positions contenues dans le jeu de données.
- **Précision sémantique** : Différence entre les valeurs des éléments du jeu de données et les valeurs de leurs homologues dans le terrain nominal. Elle porte sur la classification des objets, la codification des attributs et les relations entre les objets.
- **Exhaustivité** : Conformité de la présence ou de l'absence des éléments du jeu de données par rapport au terrain nominal. Elle s'attache à des objets, des attributs ou des relations.
- **Actualité** : Décalage entre un jeu de données et le terrain nominal à une date T. Elle décrit en quelque sorte la " fraîcheur " des données.

D'autres auteurs, comme par exemple Guptill et Morrisson (1995) différencient la précision sémantique de la précision attributaire. Egalement, Aronoff (1989) et Kresse et Fadaie (2004) proposent un élément "Usage" permettant de qualifier à quelles utilisations la base de données est censée répondre.

Element	**Aronoff** (1989)	**USA-SDTS** (1992)	**ICA** (1995)	**CEN TC287** (1998)	**ISO TC211** (2002)
1. Lineage	S	S	S	S	S
2. Positional accuracy	S	S	S	S	S
3. Attribute accuracy	S	S	S	I	S
4. Logical consistency	S	S	S	S	S
5. Completeness	S	S	S	S	S
6. Semantic accuracy			S	S	
7. Usage, purpose, constraints	S	M		S	M
8. Temporal quality	S	M	S	S	S
9. Variation in quality		I	I	S	I
10. Meta-quality		I	I	S	I
11. Resolution	S	I	I	I	M

Table 1.3 – Eléments d'évaluation de la qualité des données géographiques proposés dans cinq sources, d'après Van Oort (2005)

Van Oort (2005) a réalisé une comparaison (cf. table 1.3) de l'ensemble des éléments d'évaluation proposés par différents auteurs. Ces éléments sont classés comme étant explicitement reconnus dans la section des métadonnées sur la qualité des données géographiques (S) ou dans une autre section des métadonnées (M), ou implicitement reconnus comme un élément (I).

Comme Van Oort (2005) le montre dans le tableau de la figure 1.3, il n'y a pas d'unicité chez les différents auteurs pour évaluer la qualité des données géographiques. Par exemple, les six éléments proposés par Vauglin (1997) et David et Fasquel (1997) sont présents de manière explicite ou implicite dans la majorité des cinq sources. Cependant, et c'est ce qui nous concerne davantage dans le cadre de ce travail de thèse, Van Oort (2005) ne parle pas de précision géométrique, mais d'exactitude de positionnement, ce qui, en référence aux termes définis précédemment, est très différent.

Nous verrons qu'en effet, l'évaluation de la qualité géométrique des bases de données géographiques vectorielles se focalise sur le positionnement des objets géographiques. Ceci est particulièrement vérifiable dans le cadre de la normalisation de l'ISO.

1.2.2.3 Normalisation de l'évaluation de la qualité

Afin de communiquer les informations sur la qualité des données géographiques à l'utilisateur final, une normalisation de l'ISO a été établie. Le Comité Technique 211 (ISO TC 211) a mis en œuvre des normes internationales et des spécifications techniques sur l'information géographique et plus particulièrement les trois suivantes, qui concernent spécifiquement la qualité des données géographiques :
– ISO19113 (2002) : Principes de qualité (*Quality Principles*)
– ISO19114 (2003) : Procédures d'évaluation qualité (*Quality Evaluation Procdeures*)
– ISOTS19138 (2006) : Mesures de qualité des données (*Data Quality Measures*)

Eléments d'évaluation Les différents éléments d'évaluation de la qualité des données et les mesures qui y sont associées sont intégrés dans la norme ISO19115 (2003), qui concerne les métadonnées, sous forme d'un package optionnel "Qualité des données" (*Data Quality*). Cinq éléments d'évaluation de la qualité des données géographiques y sont différenciés : complétude, cohérence logique, exactitude temporelle, exactitude thématique et exactitude de position (cf. table 1.4).

Element	**Subelement**
Completeness	commission omission
Logical consistency	conceptual consistency domain consistency format consistency topological consistency
Positional accuracy	absolute or external accuracy relative or internal accuracy gridded data position accuracy
Temporal accuracy	accuracy of time measurement temporal consistency temporal validity
Thematic accuracy	classification correctness non-quantitative attribute correctness quantitative attribute accuracy

Table 1.4 – Eléments et sous-éléments d'évaluation de la qualité de la norme ISO 19113

Chacun de ces éléments est divisé en sous-éléments, comme défini dans la norme ISO19113 (2002). Par exemple, pour l'exactitude de position, trois sous-éléments sont différenciés : exactitude absolue, exactitude relative, ainsi qu'un sous-élément relatif à l'exactitude de position des données en mode grille (cf. table 1.4).

Bien que l'ensemble des éléments d'évaluation de la qualité des données soient décrits dans la norme ISO19113 (2002), c'est dans la norme ISO19115 (2003) qu'ils sont formalisés en UML. Ceci suppose, comme l'explique Servigne *et al.* (2005), une certaine forme de prise de possession du domaine de la qualité par les experts en métadonnées.

Procédures d'évaluation Les procédures d'évaluation préconisées par l'ISO sont décrites dans la norme ISO19114 (2003). Les étapes pour évaluer et renseigner la qualité d'une base de données dans les métadonnées y sont présentées.
La norme ISO19114 (2003) différencie les méthodes d'évaluation en deux classes : les méthodes directes et indirectes. Les méthodes directes sont celles nécessitant une comparaison avec une base de données de référence, tandis que les méthodes indirectes évaluent la qualité à partir de la base de données.
Traditionnellement, pour les mesures d'évaluation de l'exactitude de position, des méthodes directes sont appliquées, en utilisant un jeu de données de référence, censé représenter au plus proche le terrain nominal, c'est-à-dire la base de données parfaite. La notion de jeu de données de référence sera exposée dans la section 1.3.1.1.

Mesures d'évaluation De nombreuses mesures d'évaluation sont proposées dans le cadre de la spécification technique ISOTS19138 (2006). Ces mesures sont subdivisées par éléments, puis par sous-éléments. Pour l'évaluation de l'exactitude de position, trois sous-éléments sont définis :
- les mesures d'exactitude absolue (ou externe) qui évaluent la proximité des valeurs de coordonnées des objets avec des valeurs acceptées comme représentant la réalité,
- les mesures d'exactitude relative (ou interne) qui évaluent la proximité des positions relatives des objets dans une base de données avec les positions relatives respectives d'objets représentant la réalité,
- les mesures d'exactitude de la position des données raster (que nous n'abordons pas ici).

La spécification technique ISOTS19138 (2006) propose principalement des mesures d'exactitude absolue (ou externe) pour évaluer l'exactitude de position des objets vectoriels. De nombreux indicateurs y sont proposés. Ceux-ci seront exposés ultérieurement, dans la section traitant des mesures d'imprécision géométrique. Ce qu'il est possible de constater, c'est que les indicateurs proposés ne concernent que des informations relatives à l'exactitude de position, mais en aucun cas à l'impact sur les mesures de longueur ou de surface calculées à partir de la géométrie des objets vectoriels.

Toutefois les mesures proposées dans le cadre de la spécification technique ISOTS19138 (2006) n'ont en aucun cas un caractère obligatoire. Cette spécification technique est une proposition pour cinq ans. L'ensemble des normes relatives à la qualité des données géographiques font actuellement l'objet d'une étude afin de proposer une norme globale les regroupant, la norme ISODIS19157 (2012).

1.2.2.4 Communication de l'information qualité

Différents moyens de communication de la qualité des données géographiques peuvent être utilisés afin d'informer un utilisateur, et ainsi aider celui-ci à identifier si une base de données correspond à ses besoins.

Métadonnées Le moyen le plus utilisé actuellement demeure la diffusion de métadonnées, à travers les éléments proposés dans la norme ISO19113 (2002). Bien qu'utiles et nécessaires, beaucoup d'auteurs s'interrogent sur la capacité des métadonnées à véritablement informer l'utilisateur en termes de qualité externe des données géographiques (Boin et Hunter, 2008). Nous verrons effectivement par la suite qu'en plus de leur caractère optionnel dans les métadonnées, les mesures de qualité préconisées dans la spécification technique ISOTS19138 (2006) peuvent s'avérer complexes à interpréter pour l'utilisateur final.

Ainsi, d'autres méthodes de communication de la qualité des données géographiques ont été proposées ces dernières années, principalement dans le domaine de la recherche.

Faïz (1999) et Qiu et Hunter (2002) ont par exemple proposé des modèles permettant de gérer les informations qualité à différents niveaux de détail, considérant que l'information qualité proposée dans les métadonnées peut porter une dimension trop générale, ne tenant pas compte de l'hétérogénéité interne des données décrites.

Visualisation de la qualité Différents travaux ont également été menés dans le domaine de la visualisation de l'information qualité, en particulier ceux de Beard et Buttenfield (1991), qui ont proposé des méthodes permettant de représenter les éléments d'évaluation de la qualité des données géographiques avec différentes variables visuelles.
Dans le même domaine, Hope et Hunter (2007) ont étudié l'impact de la représentation de l'information qualité (en l'occurrence une information d'imprécision géométrique) sur un processus de prise de décision, montrant que la représentation choisie influence directement la décision finale.

Logiciels et outils D'autres travaux ont visé ces dernières années à développer des outils SIG intégrant la notion d'erreur dans l'utilisation des données, comme par exemple les concepts d'*error-sensitive GIS* (Unwin, 1995) ou d'*error-aware GIS* (Duckham et McCreadie, 2002).
Devillers (2004) a proposé quant à lui un modèle multidimensionnel d'information sur la qualité des données géographiques permettant d'enrichir d'indicateurs supplémentaires les informations qualité proposées dans les métadonnées, et de les communiquer sous forme de tableaux de bord à l'utilisateur.
He (2007) a également proposé un outil SIG, permettant à un utilisateur de mettre en œuvre des modèles d'erreurs développés en recherche ces dernières années.

Enfin, Brown et Heuvelink (2007) ont proposé un environnement logiciel, le *Data Uncertainty Engine*, permettant d'intégrer et de propager l'incertitude dans les modèles environnementaux. Williams *et al.* (2008) proposent également un schéma XML, baptisé *UncertML*, permettant de décrire l'incertitude et sa propagation dans un soucis d'interopérabilité entre données et modèles.

Ces nombreux initiatives montrent bien à quel point la communication d'une information sur la qualité des données géographiques constitue un point fondamental dans toute utilisation de données géographiques, mais que les moyens actuels ne garantissent pas toujours sa bonne compréhension pour l'utilisateur.

La section suivante va nous permettre de nous focaliser sur la qualification de l'imprécision géométrique des objets vectoriels, en présentant les mesures les plus couramment rencontrées ainsi que les modèles d'évaluation développés ces dernières années en recherche.

Synthèse Section Cette section a tout d'abord permis de présenter le concept de qualité des données géographiques, ainsi que les éléments permettant de l'évaluer. Nous avons vu que ces éléments d'évaluation ont varié selon les auteurs au cours des dernières années, et ce malgré la normalisation proposée par l'ISO. Nous avons surtout pu constater que ces éléments d'évaluation s'orientent davantage dans une logique de qualité interne, c'est à dire d'adéquation aux spécifications, comme l'illustrent les éléments d'évaluation proposés dans les normes de l'ISO. Afin de rendre l'information qualité plus compréhensible pour l'utilisateur final, nous avons finalement montré que de nombreuses propositions ont été réalisées ces dernières années en recherche, afin de trouver d'autres moyens de communication des informations relatives à la qualité des données géographiques.

Conclusion de la section

Nous avons vu dans cette section que la notion d'imperfection des bases de données géographiques sous-entend un ensemble de notions plus spécifiques, abordées ces dernières années par de nombreux auteurs, sans que ne se dégage de véritable consensus terminologique. Dans le cadre de ce travail de thèse, nous avons retenu la classification de l'imperfection proposée par Bouchon-Meunier (1995), en nous focalisant plus particulièrement sur la notion d'imprécision. Nous avons montré que chacune des notions proposées permet d'exprimer spécifiquement l'imperfection concernée, ce qui peut s'avérer utile dans le cadre d'un processus d'aide à la décision.

En effet, tout processus de prise de décision sollicitant l'information géographique est confronté à l'imperfection des données, pouvant rendre la décision prise incertaine. Afin de réduire cette incertitude, une évaluation de la qualité des données géographiques doit donc être entreprise.

Après avoir présenté le concept de qualité des données géographiques, différents éléments proposés dans la littérature ont été identifiés afin de faciliter son évaluation. Au niveau géométrique, nous avons notamment montré que les éléments proposés dans le cadre de la normalisation internationale préconisée par l'ISO s'attachent principalement à décrire l'exactitude de position. Egalement, nous avons vu que de nombreux moyens de communication de l'information sur la qualité des données géographiques ont été proposés ces dernières années en recherche, afin de faciliter sa compréhension par l'utilisateur final.
La section suivante nous permettra de réaliser un état de l'art sur les mesures et les modèles permettant d'évaluer l'imprécision géométrique des données géographiques.

1.3 Evaluation de l'imprécision géométrique des données géographiques

Cette section a pour objectif de présenter comment aujourd'hui est évaluée l'imprécision géométrique des données géographiques vectorielles, et comment celle-ci est communiquée à l'utilisateur. Un état de l'art des mesures et modèles développés ces dernières années sera donc proposé. Nous verrons qu'ils portent davantage sur l'estimation de l'imprécision géométrique en termes de positionnement, plutôt que sur l'impact sur les mesures géométriques de longueur et de surface.

1.3.1 Mesures d'évaluation de l'imprécision géométrique

Cette section a pour objectif de présenter comment peut être mesurée l'imprécision géométrique des objets vectoriels dans une base de données géographiques. Nous allons voir que la majorité des mesures proposées ont principalement pour objectif de quantifier les écarts de position entre primitives graphiques. Cette section présentera des mesures utilisées en recherche ou en production de données, mais également les mesures proposées dans le cadre de la normalisation ISO ou encore dans les textes règlementaires actuellement en application en France.

1.3.1.1 Mesures sur les primitives géométriques

Les mesures présentées dans cette section sont classées en fonction des primitives géométriques qu'elles qualifient, c'est à dire les primitives ponctuelles, linéaires et surfaciques. Ces mesures sont majoritairement basées sur des comparaisons avec des données de référence.

La notion de référence Afin d'évaluer la précision géométrique d'une base de données géographiques, on utilise généralement des données de référence, appelée également "Données de contrôle" (David et Fasquel, 1997). Ces données de contrôle sont définies comme *l'ensemble de données qui permettent de mesurer ou de contrôler la qualité d'un jeu de données par comparaison avec un échantillon de ce jeu de données* (David et Fasquel, 1997).

Les données de référence, qui sont donc utilisées pour réaliser des comparaisons, ont pour objectif de représenter le terrain nominal. En effet, le terrain nominal est une notion abstraite, commode à présenter de façon théorique mais qui n'est pas accessible dans la réalité. Les données de référence permettent donc de fournir une estimation "proche" (une approximation) du terrain nominal.

Pour pouvoir réaliser des comparaisons entre la base de données à évaluer et les données de référence, un échantillon d'objets représentatifs doit être extrait. Chaque objet de l'échantillon doit ensuite être mis en correspondance avec son homologue présent dans les données de référence. Cette phase dite d'appariement de données, a fait l'objet de nombreux travaux de recherche pour faciliter son automatisation (Mustière et Devogèle, 2008; Olteanu, 2008). Une fois les objets appariés, les mesures décrites ci-dessous peuvent être réalisées.

a. Primitives ponctuelles

Pour les objets représentés par des primitives ponctuelles, la distance permettant de mesurer des écarts de position la plus souvent utilisée est la **distance euclidienne**.

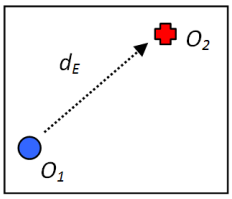


Figure 1.22 – Distance euclidienne entre deux points

Ainsi, pour deux objets O_1 et O_2 de coordonnées géographiques (x_1, y_1) et (x_2, y_2), comme illustré en figure 1.22, la distance euclidienne d_E en deux dimensions est calculée par application de l'équation 1.6.

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1.6}$$

b. Primitives linéaires

Contrairement aux objets ponctuels, les objets linéaires sont plus délicats à évaluer et plusieurs distances peuvent être utilisées. Trois distances sont ici présentées : la distance de Hausdorff, la distance de Fréchet et la distance moyenne.

Distance de Hausdorff La distance de Hausdorff (du nom du mathématicien allemand Felix Hausdorff) permet de mesurer l'écart maximal entre deux polylignes (Abbas, 1994).

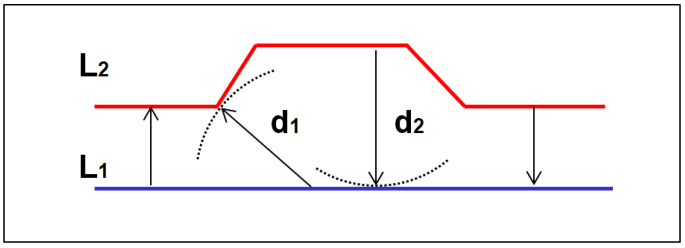


Figure 1.23 – Distance de Hausdorff entre deux polylignes

Ainsi, pour deux objets géographiques représentés par les polylignes L_1 et L_2, comme illustré dans la figure 1.23, la distance de Hausdorff d_H entre les polylignes L_1 et L_2 est définie comme suit :

$$d_H = max(d_1, d_2) \tag{1.7}$$

où d_1 représente le maximum des plus courtes distances des points de L_1 à L_2 et d_2 représente le maximum des plus courtes distances des points de L_2 à L_1.

La distance de Hausdorff est utilisée dans de nombreuses applications, qui ne s'arrêtent pas à l'étude des primitives linéaires, mais aussi à l'étude des primitives surfaciques. Elle est par exemple utilisée dans différents algorithmes d'appariement de données (Mustière et Devogèle, 2008). C'est une mesure intéressante pour quantifier les écarts de position, mais elle présente cependant un certain nombre de limites.

Tout d'abord, lorsque les deux polylignes à comparer présentent des longueurs différentes, la distance de Hausdorff va principalement porter sur les extrémités. Olteanu (2008) propose dans ce cas de n'utiliser que la première composante de la distance de Hausdorff.
De plus, la distance de Hausdorff n'est pas forcément adaptée pour étudier des polylignes présentant une forte sinuosité. Il peut arriver que cette distance soit faible alors que les deux polylignes ne se ressemblent pas du tout, comme illustré dans la figure 1.24. En effet, la distance de Hausdorff ne prend pas un compte l'ordre des sommets dans une polyligne, et ne se contente que de calculer la distance entre les sommets les plus proches. Dans ce cas, il est préférable d'utiliser d'autres distances, comme la distance de Fréchet.

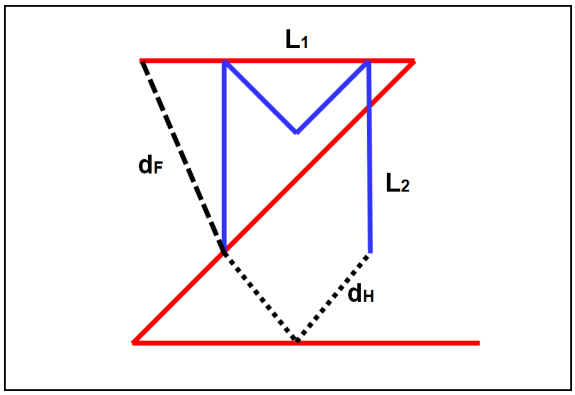


Figure 1.24 – Distances de Hausdorff et de Fréchet, d'après Badard et Lemarié (2002)

Distance de Fréchet Pour comprendre le principe de la **distance de Fréchet**, Devogèle (1997) cite l'exemple d'un maître avec son chien se déplaçant chacun le long d'une ligne : "*Ils avancent et ils arrêtent indépendamment à volonté. La distance de Fréchet entre les deux lignes est la longueur minimale de la laisse qui permet une progression simultanée*".

La distance de Fréchet d_F (Alt et Godau, 1995) est en fait basée sur la propriété que toute polyligne orientée est équivalente à une fonction continue. Elle se définit comme suit :

Soit deux polylignes f : [0, N] → V et g : [0, M] → V' et une distance d (une distance euclidienne par exemple), la distance de Fréchet d_F est définie par :

$$d_F(f,g) = \min_{\alpha:[0,1]\to[0,N];\beta:[0,1]\to[0,M]} \{\max_{t\in[0,1]}[d(f(\alpha(t)), g(\beta(t)))]\} \quad (1.8)$$

où N, M $\in \Re$ représentent le nombre de segments composant les polylignes f et g, et où V et V' sont des espaces vectoriels.
$\alpha(t)$ et $\beta(t)$ sont des fonctions continues et croissantes avec le temps et $\alpha(0) = 0$, $\beta(0) = 0$ et $\alpha(1) = N$ et $\beta(1) = M$

Le problème de la distance de Fréchet est qu'elle impose que les polylignes à comparer soient orientées dans la même direction, ce qui n'est pas évident à définir lorsque les géométries des deux polylignes sont très différentes.

Distance moyenne Une autre distance permettant de qualifier les écarts de position entre deux polylignes est la distance moyenne. La distance moyenne entre deux polylignes a été introduite par McMaster (1986) afin de comparer les effets de la généralisation sur une polyligne avec la polyligne d'origine. Cette distance permet de mesurer un écart défini entre la surface S séparant les deux polylignes L_1 et L_2 avec la moyenne de leurs longueurs (cf. figure 1.25).

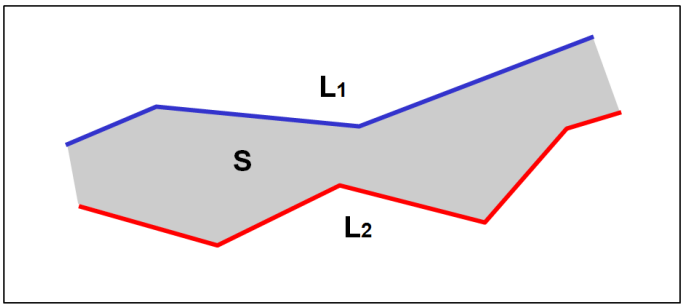


Figure 1.25 – Principe de calcul de la distance moyenne

L'équation de la distance moyenne d_M entre les polylignes L_1 et L_2 est définie comme suit :

$$d_M = \frac{S}{\frac{L_1+L_2}{2}} \tag{1.9}$$

La distance moyenne est très facile à mettre en œuvre pour comparer l'écart de position entre deux polylignes. Cependant, comme préconisé par de nombreux auteurs (Olteanu, 2008; Devogèle, 1997), il est préférable d'utiliser cette distance avec d'autres distances, comme celles de Hausdorff ou de Fréchet. En effet, en référence aux statistiques descriptives, une moyenne et un écart-type des écarts constitueraient la meilleure solution pour décrire l'écart entre deux polylignes.

Des indicateurs de forme peuvent également être utilisés pour caractériser les polylignes, comme par exemple les mesures de sinuosité proposées par Plazanet (1996). Ces indicateurs seront abordés ultérieurement au cours de cette thèse.

c. Primitives surfaciques

La comparaison des primitives surfaciques nécessite à la fois une comparaison des positions, mais également une comparaison des formes. En effet, les formes des objets surfaciques doivent pouvoir être comparées afin d'identifier les points homologues représentant les mêmes lieux d'une entité géographique. Les distances utilisées pour les polylignes peuvent également être appliquées aux primitives surfaciques. Nous allons cependant présenter quelques mesures spécifiques aux objets surfaciques, comme la distance surfacique.

Distance surfacique Pour caractériser les objets géographiques surfaciques, une distance souvent utilisée est la distance surfacique, initiée par Vauglin (1997), et illustrée dans la figure 1.26.

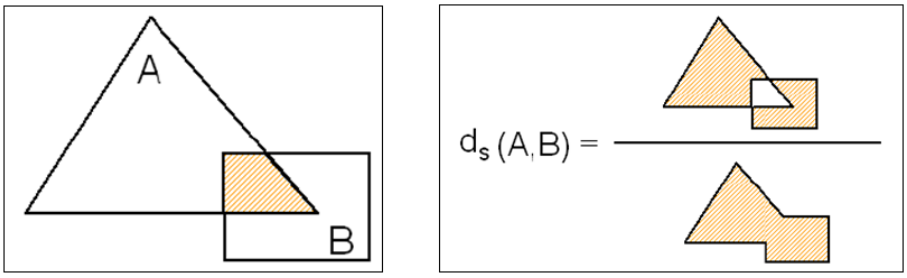


Figure 1.26 – Principe de calcul de la distance surfacique

La distance surfacique d_S entre les polygones A et B est définie comme suit :

$$d_S = 1 - \frac{S(A \cap B)}{S(A \cup B)} \tag{1.10}$$

La distance surfacique est définie dans l'intervalle [0, 1], c'est à dire que si la distance est égale à 0, les deux polygones sont égaux, si la distance est égale à 1, les deux polygones sont disjoints. Cette distance permet de calculer un taux de recouvrement entre deux polygones homologues, mais n'informe en aucun cas de la précision géométrique des objets comparés.

Indicateurs de formes Afin de qualifier les écarts de forme des objets surfaciques, de nombreuses mesures ont été créées. On peut par exemple citer la **fonction à distance radiale**, qui permet de mesurer la distance séparant le centre de masse du polygone aux points composant son contour, dans le sens trigonométrique (Cohen et Guibas, 1997), comme illustré dans la figure 1.27.

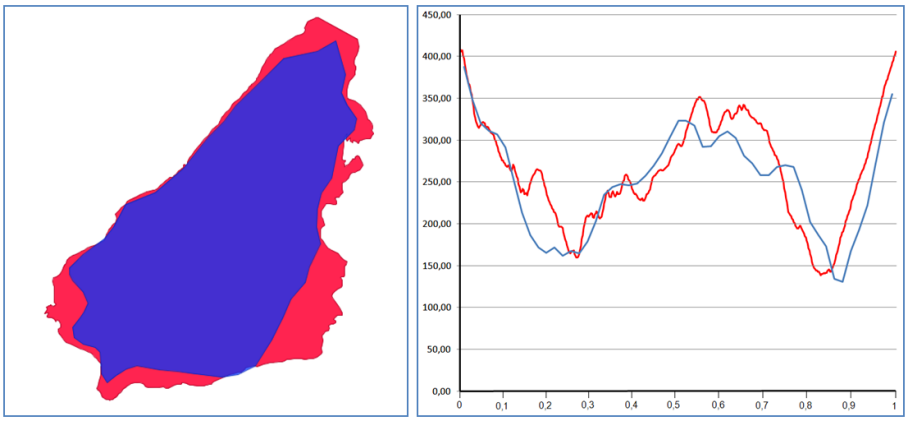


Figure 1.27 – Calcul de la distance radiale entre deux objets surfaciques

Egalement, la **fonction angulaire** (ou fonction Turning) permet de décrire un objet surfacique par les mesures des angles formés par les segments composant son contour et une demi-droite horizontale orientée selon l'axe des abscisses (Arkin *et al.*, 1991), comme illustré dans la figure 1.28.

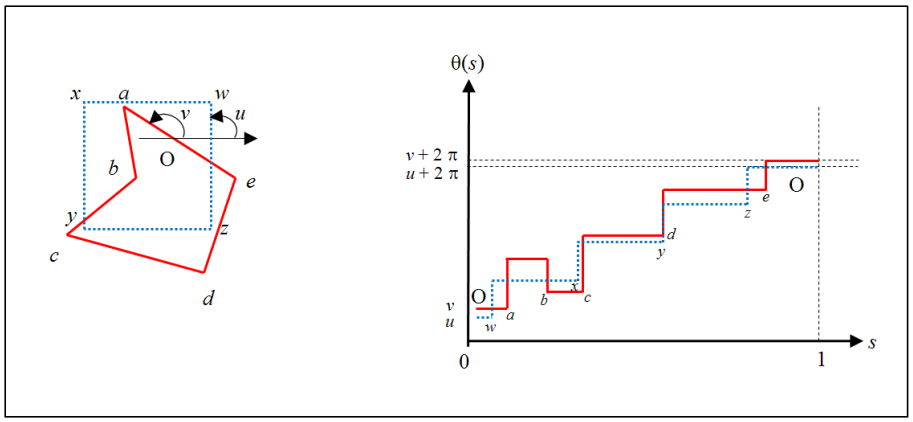


Figure 1.28 – Calcul de la fonction angulaire de deux objets surfaciques, d'après Olteanu (2008)

Enfin, d'autres mesures, comme les **mesures de compacité** ou **de circularité** ont été développées pour comparer la forme des objets surfaciques (Vauglin, 1997; Bel Hadj Ali, 2001).

1.3.1.2 Mesures proposées dans le cadre règlementaire

Cette section a pour objectif de proposer une synthèse des mesures proposées dans le cadre de la normalisation de l'ISO (par la spécification technique ISO 19138), ainsi que dans le cadre règlementaire français (par l'Arrêté du 16 septembre 2003).

Les mesures de la spécification technique ISO/TS 19138

Contexte La spécification technique ISOTS19138 (2006) propose un ensemble de mesures permettant d'évaluer la qualité des bases de données géographiques. Ces mesures sont subdivisées par éléments, puis par sous-éléments. Nous ne nous intéressons qu'aux mesures liées à l'élément "Exactitude de Position" (*Positionnal Accuracy*).

Les mesures d'exactitude de position sont subdivisées en mesures d'exactitude absolue (ou externe) et en mesures d'exactitude relative (ou interne). Malgré cette séparation, très peu de mesures d'exactitude relative sont proposées au sein de la spécification technique ISO 19138.

La table 1.5 fait l'inventaire des mesures relatives aux données vectorielles proposées dans la spécification technique ISO 19138. Ces mesures seront intégrées, dans leur majorité, dans la norme ISO 19157 (ISODIS19157, 2012) sur l'évaluation de la qualité des données géographiques. La description de l'ensemble des mesures est proposée dans l'Annexe A.

Absolute or external accuracy	
General measures for positional uncertainties	Mean value of positional uncertainties Mean value of positional uncertainties excluding outliers Number of positional uncertainties above a given threshold Rate of positional errors above a given threshold Covariance matrix
Vertical positional uncertainties	Linear error probable Standard linear error Linear map accuracy at 90% significance level Linear map accuracy at 95% significance level Linear map accuracy at 99% significance level Near certainty linear error Root mean square error Absolute linear error at 90% significance level biased vertical data
Horizontal positional uncertainties	Circular standard deviation Circular error probable Circular map accuracy standard Circular error at 95% significance level Circular near certainty error Root mean square error of planimetry Absolute circular error at 90% significance level biased data Uncertainty ellipse Confidence ellipse
Relative or internal accuracy	
	Relative vertical error Relative horizontal error

Table 1.5 – Mesures d'exactitude absolue et relative proposées dans la spécification technique ISO 19138 pour l'élément "Exactitude de Position"

Synthèse La description des indicateurs proposés dans le cadre de la spécification technique ISO 19138 confirme que pour qualifier l'imprécision géométrique, seules des mesures d'exactitude de position ont été sélectionnées par les experts. Ce que nous pouvons tout d'abord remarquer, c'est que l'ensemble des mesures proposées sont basées sur un contrôle ponctuel, même si les objets géographiques représentés sont linéaires ou surfaciques. Ainsi, aucunes distances, comme celles présentées dans la section précédente, ne sont utilisées dans la spécification technique ISO 19138.

Egalement, si certaines de ces mesures sont relativement aisées à produire (comme l'Erreur Moyenne Quadratique par exemple) et peuvent avoir une utilité pour l'utilisateur, nous pouvons constater que dans leur grande majorité, ces mesures demeurent assez complexes à produire (la Matrice de Variance-Covariance par exemple), et sont surtout destinées à un public d'experts, comme des producteurs de données par exemple.
Cette situation s'illustre par l'utilisation de nombreux indicateurs à finalité militaire, décrits dans l'accord de normalisation (STANAG, 2002) de l'agence de normalisation de l'OTAN.

Malgré cela, la section sur l'exactitude relative pourrait tout à fait convenir pour décrire l'imprécision des mesures géométriques de longueur, voire de surface.

En effet, il parait possible de proposer une mesure, basée sur des comparaisons de longueurs géométriques entre le jeu de données à évaluer et un jeu de données de référence. Après avoir identifié des couples de points homologues dans les deux jeux de données, la mesure consisterait simplement à calculer les erreurs de longueur géométrique entre les couples de points homologues. En renouvelant l'expérience sur un échantillon suffisamment important, il serait alors possible d'en déduire un écart-type, qui permettrait de qualifier l'imprécision des mesures de longueur, comme cela est déjà proposé dans les mesures existantes (uniquement basées sur des distances euclidiennes). Mais au delà de sa faisabilité, d'un point de vue purement terminologique, l'idée de classifier cette mesure au sein de l'élément "exactitude de position" peut cependant paraître discutable.

Les mesures utilisées dans l'Arrêté du 16 septembre 2003

Contexte La France s'est dotée en 2003 d'un nouveau moyen de validation des travaux topographiques réalisés par l'Etat, les collectivités locales et leurs établissement publics. Cet arrêté (qui remplace l'ancien arrêté de 1980) concerne les travaux relatifs aux points du canevas géodésique, aux objets géographiques relatifs à ce canevas, ainsi qu'aux images rectifiées. Cet arrêté repose sur la notion de "classes de précision" (notée $[xx]$) pour qualifier la précision géométrique des objets produits. Le principe d'évaluation de la précision géométrique, tel que proposé dans le cadre de l'Arrêté du 16 septembre 2003, est présenté dans l'Annexe B.

Synthèse L'arrêté de 2003 fait donc figure de texte règlementaire pour les travaux topographiques de l'Etat, les collectivités locales et leurs établissements publics. Il est intéressant de voir comment est mesurée et qualifiée l'imprécision géométrique des objets vectoriels d'une base de données géographiques à travers cet arrêté. Ce que nous pouvons remarquer, outre la complexité relative des critères permettant de définir les gabarits d'erreurs, c'est que l'ensemble des mesures est basé sur un contrôle ponctuel des objets vectoriels, même si les géométries sont linéaires ou surfaciques. Les points sont donc extraits des géométries, et mis en correspondance avec leurs homologues présents dans les données de référence. Ainsi, comme c'est le cas pour les mesures proposées dans le cadre de la spécification technique ISO 19138, aucune des méthodes de comparaison des primitives linéaires ou surfaciques présentées précédemment ne sont utilisées au niveau règlementaire en France.

Synthèse Section Nous avons vu dans cette section que de nombreuses mesures, spécifiques à chaque type de géométrie, sont proposées dans la littérature afin d'estimer l'imprécision géométrique des objets vectoriels. Nous avons cependant pu constater que les mesures proposées dans le cadre de la normalisation de l'évaluation de la qualité des données géographiques (spécification technique ISO 19138) ou encore au niveau règlementaire en France (Arrêté du 16 septembre 2003) ne proposent que des mesures basées sur un contrôle ponctuel, au delà de la complexité des indicateurs proposés. Mais surtout, ces mesures ne permettent que d'estimer l'imprécision géométrique en termes de positionnement, mais pas en termes d'imprécision des mesures géométriques.

1.3.2 Les modèles d'évaluation de l'imprécision géométrique

Depuis le début des années 80, la recherche sur la modélisation de l'imprécision géométrique des données géographiques a fait l'objet de très nombreuses contributions. Vauglin (1997) différencie deux grandes écoles de pensées, que Chrisman (1989) avait déjà identifiées : l'école des simulations et l'école des tests.

Ecole des simulations L'école des simulations (incarnée par Goodchild) préconise une évaluation de la précision géométrique et de ses conséquences par simulations statistiques. Elle a pour objectif de rechercher un modèle qui permette d'approcher le processus de génération des erreurs dans la manipulation des données géographiques. Le modèle stochastique utilisé doit ainsi permettre de produire des cartes aléatoires simulant la réalité. La théorie développée par cette école découle des statistiques mathématiques et de très nombreux auteurs participent à cette tendance.

Ecole des tests L'école des tests (incarnée par Chrisman) préconise, quant à elle une approche par comparaison avec des données réelles. Le principe de cette école consiste à la fois à tester la qualité géométrique d'objets géographiques en réalisant des superpositions et des comparaisons de plusieurs occurrences de la même information (Chrisman, 1991), afin d'en dégager des lois statistiques sur les imprécisions géométriques.

La section qui suit a pour objectif de présenter différents modèles d'évaluation des imprécisions géométriques développés en recherche, non pas par rapport à leur école de pensée d'appartenance, mais davantage par rapport à l'application considérée. Ainsi nous distinguerons dans cette section les modèles permettant la modélisation des imprécisions géométriques en termes de positionnement, puis en termes de mesures géométriques de longueur ou de surface.

1.3.2.1 Modèles relatifs au positionnement

Si l'imprécision sur la localisation des sommets constituant la géométrie d'un objet vectoriel est en général mesurée et estimée à l'aide de l'Erreur Moyenne Quadratique, de nombreux autres modèles ont été développés au cours des dernières années.

a. Approches probabilistes

Le modèle le plus utilisé, est celui de la **bande Epsilon**, proposé par Perkal (1954). Ce modèle considère que chaque point saisi peut se trouver équiprobablement n'importe où dans un disque de rayon epsilon centré sur ses coordonnées. La bande epsilon propose donc d'ajouter une valeur de tolérance pour chaque objet.

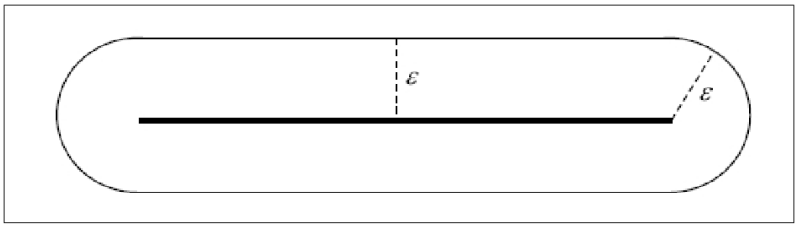


Figure 1.29 – Le modèle "Epsilon-Band"

Le modèle de la bande Epsilon a été introduit dans le domaine de l'information géographique par Chrisman (1982) pour les objets linéaires, sous l'appellation "*epsilon-band model*" (cf. figure 1.29). Le principal problème du modèle de la bande epsilon demeure dans le calibrage de la valeur epsilon.

Cependant ce modèle a inspiré de nombreux travaux, comme par exemple ceux de Goodchild et Hunter (1997) qui proposent un modèle permettant de comparer deux jeux de données linéaires (cf. figure 1.30), en étudiant le taux de recouvrement de la polyligne comparée par la bande epsilon autour de la polyligne de référence. Ce modèle a été prolongé par Heo *et al.* (2009) pour l'étude de la variation du trait de côte, ou encore par Haklay (2010) pour comparer des objets routiers.

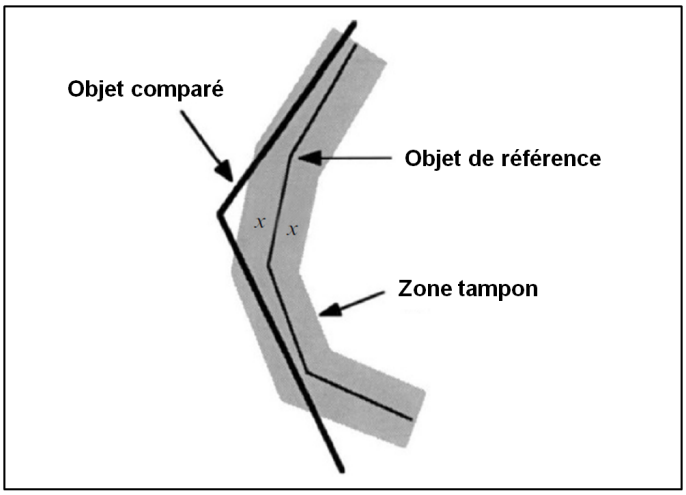


Figure 1.30 – Le modèle proposé par Goodchild et Hunter (1997)

De nombreuses variantes du modèle de la bande epsilon ont par la suite été proposées. Par exemple, Wolf et Ghilani (1997) proposent un modèle basé sur une distribution normale bivariée, où l'imprécision de positionnement d'un point prend la forme d'une ellipse (cf. figure 1.31), en considérant qu'il existe une corrélation d'erreur entre les coordonnées en x et en y.

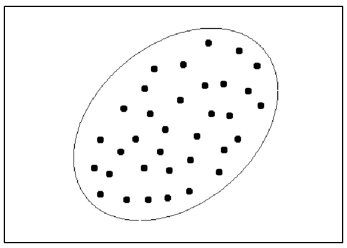


Figure 1.31 – Le modèle d'imprécision en forme d'ellipse avec un échantillon de points

Bolstad *et al.* (1990) proposent également un modèle d'erreur sur les données saisies manuellement, en utilisant une distribution normale en deux dimensions.

Caspary et Scheuring (1992) ont proposé un modèle dans la continuité de celui de la bande epsilon proposé par Chrisman (1982), mais au sein duquel la bande est resserrée dans la partie centrale de la ligne (cf. figure 1.32).

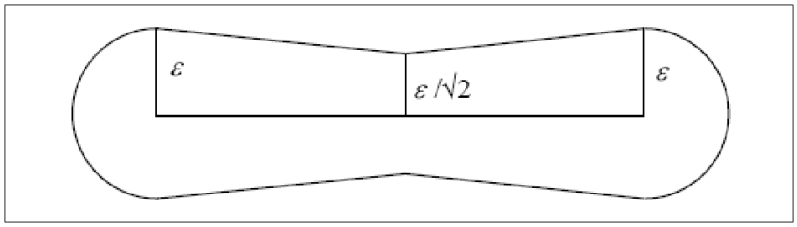


Figure 1.32 – Le modèle "Error Band"

Ce modèle, baptisé "*error-band model*" est basé sur le modèle de Dutton (1992). Dutton (1992) a proposé un modèle probabiliste de positionnement d'une polyligne, suite à une expérimentation basée sur la génération d'erreurs aléatoires aux extrémités d'une ligne (cf. figure 1.33).

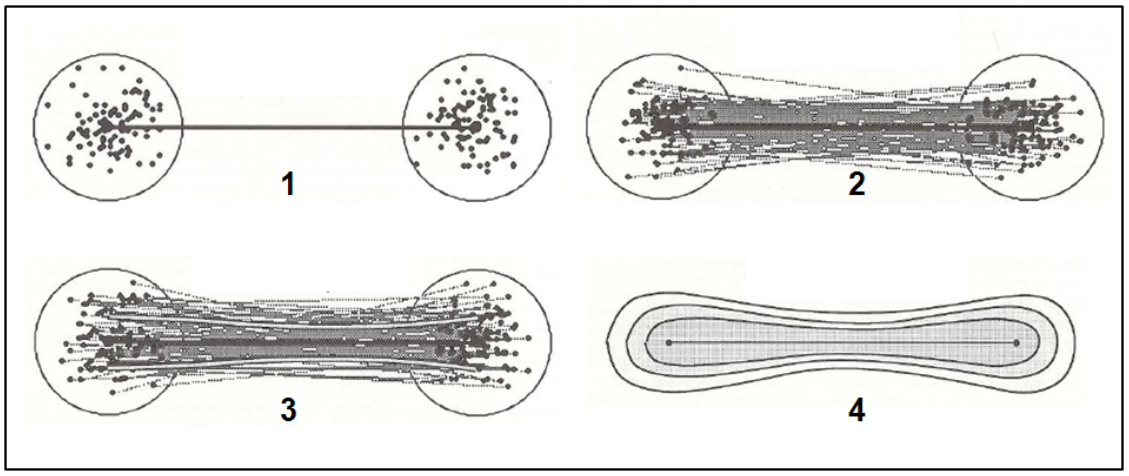


Figure 1.33 – Le modèle proposé par Dutton (1992)

Shi *et al.* (1999) ont prolongé ces précédentes contributions en proposant un modèle d'erreur probabiliste sur les segments de polylignes. Dans ce modèle, les erreurs de position des extrémités du segment suivent une distribution normale en deux dimensions, et ne sont pas corrélées. La probabilité des erreurs de position sur le segment de ligne peut ainsi être déduite en fonction de celle de ses extrémités (cf. figure 1.34).

Le modèle de Shi *et al.* (1999) a été prolongé par Shi et Liu (2000) en prenant en compte la corrélation entre les points successifs composant la polyligne, et en considérant que la probabilité de distribution des sommets localisés aux extrémités est identique. Ce modèle appelé "*G-Band model*", présente également une probabilité d'erreurs davantage resserrée au milieu de la ligne qu'à ses extrémités.

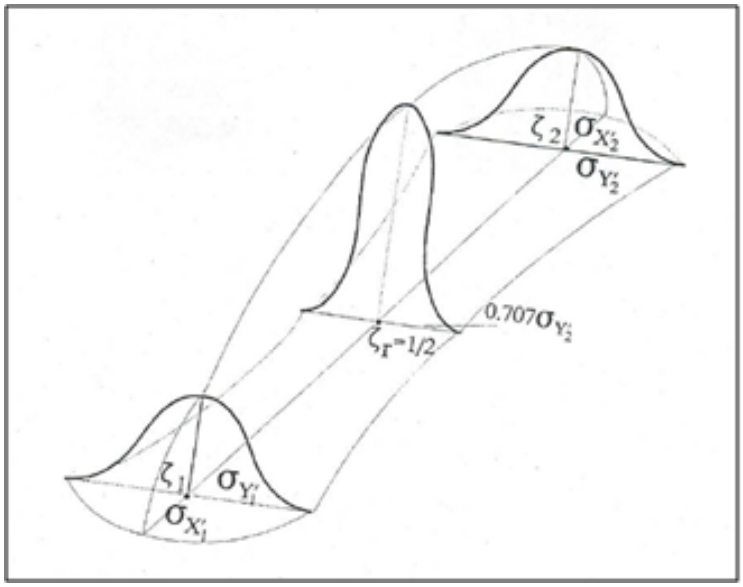


Figure 1.34 – Le modèle proposé par Shi *et al.* (1999)

Les résultats théoriques de ces modèles demeurent cependant criticables. Comme l'explique Vauglin (1997), ces modèles ont tendance à faire croire que les points intermédiaires des segments sont plus précis que les sommets. Or selon lui, la géométrie des objets vectoriels est construite à partir de ces sommets, sur lesquels l'opérateur de saisie porte une grande attention, alors que les segments ne sont que le résultat d'une interpolation entre ces sommets. On pourrait donc plutôt croire que l'imprécision de positionnement est plus importante entre les sommets, que sur les sommets à proprement parler.

Franck et Navratil (2011) résument très schématiquement cette idée par trois types de zones de tolérance (cf. figure 1.35) :
- les zones de tolérance inscrites dans une bande Epsilon (*epsilon band as a confidence region*),
- les zones de tolérance en forme d'os (*bone shaped confidence region*),
- les zones de tolérance en forme de double-nœud (*double bow confidence region*).

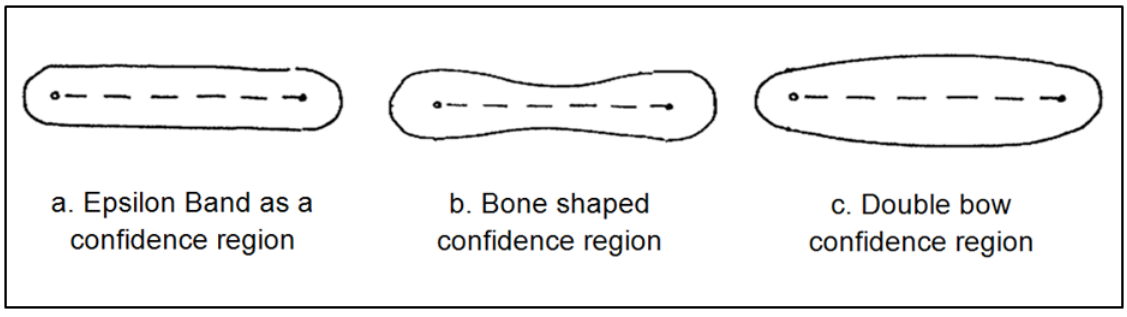


Figure 1.35 – Les trois types de modélisation des zones de tolérance, d'après Franck et Navratil (2011)

En référence à ce constat, Vauglin (1997) propose de ne conserver la bande epsilon que comme zone de tolérance mais considère cependant qu'il n'est pas possible de modéliser l'erreur géométrique par une simple loi uniforme dans les données géographiques, du fait des différents processus intervenus lors de la construction de la géométrie des objets.

Afin de modéliser l'erreur de positionnement des polylignes, Vauglin (1997) a proposé un modèle d'erreur nommé "*modèle GES*". Ce modèle suit une loi Gaussienne Exponentielle Symétrique (cf. figure 1.36), qui est en fait le mélange entre :
- une loi Gaussienne représentant la distribution des erreurs de pointé lors de l'acquisition de la géométrie de l'objet,
- une loi Exponentielle Symétrique représentant la distribution des erreurs de généralisation liées à la construction des segments de polylignes entre les sommets.

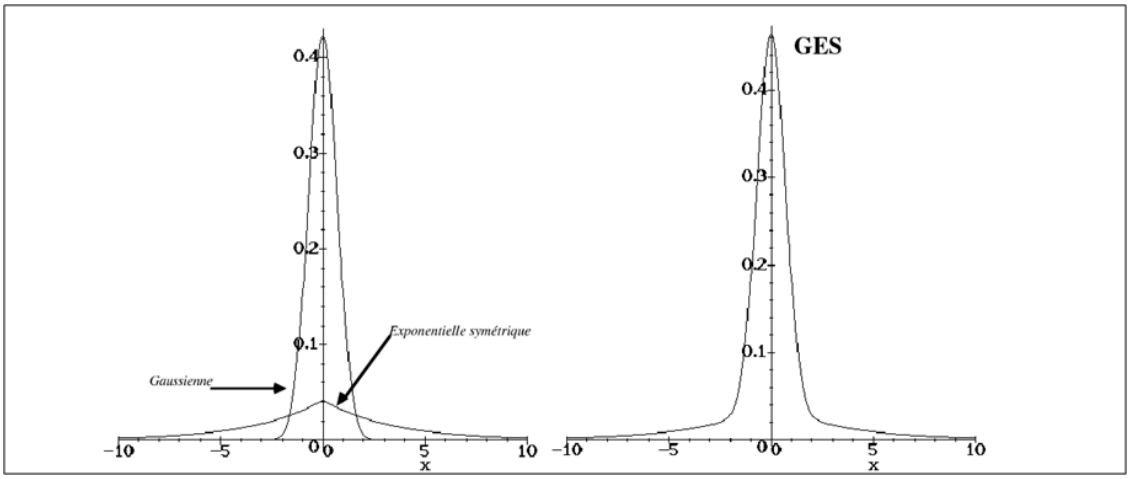


Figure 1.36 – Le modèle GES proposé par Vauglin (1997)

Leung et Yan (1998) ont également développé un modèle basé sur les points composant la géométrie des objets vectoriels, dénommé "*point-based error model*". Ce modèle est un modèle probabiliste qui peut prendre la forme d'une bande epsilon à partir d'un certain seuil. L'hypothèse de base est que l'erreur de localisation des points varie selon une distribution normale circulaire, et qu'elle se propage aux lignes et aux polygones. A partir de cette hypothèse, pour chaque couple de points (a, b), l'imprécision de localisation pour chaque point (x, y) est donnée par l'équation 1.11, où σ est un indicateur d'exactitude de position.

$$f(x,y) = \frac{1}{2\pi\sigma^2} exp\left[-\frac{(x-a)^2+(y-b)^2}{2\sigma^2}\right] \tag{1.11}$$

L'intérêt de ce modèle est qu'il peut être propagé d'un simple objet à l'ensemble des objets d'une base de données. Sa principale limite est qu'il ne prend pas en compte l'hétérogénéité en termes d'exactitude de localisation entre les différents objets de la base de données.

b. Approches par simulations

Face à ces approches analytiques, de nombreux auteurs privilégient des approches par simulations Monte-Carlo, les considérant plus robustes et surtout applicables dans un grand nombre d'applications.

Par exemple, Hunter et Goodchild (1996) ont proposé un modèle nommé "*Vector Uncertainty Model*" permettant de simuler des erreurs de positionnement sur chaque point composant la géométrie des objets, en suivant des ditributions normales indépendantes en x et y. A l'aide d'un grand nombre de réalisations d'objets simulés, ces méthodes permettent d'estimer l'imprécision en termes de positionnement, mais également pour d'autres opérations, comme les intersections entre objets, ou encore les mesures de longueur ou de surface.

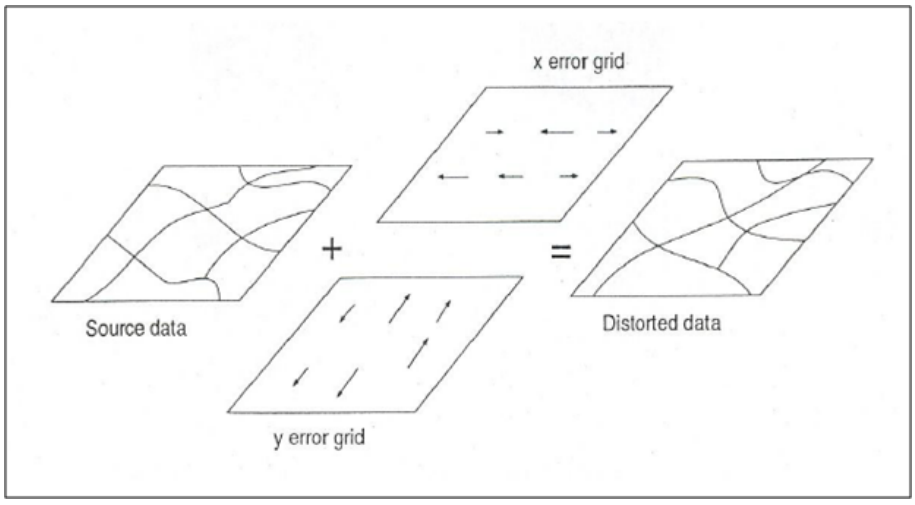


Figure 1.37 – Le modèle proposé par Hunter et Goodchild (1996)

Le principal défaut de ces approches par simulations réside dans la génération d'incohérences topologiques lors de l'introduction d'erreurs dans le positionnement des sommets de la géométrie des objets simulés, comme le souligne Bonin (2002). Il est donc nécessaire d'introduire des corrélations entre les sommets, afin d'obtenir des objets bruités plus réalistes. Egalement, le paramétrage du bruit sur les coordonnées des points composant la géométrie des objets doit être défini. L'Erreur Moyenne Quadratique est souvent utilisée dans cette optique (Hunter et Goodchild, 1996).

Ces modèles par simulations d'erreurs de positionnement ont été repris par un très grand nombre d'auteurs ces dernières années, comme par exemple De Bruin *et al.* (2008) qui utilisent des simulations afin de modéliser les erreurs de délimitation des parcelles agricoles saisies par relevés GPS. On peut noter que l'expérimentation de De Bruin *et al.* (2008) fut réalisée à l'aide du logiciel DUE (*Data Uncertainty Engine*), qui intègre un environnement de modélisation de l'imprécision de positionnement par simulations Monte-Carlo (Heuvelink *et al.*, 2007; Brown et Heuvelink, 2007).

Cet environnement propose d'ailleurs une approche intéressante, déjà évoquée par Harrie et Sarjakoski (2002), qui différencie les objets géographiques selon leur nature. Deux types d'objets sont ici distingués :
– les objets "déformables", comme une limite de forêt par exemple,
– les objets "rigides", comme un bâtiment par exemple.

Différentes méthodes de simulation sont proposées en fonction de la nature des objets, comme illustré dans la figure 1.38. Le logiciel DUE permet également de prendre en compte la corrélation des erreurs entre sommets successifs d'une géométrie lors de la réalisation de simulations.

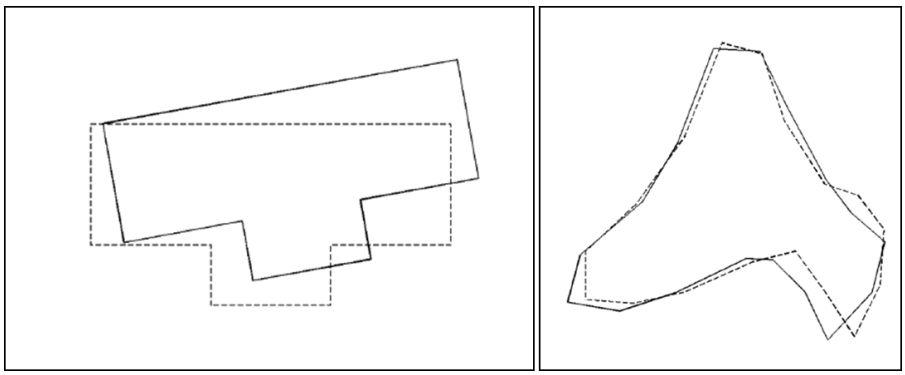


Figure 1.38 – Objets rigides et déformables, d'après Heuvelink *et al.* (2007)

Pour terminer cette section, nous pouvons noter le développement de modèles permettant d'étudier les relations entre les effets de la simplification des polylignes sur l'imprécision en termes de positionnement, comme ceux proposés par Veregin (2000) ou Cheung et Shi (2004).

c. Approches mixtes

Egalement, nous pouvons citer des contributions mixtes, permettant de modéliser l'imprécision sur les opérations d'intersection entre objets géographiques, comme par exemple entre points et polygones. Leung et Yan (1998) ont proposé un cadre général où les polygones et les points sont caractérisés comme étant précis, flous, ou aléatoires, pour ainsi déterminer neuf cas d'analyse d'intersections possibles. Egalement, Cheung *et al.* (2004) ont proposé un modèle, au sein duquel la localisation des polygones est considérée comme précise, et celle des points comme imprécise. Ce modèle combine à la fois approches par tests et par simulations.

Les différents modèles présentés dans cette partie ne constituent pas une revue de littérature exhaustive sur la modélisation de l'imprécision géométrique sur le positionnement des objets vectoriels, mais ils permettent de bien illustrer combien les contributions dans ce domaine sont nombreuses. Nous allons voir que c'est nettement moins le cas en ce qui concerne les modèles permettant de modéliser l'impact des imprécisions géométriques sur les mesures de longueur ou de surface.

1.3.2.2 Modèles relatifs aux mesures géométriques

Chrisman et Yandell (1988) ont proposé un modèle statistique sur l'effet des erreurs de positionnement des sommets d'une géométrie polygonale sur le calcul d'aire. Ce modèle part de l'hypothèse que les sommets sont soumis à des erreurs indépendantes et identiquement distribuées. Cependant aucune hypothèse n'est formulée sur la nature de la distribution des erreurs (normale ou log-normale par exemple). Le modèle statistique permet de formuler une estimation de la variance de l'aire A d'un polygone (calculée avec l'équation 1.2).

La variance de l'aire A est déterminée dans l'équation 1.12.

$$\begin{aligned}\sigma_A^2 &= 0.5n{\sigma^2}_x{\sigma^2}_y(1-\rho^2) \\ &+ 0.25\sum_{i=1}^{n}\left[{\sigma^2}_y(x_{i+1}-x_{i-1})^2 - {\sigma^2}_x(y_{i+1}-y_{i-1})^2\right] \\ &- 0.5\sum_{i=1}^{n}\rho\sigma_x\sigma_y(y_{i+1}-y_{i-1})(x_{i+1}-x_{i-1})\end{aligned} \tag{1.12}$$

où :
- n est le nombre de sommets du polygone
- ${\sigma^2}_x$ et ${\sigma^2}_y$ sont respectivement les variances en x et en y de l'erreur de localisation des sommets du polygone
- ρ est le coefficient de corrélation des erreurs entre les coordonnées x et y

Prisley *et al.* (1989) ont également développé une expression de la moyenne et de la variance sur l'erreur de calcul d'aire d'un polygone.
Griffith (1989) a lui aussi discuté du problème des erreurs de digitalisation sur les calculs d'aire et de longueur, en les exprimant sous forme matricielle, et en soulignant quelques pistes de recherches futures.
Kiiveri (1997) a également proposé un modèle général d'imprécision sur le positionnement, mais propose aussi une expression de l'imprécision en termes de mesures de longueur, de périmètre et d'aire.

Goodchild *et al.* (1999) ont également proposé un modèle basé sur des simulations, dédié aux utilisateurs de données géographiques afin qu'ils puissent eux-même évaluer la variance de l'aire des polygones en fonction de l'imprécision de positionnement des sommets composant leur géométrie.

Plus récemment, Goodchild (2004) a proposé un cadre général d'analyse sur les erreurs de mesure dans les SIG, décomposé en quatres parties :
- modèles et lois de propagation d'erreurs de position,
- intersections de points avec des polygones,
- intersections entre polygones,
- mesures de longueur et de surface.

La modélisation de l'imprécision sur les mesures de longueur et d'aire est plus largement détaillée dans Leung *et al.* (2004).

De Bruin *et al.* (2008) ont proposé un modèle permettant d'estimer l'imprécision des mesures d'aire de parcelles agricoles saisies par GPS, afin d'évaluer le manque à gagner des agriculteurs par rapport à des mesures d'aire réalisées sur des parcelles cadastrales sur cartes. Cette étude utilise également le logiciel "*Data Uncertainty Engine*" afin de simuler les erreurs de localisation des limites de parcelles.

Enfin, Julien (2008) a proposé un modèle théorique d'estimation de l'imprécision des mesures géométriques de longueur et de surface, en modélisant l'erreur de localisation de chaque point de la géométrie par une loi normale indépendante en x et y, et en prenant en compte l'angularité entre segments successifs.
Pour exprimer l'imprécision des mesures de longueur $\sigma(L)$, à partir d'une erreur moyenne quadratique ϵ_q, le modèle proposé par Julien (2008) prend en compte les angles inscrits entre segments successifs θ_i, conformément à l'équation 1.13.

$$\sigma(L) = \sqrt{1 + 2 \sum_{2 \leq i \leq n-1} sin^2 \frac{\theta_i - \theta_{i-1}}{2}} * \epsilon_q \tag{1.13}$$

Ce modèle a été appliqué à l'étude de l'imprécision des longueurs d'objets routiers dans Girres et Julien (2010).

Ainsi, comme nous avons pu voir dans cette section, les modèles d'imprécision géométrique traitant des mesures de longueur et de surface sont beaucoup plus rares que leurs homologues permettant d'estimer l'imprécision de positionnement, pour lesquels ils dépendent généralement, comme le montrent les contributions de Kiiveri (1997) ou Goodchild (2004). Le principal point commun entre ces modèles est qu'ils se basent sur des erreurs de positionnement des sommets, qui suit généralement une distribution gaussienne indépendante en x et en y. L'utilisation de simulations d'erreurs offre une possibilité robuste et aisée à mettre en œuvre afin d'évaluer l'imprécision des mesures de surface ou de longueur réalisées en aval. Cependant des approches par tests peuvent également être très utiles, en particulier lorsque différents processus générateurs d'erreurs interviennent sur la géométrie des objets, rendant difficile le paramétrage du bruit dans une approche par simulations. Cependant, malgré tout l'intérêt que présentent ces différents modèles, nous allons voir par la suite qu'ils ne sont pas véritablement adaptés pour estimer les imprécisions des mesures géométriques de longueur ou de surface, en particulier dans un contexte opérationnel.

Synthèse Section Nous avons vu dans cette section que de nombreux modèles d'évaluation de l'imprécision géométrique ont été développés ces dernières années en recherche. Cependant, la grande majorité des modèles proposés s'attache à évaluer l'imprécision relative au positionnement des objets géographiques. Les contributions proposant des modèles d'estimation de l'imprécision des mesures géométriques sont, à l'inverse, beaucoup moins nombreuses.

Conclusion de la section

Cette section a permis de réaliser un état de l'art des principales mesures et modèles développés afin de qualifier l'imprécision géométrique des objets vectoriels.

Concernant les mesures, nous avons vu que pour chaque type de primitives graphiques (ponctuelles, linéaires et surfaciques), de nombreux indicateurs sont disponibles, principalement basés sur des comparaisons avec des données de référence. Cependant, dans le cadre règlementaire français, la qualification de l'imprécision géométrique des objets vectoriels demeure basée sur un contrôle ponctuel.

Concernant les différents modèles d'imprécision géométrique développés ces dernières années en recherche, nous avons vu qu'ils cherchaient majoritairement à modéliser l'imprécision géométrique en termes de positionnement mais beaucoup plus rarement l'imprécision des mesures de longueur et de surface. Malgré cela, nous avons pu voir que deux approches se différencient : l'approche par simulations et l'approche par tests. Bien que l'approche par simulations s'avère être la plus populaire ces dernières années, l'approche par tests peut se montrer très intéressante lorsque la complexité des erreurs ne peut pas se limiter à des simulations en intégrant un bruit aléatoire.

Conclusion du chapitre

Ce premier chapitre a permis de réaliser un état de l'art non-exhaustif sur l'évaluation de l'imprécision géométrique des données géographiques vectorielles.

Nous avons tout d'abord montré que les données géographiques vectorielles constituent une représentation simplifiée et imparfaite du monde réel, et avons cherché à en expliquer les causes. Deux processus majeurs ont été identifiés pour expliquer ces causes d'imperfections : le processus d'abstraction du monde réel et le processus de production des données.
Le processus d'abstraction a pour objectif de réduire la complexité du monde réel, tout en satisfaisant des objectifs en termes d'utilisation des données, de niveau de détail des objets, ou encore de volume d'information. Les spécifications permettent de formaliser ce processus d'abstraction, qui conditionne étroitement le processus de production mobilisé pour saisir la géométrie des objets. Nous avons ensuite montré que chaque processus de production a pour effet de générer des erreurs, plus ou moins grandes, qui impactent la géométrie des objets, et par conséquence, les mesures géométriques calculées.

Nous avons vu par la suite que l'imperfection des données géographiques peut générer une incertitude quant aux décisions prises par un utilisateur sollicitant ces données. Afin de réduire cette incertitude, il est nécessaire d'évaluer et de communiquer la qualité des données géographiques. Après avoir présenté les différents éléments permettant de qualifier la qualité des données géographiques, ce travail de thèse s'est focalisé sur l'évaluation de l'imprécision géométrique des données.

Nous avons vu que de nombreuses mesures et de modèles développés ces dernières années, tant dans le domaine de la recherche que dans le cadre règlementaire, permettent de qualifier l'imprécision géométrique des données géographiques. Cependant, nous avons montré que ces modèles et mesures cherchent majoritairement à quantifier l'imprécision géométrique en termes de positionnement des objets, plutôt qu'en termes de mesures géométriques.

Ce dernier constat nous incite à explorer davantage la question de l'estimation de l'imprécision des mesures géométriques, étant donné que de nombreuses applications, sollicitées dans le cadre de prises de décision, sont basées sur des mesures géométriques de longueur et de surface.

Chapitre 2

Proposition : un modèle d'estimation de l'imprécision des mesures géométriques

Introduction

Nous avons vu dans le chapitre précédent que les bases de données géographiques sont inévitablement entachées d'imperfections. Ces imperfections résultent de différents processus impactant notamment la géométrie des objets vectoriels. Nous en avons identifié deux : le **processus d'abstraction** du mondé réel et le **processus de production** des données. Dans la pratique l'impact de ces deux processus s'accumule, générant des imprécisions géométriques sur les objets restitués, qui se répercutent sur les mesures géométriques de longueur ou de surface.

Afin de réduire l'incertitude des décisions prises lors de l'utilisation de données géographiques, des indicateurs permettant de qualifier la qualité d'une base de données géographiques peuvent être produits. En ce qui concerne la dimension géométrique des objets vectoriels, ces indicateurs se focalisent uniquement sur l'exactitude de position des objets, mais jamais sur l'imprécision des mesures géométriques. Or, de nombreux usages de l'information géographique sollicitent aujourd'hui des mesures géométriques de longueur ou de surface, sans jamais que ne soit communiquée quelque information sur l'imprécision de la mesure calculée. En fait cela peut se comprendre, car comme nous allons le voir, produire un indicateur sur l'imprécision des mesures géométriques est une tâche complexe, étant donné que de nombreux processus impactent ces mesures.

Cette section va nous permettre de définir les principes généraux d'un modèle d'estimation de l'imprécision des mesures géométriques calculées à partir d'objets de base de données géographiques vectorielles. Une synthèse des besoins et moyens actuels sera proposée dans un premier temps, justifiant le développement d'outils plus adaptés aux besoins des utilisateurs. La proposition générale d'un modèle d'estimation sera donc exposée dans un second temps, en définissant tout d'abord les objectifs principaux du modèle, son fonctionnement envisagé, ainsi que les principes régissant son développement.

2.1 Besoins et moyens actuels d'évaluation de l'imprécision géométrique

Les utilisateurs de données géographiques doivent composer avec leurs imperfections, et en maitriser les conséquences. A cet égard, différents outils (métadonnées, mesures et modèles) sont proposés afin d'évaluer et de communiquer la qualité de ces données géographiques, en particulier sur leur dimension géométrique. Nous allons voir dans cette synthèse que les outils proposés aux utilisateurs ne permettent pas de réduire entièrement l'incertitude liée à l'usage de données géographiques, notamment concernant les mesures géométriques de longueur et de surface. Afin de renseigner l'utilisateur de manière pertinente, il convient tout d'abord d'identifier ses besoins potentiels. Ainsi, nous essaierons d'identifier les besoins des utilisateurs de données géographiques en termes d'imprécision géométrique, ce qui nous permettra de mieux comprendre les limites des moyens actuels d'évaluation. Egalement, bien que de nombreux modèles aient été développés en recherche ces dernières années sur ce sujet, nous verrons qu'ils ne traitent que partiellement du problème d'estimation de l'imprécision des mesures géométriques. Dans ce contexte, ceci nous amènera finalement à définir le positionnement de ce travail de thèse.

2.1.1 Métadonnées, qualité géométrique et besoins des utilisateurs

Nous allons voir dans cette section en quoi les métadonnées, qui constituent aujourd'hui le principal moyen permettant de renseigner les utilisateurs sur la qualité des données géographiques, ne s'avèrent pas adaptées à tous les besoins des utilisateurs. Nous verrons que ceci est particulièrement le cas concernant les informations relatives à la qualité géométrique des bases de données géographiques. Ce constat nous permettra d'identifier les attentes des utilisateurs en termes d'information sur l'imprécision géométrique.

2.1.1.1 Métadonnées et qualité externe

Les métadonnées constituent le principal outil de renseignement de l'imprécision géométrique pour les utilisateurs de bases de données géographiques. Or comme nous allons le voir, ces métadonnées sont généralement complexes à renseigner et surtout à interpréter pour un utilisateur. De plus, concernant l'imprécision géométrique des données géographiques, elles ne se focalisent que sur l'exactitude de position.

Optionalité de l'évaluation de la qualité Comme il a été mentionné précédemment, la documentation de la sous-partie sur les éléments d'évaluation de la qualité des données géographiques est optionnelle dans la norme ISO 19115. On peut ainsi légitimement douter que les producteurs de données puissent investir suffisamment de temps et de moyens afin d'évaluer et documenter les différents éléments proposés, surtout si l'on prend en compte la complexité des mesures proposées dans la spécification technique ISO 19138. En effet, en prenant l'exemple des mesures proposées pour évaluer l'exactitude de position, si le calcul de l'Erreur Moyenne Quadratique est un indicateur communément renseigné par les producteurs de données, on doute que ce soit le cas pour des indicateurs comme la Matrice de Variance-Covariance par exemple.

Difficulté d'interprétation Au delà de la difficulté à produire ces indicateurs, on peut s'interroger sur la capacité d'intégration et d'interprétation de ces indicateurs par l'utilisateur final. En effet, comment un utilisateur peut-il interpréter ces informations, afin de savoir si la base de données qu'il utilise est adaptée aux applications qu'il souhaite réaliser ? Les métadonnées sont pourtant censées être destinées à l'utilisateur, afin que celui-ci évite les mauvais usages. Cependant, on s'aperçoit rapidement, à la lecture de la spécification technique ISO 19138, que les mesures proposées n'offrent pas la possibilité d'évaluer le degré d'adéquation de la base de données avec les utilisations qui en seront faites. On voit donc que l'évaluation de la qualité externe, à travers le concept de "*fitness for use*", n'est pas réellement mise en œuvre dans les métadonnées, comme le soulignent Servigne *et al.* (2005). Cela rejoint également le propos de Boin et Hunter (2008) qui s'interrogent sur la capacité des métadonnées à véritablement informer l'utilisateur sur la qualité externe des données géographiques.

Bien que ce constat soit globalement le même pour l'ensemble des éléments d'évaluation de la qualité, qui s'attachent davantage à évaluer la qualité interne des bases de données géographiques, ceci illustre parfaitement le paradoxe porté par les métadonnées, que l'on peut considérer comme des "informations produites par des experts, pour des experts". Gervais (2004) estime que ces informations sont même difficiles à comprendre par des experts en géomatique.

Pour compenser l'absence de qualité externe, un sous-élément "utilisabilité" a été prévu dans la future norme ISO 19157, permettant à l'utilisateur d'enrichir les métadonnées d'indicateurs spécifiquement développés afin de tester l'adéquation de la base de données avec les usages prévus. On peut par contre imaginer qu'il existe potentiellement autant d'indicateurs que d'usages prévus, ce qui demeure délicat à formaliser.

Métadonnées et imprécision géométrique Dans la spécification technique ISO 19138, les indicateurs d'imprécision géométrique ne portent que sur l'erreur de positionnement, comme nous avons pu le souligner auparavant. De plus, ces indicateurs ne portent que sur du contrôle ponctuel, bien que de nombreuses mesures plus adaptées aux géométries linéaires et surfaciques aient été développées. Nous ne nous attarderons pas sur la pertinence des mesures proposées dans la spécification technique ISO 19138, mais préférons nous focaliser sur un autre point.

Comme de nombreux indicateurs permettent déjà de qualifier l'erreur de positionnement, il semble pertinent de proposer à l'utilisateur des informations liées à d'autres besoins, comme par exemple l'évaluation de l'imprécision des mesures géométriques de longueur ou de surface. Nous avons évoqué à ce titre qu'il serait intéressant d'intégrer des indicateurs d'imprécision des mesures géométriques dans les métadonnées, au sein de la partie sur les mesures d'exactitude relative. Pour le moment, les indicateurs proposés ne portent que sur des calculs de distance euclidienne entre deux points de la base de données à renseigner, en comparaison avec sa référence. Il semblerait pertinent de compléter les indicateurs proposés d'une mesure similaire, basée cette fois sur un calcul de longueur géométrique. Cette proposition se heurte cependant à l'optionalité du renseignement de la sous-partie sur la qualité dans les métadonnées, ainsi qu'à la disponibilité de données de référence pour pouvoir mesurer cet indicateur.

Constat Les différents éléments évoqués dans cette section nous permettent donc de poser un triple constat sur l'intérêt des métadonnées en termes de qualité géométrique et de besoins utilisateurs :
- la sous-partie sur la qualité est renseignée de manière optionnelle,
- les indicateurs proposés sont complexes et ne permettent pas facilement de juger de l'adéquation de la base de données avec les besoins des utilisateurs,
- les indicateurs d'imprécision géométrique ne portent que sur l'erreur de position.

On voit bien à travers ce triple constat que les métadonnées, malgré leur intérêt, ne constituent pas l'outil le plus adapté pour renseigner l'utilisateur en termes d'adéquation à ses besoins, et donc de qualité externe. Ceci est particulièrement vrai pour l'information sur l'imprécision géométrique. Afin de mieux comprendre les enjeux liés à ce problème, nous allons tenter d'identifier les différents niveaux de besoins des utilisateurs en termes d'information sur l'imprécision géométrique.

2.1.1.2 Imprécision géométrique et besoins utilisateurs

L'estimation de l'imprécision géométrique se focalise presque exclusivement sur le positionnement des objets. Les nombreuses classifications des éléments d'évaluation de la qualité des données géographiques, proposées dans diverses contributions (Van Oort, 2005), ne prennent effectivement en compte que l'exactitude de position afin de renseigner l'imprécision géométrique. Pourtant, on peut malgré tout considérer qu'il existe une typologie des besoins utilisateurs en termes d'estimation de l'imprécision géométrique des données géographiques vectorielles.

Typologie des besoins Cette typologie peut être divisée en trois catégories, qui rejoignent les principaux axes de recherches dans la modélisation de l'imprécision géométrique. Ces besoins utilisateurs concernent l'estimation de :

1. l'imprécision géométrique en termes de positionnement,
2. l'imprécision des mesures géométriques de longueur et de surface,
3. l'impact de l'imprécision géométrique sur des opérations topologiques (intersection, union...).

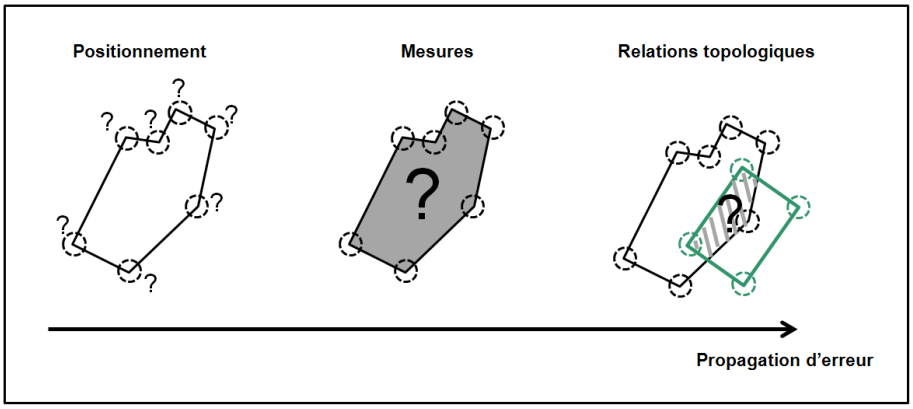


Figure 2.1 – Imprécision géométrique et besoins utilisateurs

On peut en effet considérer que la propagation de l'erreur de positionnement initiale de la géométrie des objets vectoriels impacte :
- dans un premier temps, les mesures géométriques qui seront réalisées à partir de cette géométrie imprécise,
- dans un second temps, les opérations topologiques (union, intersection) réalisées avec d'autres objets, portant eux-même une erreur de positionnement (cf. figure 2.1).

Inadéquation des moyens d'estimation Malgré ces conséquences évidentes de la propagation de l'erreur de positionnement, les indicateurs permettant d'estimer l'imprécision géométrique en termes de positionnement ne sont pas vraiment pertinents en vue d'estimer l'imprécision des mesures géométriques de longueur et de surface.

Par exemple, comme illustré dans la figure 2.2, un objet mal positionné dans une base de données géographiques (par rapport à sa référence) ne verra pas forcément la mesure géométrique qu'il porte (ici une mesure d'aire) impactée par l'erreur de positionnement, même si celle-ci est importante. A l'inverse, l'erreur de positionnement peut parfois être faible, et l'erreur de mesure géométrique importante.

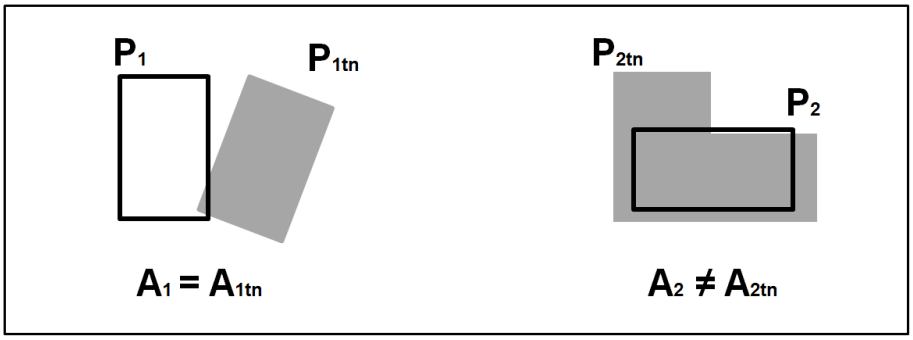


Figure 2.2 – Un objet mal positionné portant une erreur de mesure faible, et un objet bien positionné portant une erreur de mesure importante

Cet exemple justifie le fait que l'estimation de l'imprécision des mesures géométriques de longueur et de surface ne peut se contenter d'exploiter des indicateurs utilisés pour qualifier l'imprécision en termes de positionnement. Il semble ainsi nécessaire de développer des méthodes spécifiques à l'estimation de l'imprécision des mesures géométriques. Cependant, aucun moyen n'est véritablement fourni à l'utilisateur pour l'estimer aujourd'hui. Ce constat justifie pourquoi, dans le cadre de ce travail de thèse, nous avons décidé de nous focaliser sur l'estimation de l'imprécision des mesures géométriques de longueur et de surface, et de mettre au point des méthodes permettant de la communiquer à l'utilisateur.

L'absence d'indicateurs sur l'imprécision des mesures géométriques peut se révéler dommageable, puisque de nombreuses applications SIG sont aujourd'hui basées sur des mesures de longueur ou de surface, comme par exemple le calcul d'un itinéraire routier (qui nécessite la longueur, cf. 2.3), ou les calculs de densité de population ou de taxe foncière (qui nécessitent la surface).

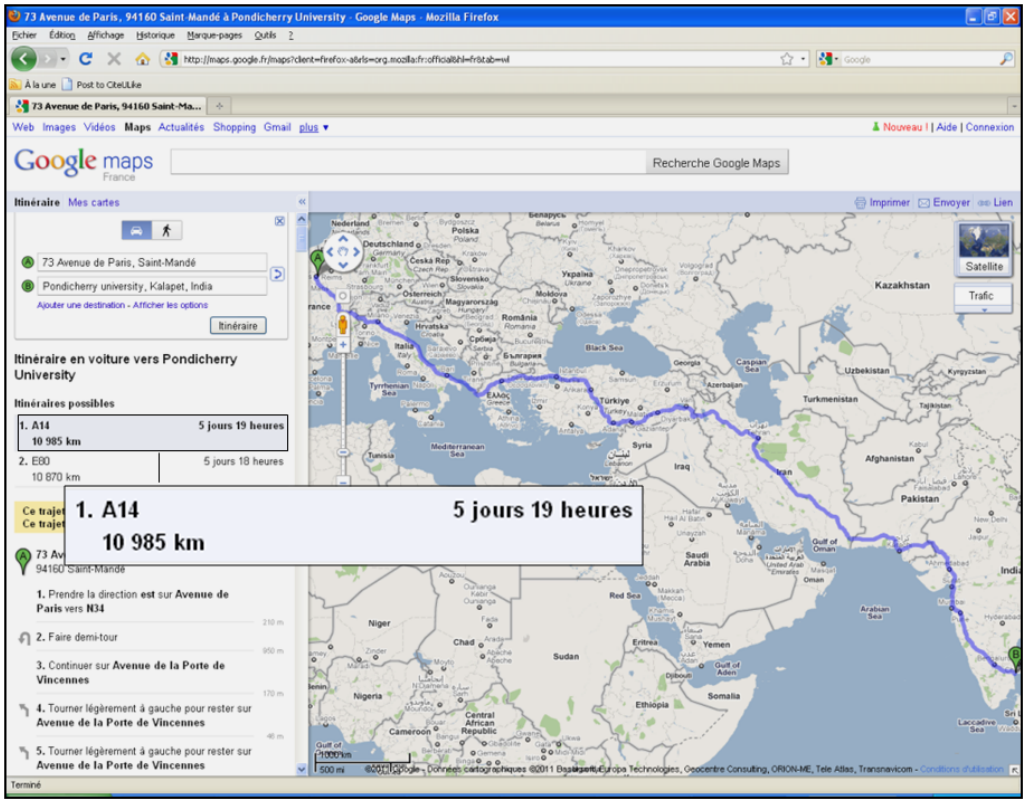


Figure 2.3 – Un calcul d'itinéraire routier réalisé avec Google Map : Quelle est l'imprécision de la mesure ?

Autant l'imprécision sur les mesures géométriques peut s'avérer faible sur des bases de données topographiques, saisies à grande échelle avec des méthodes d'acquisition de haute précision, autant elle peut être très importante sur des bases de données saisies à moyenne échelle, ayant parfois subi les effets de la généralisation cartographique. L'absence d'indicateurs sur l'imprécision des mesures géométriques peut donc s'avérer dommageable pour l'utilisateur final, afin de réduire l'incertitude liée à sa prise de décision.

Pour pouvoir estimer l'imprécision sur les mesures de longueur ou de surface réalisées à partir d'objets géographiques vectoriels, nous avons vu que peu de modèles ont été proposés. Nous allons voir dans la section suivante l'intérêt et les limites de ces modèles dans le cadre de ce travail de thèse.

Synthèse Section Plusieurs niveaux de besoins peuvent être attendus par les utilisateurs en termes d'évaluation de la qualité géométrique des données géographiques. Les métadonnées, qui constituent le principal moyen de communication de cette information, ne portent que sur l'imprécision relative au positionnement des objets géographiques, mais pas sur l'imprécision des mesures géométriques portées par ces objets.

2.1.2 Limites des modèles d'évaluation existants

Nous allons voir dans cette section que les modèles d'estimation de l'imprécision des mesures géométriques existants ne permettent pas de modéliser l'impact de tous les processus de production, mais également certains impacts classiques impliqués par le processus d'abstraction.

2.1.2.1 Limites relatives à l'impact du processus de production

La majorité des modèles développés en recherche ces dernières années ont pour vocation de modéliser l'imprécision géométrique des objets vectoriels en termes de positionnement. Beaucoup plus rares sont ceux qui ont étudié l'impact de ces imprécisions géométriques sur des mesures de surface ou de longueur, comme Chrisman et Yandell (1988), Kiiveri (1997), ou Goodchild (2004), qui ont cherché à les formaliser par application d'un modèle de propagation d'erreurs de positionnement. Ces modèles sont généralement basés sur une probabilité de distribution des erreurs de position suivant une loi normale, indépendante en x et y. Les approches par simulations, relativement simples à mettre en œuvre, peuvent ainsi être utilisées pour modéliser certaines sources d'erreurs affectant la géométrie des objets géographiques, comme par exemple les erreurs de pointé (Bolstad *et al.*, 1990) ou encore les imprécisions d'acquisition par relevés GNSS (De Bruin *et al.*, 2008). Il a déjà été montré que l'utilisation de ces approches par simulations offre une possibilité intéressante afin d'estimer l'impact de ces erreurs sur des mesures de longueur ou de surface, en exploitant un grand nombre d'objets simulés.

Cependant, comme nous l'avons évoqué auparavant, l'impact des processus de production sur la géométrie des objets vectoriels peut se manifester sous des formes très variées, et leur modélisation ne se limite pas à l'application d'un bruit aléatoire sur les sommets de la géométrie. En effet, l'impact de la généralisation cartographique peut se manifester par des modifications de forme importantes (et de manière hétérogène), générant des erreurs de mesure considérables. Par exemple, en zones de montagne, les processus de généralisation cartographique peuvent mener à la suppression de virages sur des objets routiers afin de faciliter la lecture de la carte (Mustière, 2001). On comprend bien par ce type d'exemple que les approches par simulations ne s'avèrent pas adaptées pour modéliser l'impact de toutes les erreurs potentiellement affectées par les processus de production, en particulier le processus de généralisation cartographique.

La figure 2.4 permet d'illustrer ce propos. Elle montre que l'application d'un modèle d'erreurs aléatoires sur une route généralisée au 1 :100.000 ne permet pas de modéliser l'imprécision géométrique, tant en termes de positionnement que de mesure de longueur, par rapport à sa référence topographique. En effet, jamais les routes simulées (en traits bleus fins) ne ressemblent à la réalité (en rouge).

Egalement, il est difficilement concevable d'appliquer un même modèle d'erreur par simulations sur des objets saisis manuellement par un opérateur (dans le cadre de relevés photogrammétriques par exemple) et sur des objets saisis par acquisitions GNSS. En effet, les erreurs impliquées par chacun des deux processus de saisie ont des causes et des effets totalement différents. Les regrouper dans un même modèle par pragmatisme peut ainsi s'avérer dangereux.

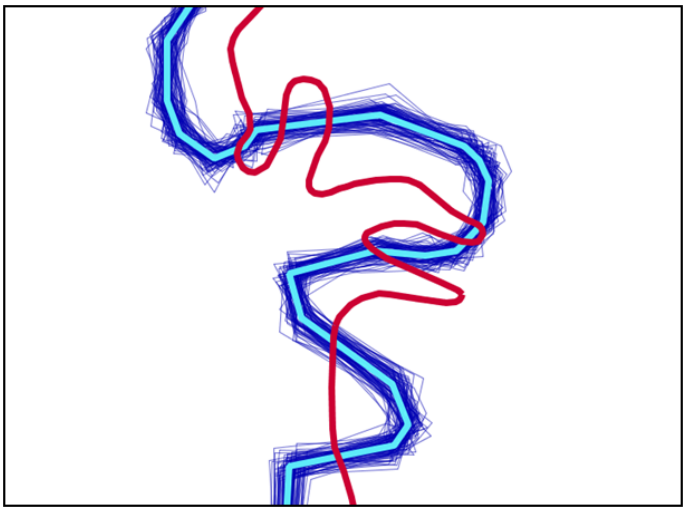


Figure 2.4 – Inadéquation d'une approche par simulations pour modéliser l'impact de la généralisation cartographique

Pour illustrer ce propos, la figure 2.5 présente le résultat d'une étude de comparaison des traces saisies par trois récepteurs GPS différents, mais équipés de la même puce (Ray (2009), voir également Ray *et al.* (2010)).

Cette figure montre que les traces produites par les trois récepteurs GPS présentent toutes des erreurs, dont certaines n'auraient pas été générées avec un processus de restitution manuel par un opérateur de saisie. Ceci démontre bien que la modélisation des impacts de l'erreur de saisie par un opérateur-restituteur ou par un récepteur GPS ne peuvent être envisagée de manière identique.

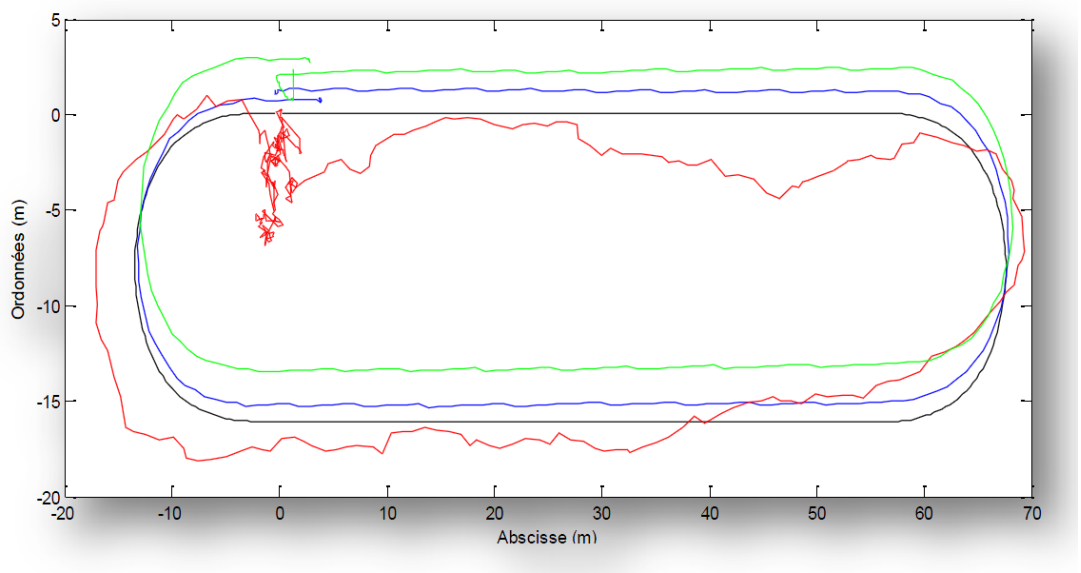


Figure 2.5 – Comparaison des traces de trois récepteurs GPS (GlobalSat BT-338 en rouge, Emtac BTGPS S3 en bleu, et Royaltek RBT-2110 en vert), d'après Ray (2009)

Chaque processus de production nécessite donc un modèle d'erreur qui lui est propre, modélisant l'erreur associée (et son impact sur les mesures géométriques) en fonction des spécificités du processus impliqué. Nous considérons donc qu'il n'existe pas de modèle d'erreur universel sur les données géographiques.
Mais au delà de l'impact des processus de production, les différents modèles d'erreur développés en recherche ces dernières années ne prennent pas non-plus en compte certains impacts impliqués par le processus d'abstraction du monde réel.

2.1.2.2 Limites impliquées par le processus d'abstraction

Au delà de l'impact des processus de production, les mesures géométriques de longueur et de surface calculées à l'aide de logiciels SIG classiques sont soumises à d'autres effets qui ne sont généralement pas abordés dans les modèles exposés précédemment. On peut notamment en retenir trois, impliquées par le processus d'abstraction du monde réel, à savoir les impacts :
– de la projection cartographique,
– de la non-prise en compte du terrain,
– de l'approximation polygonale des courbes.

Impact de la projection cartographique En règle générale, la géométrie des objets vectoriels est saisie en coordonnées planes, comme nous avons pu le voir dans les extraits de spécifications. De la même manière, les mesures de longueur et de surface sont réalisées par commodité dans un plan, suite à une projection de l'ellipsoïde terrestre (qui constitue déjà une approximation de la surface de la Terre). Comme la projection d'une surface courbe dans un plan ne peut se faire sans générer des déformations, les objets géographiques en sont impactés, tout comme les mesures géométriques de longueur et de surface (cf. figure 2.6). Ces impacts se ressentent particulièrement sur les objets géographiques de grandes dimensions. Ainsi, l'impact de la projection cartographique doit être pris en compte dans l'estimation de l'imprécision des mesures géométriques, ce qui n'est pas le cas dans les modèles présentés précédemment.

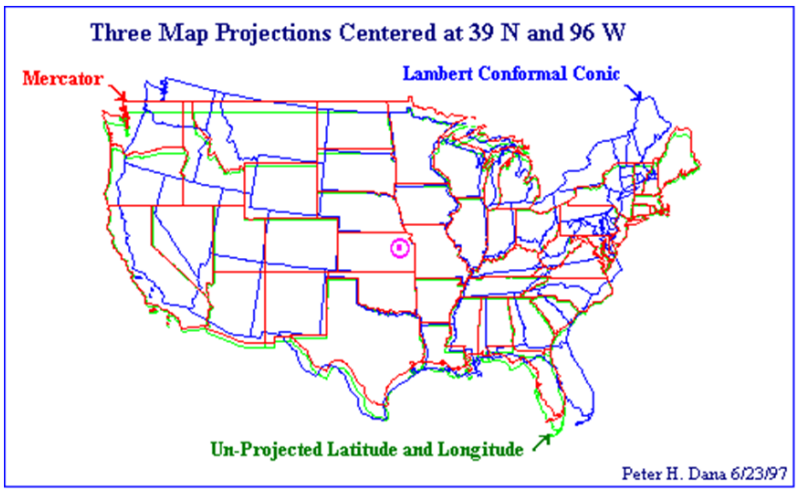


Figure 2.6 – Impact de trois projections cartographiques sur les Etats-Unis, d'après Dana (1997)

Impact de la non-prise en compte du terrain Bien que paraissant évidente, la non-prise en compte du terrain peut s'avérer particulièrement génératrice d'erreurs en termes de mesures de longueur ou de surface. Pourtant, de nombreuses bases de données géographiques ne disposent pas de coordonnées tridimensionnelles sur la géométrie des objets géographiques (principalement du fait des moyens financiers qu'implique leur saisie). Ceci engendre une sous-estimation systématique des mesures géométriques, en particulier dans les zones où les pentes sont importantes (cf. figure 2.7).

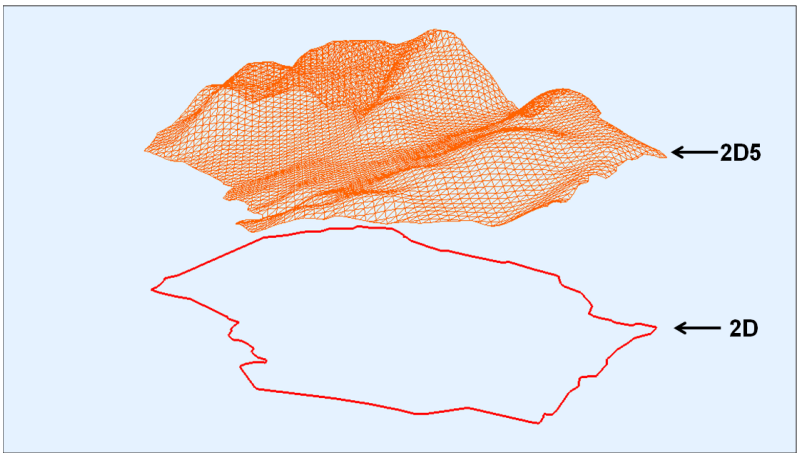


Figure 2.7 – Représentation 2D et 2D5 d'une limite communale

Ainsi, la prise en compte du relief sur la géométrie des objets vectoriels s'avère indispensable pour pouvoir modéliser de manière réaliste l'imprécision des mesures de longueur et de surface. Malgré cela, la majorité des modèles présentés précédemment ne prennent généralement pas en compte cette composante, et ne proposent que des modèles d'imprécision basés sur des mesures réalisées en deux dimensions.

Impact de l'approximation polygonale des courbes Dans une base de données géographiques, la géométrie des objets vectoriels est généralement représentée par une succession de segments de droites. Cependant dans la réalité, les entités du monde réel peuvent présenter des formes courbes (cf. figure 2.8).

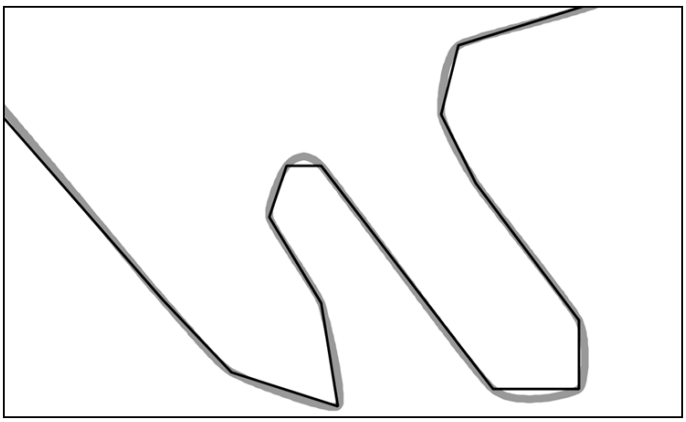


Figure 2.8 – Impact de l'approximation polygonale sur la géométrie d'une route

L'approximation polygonale de la géométrie de ces objets a dans ce cas pour conséquence de sous-estimer systématiquement les longueurs calculées. Les modèles présentés précédemment ne prennent pas non-plus en compte l'impact de l'approximation polygonale sur les mesures géométriques de longueur ou de surface.

Bilan Nous venons de présenter trois sources potentielles d'erreurs sur les mesures géométriques de longueur et de surface. Ces sources d'erreurs peuvent être considérées comme liées au processus d'abstraction du monde réel qui définit les **règles de représentation** des objets d'une base de données géographiques. En effet, ce sont dans les spécifications que sont décidés dans quel système de coordonnées est saisie la géométrie des objets ou si une information d'altitude doit être saisie ou non. L'impact de l'approximation polygonale est quant à lui inhérent à la représentation en mode vectoriel dans la majorité des logiciels SIG. Afin d'estimer l'imprécision des mesures géométriques calculées à partir de la géométrie des objets vectoriels, il convient de prendre en compte l'impact de ces règles de représentation, du fait des conséquences que nous venons d'évoquer.

Nous voyons donc bien, à travers ce constat, que les mesures géométriques de longueur et de surface sont impactées par différentes contributions qui ne se limitent pas aux processus de production. Les différents modèles développés ces dernières années en recherche ne prennent pas en compte toutes les causes des erreurs de mesure. Ces modèles se contentent généralement de modéliser une erreur globale, sans chercher à identifier les différents impacts impliqués dans l'erreur finale. Cependant, ces modèles "globaux" ne sont valables que dans des conditions identiques à celles dans lesquelles ils ont été élaborés. Si une cause d'erreur supplémentaire est intégrée sur les objets traités, ces modèles ne sont donc a priori plus valables. Cela confirme l'idée que la modélisation de l'imprécision des mesures géométriques de longueur et de surface est complexe, puisqu'elle nécessite d'intégrer les impacts de processus hétérogènes affectant la géométrie des objets.

Synthèse Section Cette section a permis de montrer que les modèles d'estimation de l'imprécision des mesures géographiques développés ces dernières années en recherche ne permettent pas de modéliser les impacts de tous les processus de production. Egalement ces modèles ne prennent pas en compte les impacts de règles de représentation, impliquées par le processus d'abstraction du monde réel, qui affectent les mesures géométriques de longueur et de surface.

2.1.3 Positionnement de la thèse

Comme nous venons de le montrer dans cette synthèse, renseigner un utilisateur de l'imprécision des mesures géométriques calculées à partir des objets d'une base de données géographiques est un sujet complexe. Nous allons dans cette section présenter le positionnement général de ce travail de thèse, conformément aux constats que nous venons d'évoquer.

Qualité externe La démarche dans laquelle s'inscrit ce travail de thèse s'oriente dans une logique de développement de moyens et de méthodes pour l'évaluation de la qualité externe des données géographiques. Il est en effet très délicat pour un utilisateur d'évaluer le degré d'adéquation d'une base de données géographiques avec ses besoins. La majorité des éléments d'évaluation, et des indicateurs associés sont développés dans une logique de qualité interne, davantage destinée à évaluer l'adéquation des bases de données géographiques avec leurs spécifications.

Nous cherchons donc, dans le cadre de ce travail de thèse, à mettre en œuvre des moyens d'évaluation permettant d'informer l'utilisateur de ce qu'il peut attendre de ses données géographiques. Il s'agit donc de développer un modèle accessible d'usage, qui permette d'informer un utilisateur pour que celui-ci réduise l'incertitude inhérente à un processus de prise de décision. Nous nous sommes pour cela focalisé sur certains besoins spécifiques, relatifs à l'estimation de l'imprécision géométrique des objets vectoriels.

Besoins identifiés En termes d'imprécision géométrique, un premier constat s'est imposé, montrant que les indicateurs proposés se limitent à renseigner l'exactitude de position des objets. Or, comme nous l'avons montré, les besoins actuels ne se limitent pas à quantifier l'erreur de positionnement des objets géographiques. Nous avons pu exposer que renseigner l'imprécision des mesures géométriques de longueur et de surface porte un véritable intérêt, étant donné que de nombreuses applications sont aujourd'hui basées sur ces mesures. Cependant, estimer l'imprécision géométrique en termes de positionnement ou de mesures de longueur et de surface n'impliquent pas les mêmes moyens et méthodes.

Moyens utilisés Pour estimer l'imprécision des mesures géométriques, nous avons montré qu'il était possible d'enrichir la partie sur les mesures d'exactitude relative proposées dans la spécification technique ISO 19138. Cependant, l'enrichissement permettant d'estimer l'imprécision de mesures géométriques suppose des comparaisons préalables avec une base de données de référence, qui n'est pas toujours disponible. De plus, étant donné que le renseignement des éléments d'évaluation de la qualité est optionnel dans la norme ISO 19115, on peut douter que ces informations soient saisies systématiquement. Bien que cette solution demeure optimale et envisageable, nous nous orientons davantage dans le cadre de ce travail de thèse vers le développement d'outils facilement mobilisables pour un utilisateur, pouvant être intégrés dans une interface SIG.

Choix de modélisation L'état de l'art des modèles d'estimation de l'imprécision des mesures géométriques montre que les modèles existants, généralement basés sur des approches par simulations, se révèlent adaptés pour modéliser l'impact de certains processus de production, comme les erreurs de pointé, voire d'acquisition GNSS. Ces mêmes modèles ne s'avèrent par contre pas adaptés pour modéliser l'impact de processus de production plus

complexes, comme la généralisation cartographique. Au delà des processus de production, ces modèles ne prennent pas non-plus en compte l'impact des règles de représentation, impliquées par le processus d'abstraction du monde réel. Ainsi, le modèle mis en œuvre dans le cadre de ce travail de thèse cherchera à estimer les impacts de ces processus hétérogènes, qui dans la pratique s'accumulent et génèrent une erreur globale de mesure.

Problématiques Ce positionnement général nous amène donc à formuler un certain nombre de questions qui guideront l'ensemble de ce travail de thèse :
- Comment modéliser l'impact des différents processus impliqués dans les erreurs de mesures géométriques de longueur et de surface ?
- Comment combiner ces différents impacts afin de produire une estimation globale de l'imprécision des mesures ?
- Comment mettre en œuvre cette évaluation de manière opérationnelle pour l'utilisateur ?

La section suivante nous permettra de poser les bases de l'élaboration de ce modèle d'estimation de l'imprécision des mesures géométriques, en définissant les objectifs, les moyens et les méthodes utilisées pour le mettre en œuvre, et que nous développerons en détail dans les chapitres suivants.

Synthèse Section L'objectif de ce travail de thèse consiste à développer un modèle d'estimation de l'imprécision des mesures géométriques, permettant de prendre en compte les impacts de processus de production hétérogènes, ainsi que de règles de représentation impliquées par le processus d'abstraction du monde réel. La mise en œuvre de ce modèle s'inscrit dans une démarche de développement de moyens et de méthodes pour l'évaluation de la qualité externe des données géographiques.

2.2 Spécifications générales du modèle d'estimation

L'objectif de ce travail de thèse réside dans l'élaboration d'un modèle permettant à un utilisateur d'estimer l'imprécision des mesures de longueur et de surface calculées à partir de la géométrie d'objets vectoriels. Ce modèle se définit par conséquent comme un **outil d'aide à la décision**, élaboré dans une logique d'évaluation de la qualité externe. Nous présenterons dans cette section les objectifs attendus du modèle, puis présenterons son fonctionnement envisagé, et enfin l'approche utilisée pour sa modélisation.

2.2.1 Objectifs du modèle d'estimation

Afin de bien délimiter les capacités du modèle d'estimation, nous présenterons dans cette section les objectifs attendus du modèle d'estimation. Ces objectifs peuvent être résumés en trois catégories, comme illustré dans la figure 2.9 :
- la définition des données à évaluer (entrée),
- les impacts qui seront modélisés (évaluation),
- la communication des résultats (sortie).

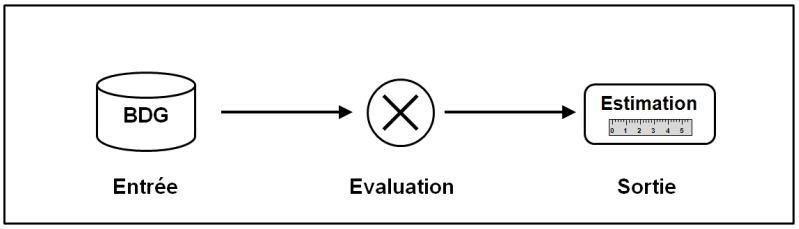


Figure 2.9 – Principe général du processus d'évaluation

2.2.1.1 Bases de données à évaluer

De nombreuses applications sollicitent aujourd'hui les mesures géométriques de longueur ou de surface. On peut citer à titre d'exemples les calculs d'itinéraires sur un réseau routier (qui nécessitent la longueur), ou encore les calculs de densité de population au sein d'une entité administrative (qui nécessitent la surface). Si pour certaines applications l'incertitude relative au résultat de la mesure ne s'avère pas trop lourde de conséquences, cela peut être le cas dans des applications plus stratégiques, comme par exemple pour les calculs de primes en fonction d'une surface exploitée en milieu agricole. La connaissance de l'imprécision de la mesure géométrique peut ainsi s'avérer pertinente pour l'utilisateur.

De nombreux indicateurs d'exactitude de position sont proposés dans les métadonnées. Ces indicateurs sont généralement basés sur des comparaisons préalables avec un jeu de données de référence. Nous avons vu qu'il pouvait être envisageable de produire des indicateurs d'imprécision des mesures géométriques dans la partie traitant de l'exactitude relative. Cependant, la production de ce type d'indicateur implique également des comparaisons préalables avec une base de données de référence. Or, parfois ces informations ne sont pas disponibles.

Principe d'évaluation Nous pouvons considérer que dans le pratique, et dans le cadre d'une prise de décision, un utilisateur exploite généralement la base de données la plus

juste, et parfois la seule réellement à disposition sur la zone ou l'étendue concernée. Il n'est donc pas toujours possible de disposer d'une référence pour réaliser des comparaisons, et ainsi produire un indicateur d'imprécision des mesures géométriques. C'est pourquoi nous pensons que le modèle développé doit permettre d'estimer l'imprécision des mesures géométriques de longueur ou de surface **sans données de référence** (cf. figure 2.10), étant donné que celles-ci ne sont pas toujours disponibles, voire existantes. Pour pallier l'absence de données de référence, des connaissances sur les données évaluées seront intégrées dans le processus d'évaluation.

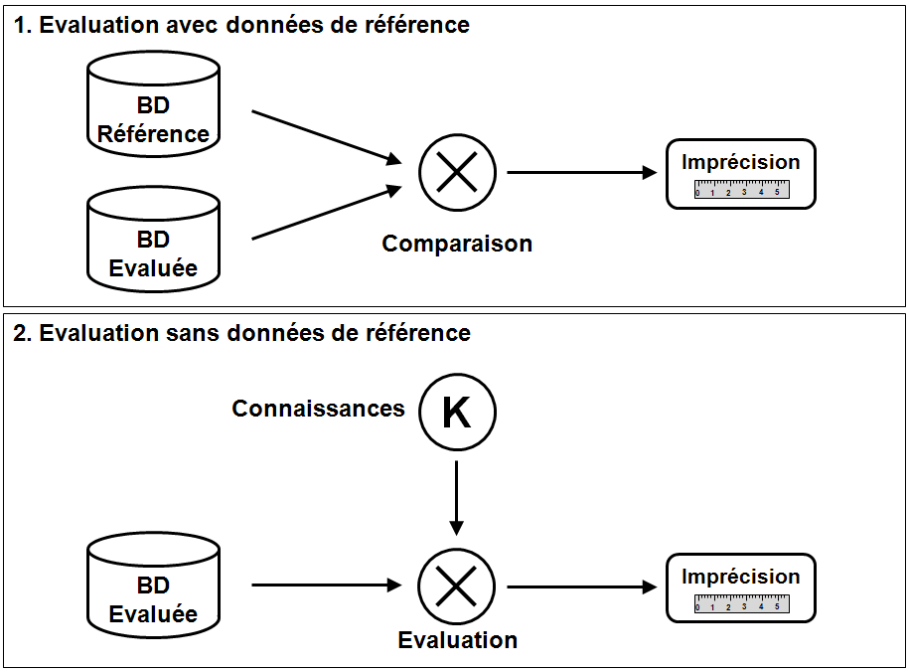


Figure 2.10 – Un processus d'évaluation sans données de référence

Classes d'objets concernées Le modèle développé doit permettre d'estimer l'imprécision des mesures calculées à partir de la géométrie d'objets représentant des **informations de référence**. Par information de référence, nous entendons des classes d'objets comme les routes, les voies ferrées ou encore des parcelles cadastrales ou des limites administratives. Cependant, certaines informations de référence présentent des limites floues. C'est par exemple la cas des forêts, ou encore la délimitation d'un trait de côte. Pour l'illustrer, la figure 2.11, proposée par Comber *et al.* (2005), présente la définition d'une forêt dans une sélection de pays du monde, en fonction de la couverture de canopée et de la hauteur des arbres. Nous voyons bien qu'il n'existe pas d'unicité sur la définition des forêts.

Par conséquent, les classes d'objets présentant des limites floues ne seront pas traitées par ce modèle. En effet, il s'avèrerait délicat d'estimer l'imprécision de mesures de longueur ou de surface sur des entités ayant déjà des limites mal définies dans le monde réel. Les expérimentations qui seront réalisées dans le cadre de la mise en œuvre du modèle porteront principalement sur des classes d'objets aux limites bien définies, comme par exemple les routes pour les géométries linéaires, ou encore sur des limites administratives pour les géométries polygonales.

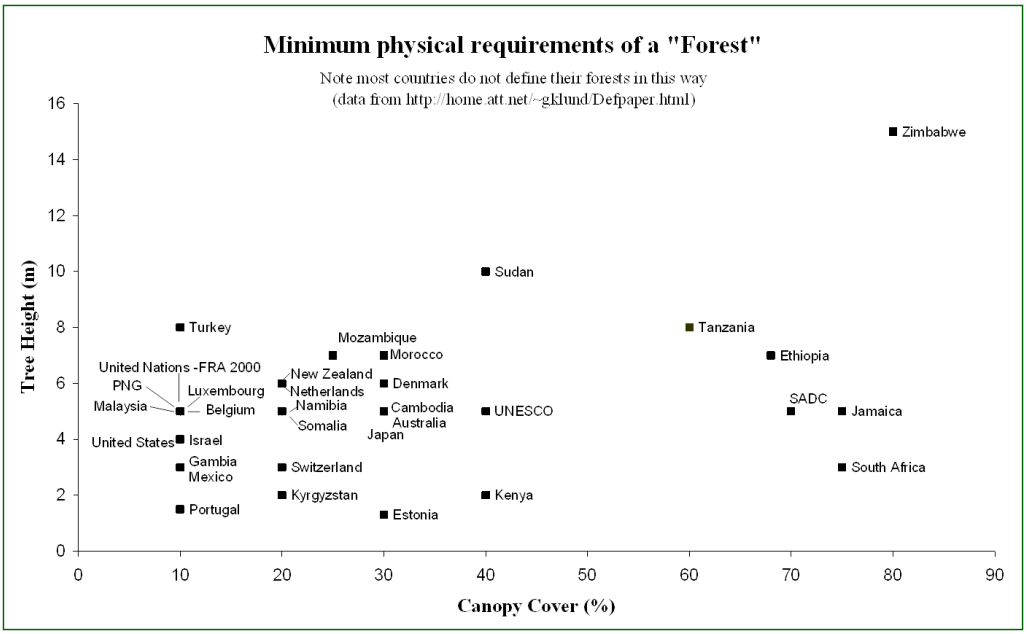


Figure 2.11 – La définition d'une forêt selon plusieurs pays, d'après Comber *et al.* (2005)

Echelles concernées Le niveau de détail d'une base de données géographiques est particulièrement conditionné par l'échelle de raisonnement à laquelle celle-ci est censée convenir. Le niveau d'imprécision géométrique des objets vectoriels en est donc intimement lié. Ainsi, nous pouvons partir de l'hypothèse que l'erreur de mesure croit lorsque l'échelle de représentation d'une base de données géographiques diminue. De ce fait, les données topographiques, généralement représentées à des échelles comprises entre le 1 :5.000 et le 1 :15.000 génèrent a priori une plus faible erreur sur les mesures que des données destinées à des représentations à moyenne échelle, c'est-à-dire du 1 :50.000 au 1 :250.000. C'est donc pour les **données représentées à moyenne échelle** que le modèle développé se verra le plus adapté. Au delà, dans les gammes de plus petites échelles, l'extrême niveau de simplification des géométries rend les mesures de longueur ou de surface inappropriées. Ce type d'utilisation des données peut être considéré comme une erreur d'usage, comme l'a souligné Beard (1989).

Sélection des objets Le modèle développé doit permettre d'estimer l'imprécision des mesures géométriques sur :
– l'ensemble des objets d'une base de données géographiques,
– une sélection d'objets, définie manuellement ou suite à un traitement automatique.
Nous chercherons cependant à faire en sorte que les expérimentations réalisées s'orientent vers des cas d'utilisation précis. Par exemple, pour une classe d'objets de tronçons de route, le modèle sera exploité en aval d'un calcul d'itinéraire (donc sur une sélection d'objets), afin d'évaluer l'imprécision de longueur de l'itinéraire calculé. Ce cas d'utilisation étant de nos jours extrêmement fréquent dans de nombreuses applications. Il demeurera également possible d'exploiter le modèle sur l'ensemble de la base de données, ce qui constitue également un cas d'utilisation classique, pour un gestionnaire de réseau par exemple. Ces conditions d'utilisation impliquent le déploiement d'une interface homme-machine, permettant à l'utilisateur de sélectionner aisément les objets qu'il souhaite évaluer.

2.2.1.2 Identification des impacts à évaluer

Contrairement aux modèles d'évaluation présentés précédemment, qui proposent des méthodes générales d'estimation de l'imprécision des mesures de longueur et de surface, le modèle mis en œuvre dans le cadre de cette thèse a pour objectif d'estimer l'impact de différents processus impliqués dans l'erreur de mesure finale. Deux processus ont particulièrement été soulignés : le processus d'abstraction, à travers les règles de représentation, et le processus de production (cf. figure 2.12).

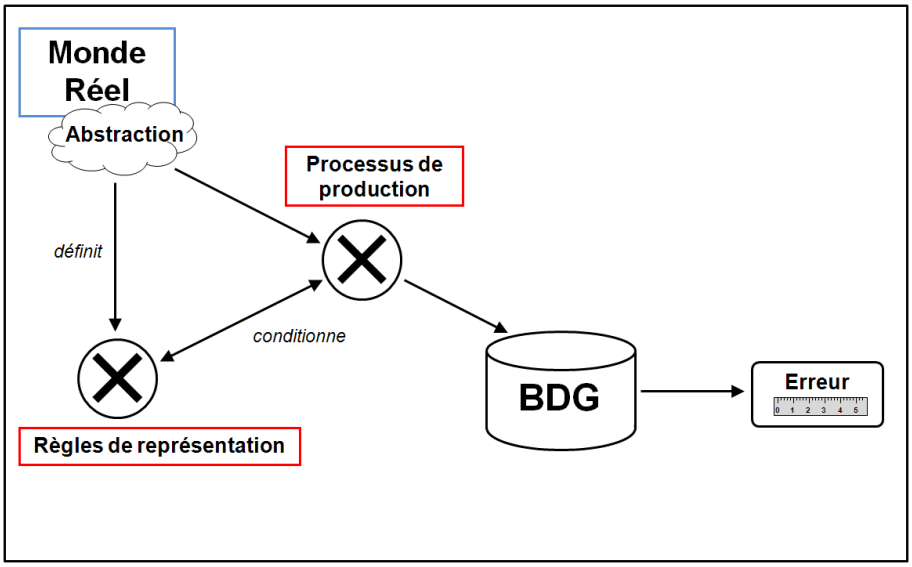


Figure 2.12 – Processus impliqués dans les erreurs de mesures géométriques

Ces deux processus affectent les mesures géométriques calculées à partir de la géométrie des objets vectoriels, il est donc nécessaire de modéliser leurs impacts respectifs dans l'erreur finale de mesure. Nous allons donc présenter dans cette section les différents impacts, induits par ces deux processus, que le modèle cherchera à évaluer. L'identification de ces impacts avait été proposée dans Girres et Ruas (2010).

Processus de production Comme nous l'avons vu dans le chapitre 1, de nombreux processus de production peuvent être mis en œuvre afin de saisir une base de données géographiques, en fonction du niveau de détail recherché, de l'étendue, et des moyens financiers disponibles. Suite à la présentation des erreurs liées à chaque processus de production, il est cependant possible de différencier plusieurs "types d'erreurs" modélisables qui affectent la géométrie des objets vectoriels. Nous en avons identifié deux, dont l'impact sur les mesures géométriques sera intégré au modèle :

1. **l'erreur de pointé**, qui concerne différents processus de production déjà cités auparavant : les relevés topographiques, la restitution photogrammétrique ou encore la digitalisation manuelle.
2. **la généralisation cartographique**, qui se différencie totalement des autres processus de par la diversité de ses impacts potentiels sur la géométrie des objets.

Comme il est possible de le constater, l'impact de l'imprécision des acquisitions par relevés GNSS n'a pas été pris en compte dans le cadre de l'élaboration du modèle, étant donné la difficulté de modéliser ses effets sans avoir connaissance de l'ensemble des éléments générateurs d'erreurs, notamment ceux liés au capteur et à son environnement direct.
De la même manière, l'impact des relevés Lidar et du traitement d'images de télédétection ne seront pas abordés dans le cadre de l'élaboration du modèle d'évaluation, par manque de recul sur les modèles induits par ces types de processus de production.

Règles de représentation Comme évoqué dans la section traitant des limites des modèles existants, les impacts des règles de représentation affectant les mesures géométriques de longueur et de surface ne sont généralement pas pris en compte. Dans le cadre de ce travail, nous traiterons de l'impact de trois règles de représentation déjà mentionnées précédemment :

1. **la projection cartographique**, qui se manifeste par une sur ou sous-estimation des mesures géométriques de longueur et de surface.
2. **la non-prise en compte du terrain**, qui génère une sous-estimation systématique des mesures géométriques de longueur et de surface.
3. **l'approximation polygonale des courbes**, qui sous-estime systématiquement les mesures géométriques de longueur, lorsque les entités du monde réel à représenter sont courbes.

Nous nous limiterons à ces trois règles de représentation dans le cadre du développement du modèle. Bien évidemment, le processus d'abstraction se manifeste par d'autres règles qui peuvent impacter les géométries produites (le choix de modéliser une route par une polyligne plutôt que par un polygone par exemple), mais celles-ci ne seront pas prises en compte dans le cadre du développement du modèle.

La modélisation des différents impacts des processus de production et des règles de représentation que nous venons de lister, constituent les **fonctions du modèle** que nous chercherons à développer afin d'estimer l'imprécision des mesures géométriques. Le développement de ces fonctions, qui constitue le cœur de ce travail de thèse, sera présenté dans les chapitres 3 et 4. Ainsi, après avoir identifié l'ensemble des processus impactant les mesures géométriques de longueur et de surface, nous allons présenter dans la section suivante les résultats attendus par le modèle ainsi proposé.

2.2.1.3 Communication des résultats

En considérant que les mesures de longueur et de surface calculées à partir de la géométrie des objets vectoriels peuvent être impactées par différentes contributions (les processus de production et les règles de représentation que nous venons de lister), différentes stratégies de communication de l'erreur de mesure peuvent être proposées.

Dans un premier temps, les contributions à l'erreur finale peuvent être communiquées indépendamment les unes des autres. Mais afin de proposer une information utile dans le cadre d'un processus d'aide à la décision, il s'avèrera nécessaire dans un second temps de communiquer à l'utilisateur une estimation globale de l'imprécision de la mesure calculée, à partir des impacts modélisés de manière indépendante.

Détail des erreurs par sources Etant donné que plusieurs sources potentielles d'erreurs impactent les mesures géométriques, il peut-être possible de proposer à l'utilisateur une estimation de chaque contribution sous forme de tableau de bord. Une modélisation de l'impact de chaque contribution permettra d'évaluer leur part respective (cf. figure 2.13). Si ce type de communication de l'erreur n'est pas le plus pertinent dans le cadre d'un processus d'aide à la décision, il peut par contre porter une dimension pédagogique très intéressante afin de sensibiliser l'utilisateur sur le poids que porte chaque contribution dans l'erreur finale.

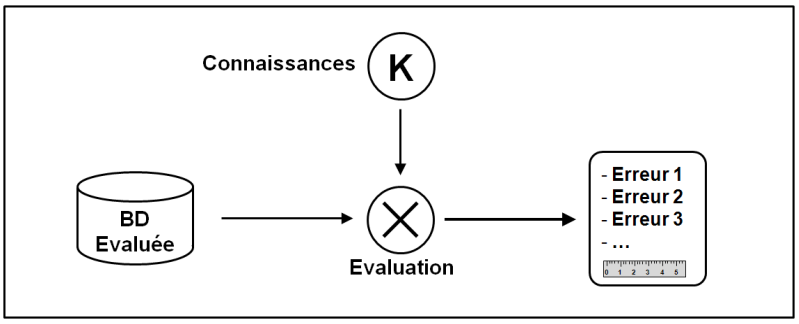


Figure 2.13 – Un processus d'évaluation détaillant chaque contribution à l'imprécision de la mesure

Combinaison des erreurs La finalité de ce travail de thèse réside dans la possibilité de proposer à l'utilisateur une estimation de l'imprécision des mesures géométriques de longueur ou de surface, en prenant en considération l'ensemble des composantes de l'erreur finale. La communication d'une estimation globale demeure la forme la plus pertinente dans le cadre d'un processus d'aide à la décision. Cependant, la diversité des sources potentielles d'erreurs impliquées met en évidence la complexité de leur combinaison afin de produire un indicateur global. Ainsi, différents scénarios de combinaison des erreurs peuvent être identifiés :

- une agrégation de l'ensemble des contributions,
- une sélection des contributions, en ne prenant en compte que les plus significatives,
- une combinaison de l'ensemble des contributions.

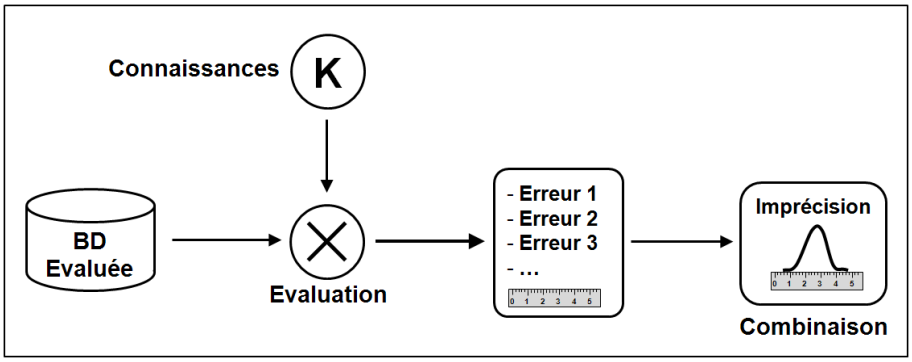


Figure 2.14 – Un processus d'évaluation produisant une estimation agrégée de l'imprécision de la mesure

L'objectif final est de pouvoir communiquer à l'utilisateur une information exprimée sous forme d'imprécision de mesure, en prenant en compte les impacts des composantes de l'erreur identifiées (cf. figure 2.14). L'objectif est de pouvoir proposer à l'utilisateur une information formulée de la manière suivante : *"La mesure calculée initialement est de X mètres. Cependant, nous considérons qu'elle est plutôt comprise entre (X+A) mètres et (X+B) mètres"*.

Synthèse Section Le modèle d'estimation de l'imprécision des mesures géométriques se définit comme un outil d'aide à la décision, facile à prendre en main pour l'utilisateur final. Ce modèle a pour objectif d'estimer l'impact des processus de production et des règles de représentation sur les mesures géométriques sans données de référence. La communication des résultats est envisagée sous deux formes : chaque contribution à l'erreur finale est tout d'abord quantifiée de manière indépendante, puis une estimation globale de l'imprécision de mesure est ensuite proposée à l'utilisateur.

2.2.2 Fonctionnement du modèle

Le modèle d'estimation de l'imprécision des mesures géométriques est pensé comme un outil d'aide à la décision. Il est donc souhaitable que cet outil soit suffisamment intuitif pour être assimilé facilement par un utilisateur, nécessitant peu de pré-requis, que ce soit au niveau de son paramétrage ou encore du formatage des données à évaluer. Cette section a pour objectif de présenter le principe général de fonctionnement du modèle, en définissant la démarche d'évaluation de l'imprécision des mesures, et en proposant les bases d'une interface homme-machine permettant de la mettre en œuvre simplement.

2.2.2.1 Principe d'évaluation

Pour que l'estimation de l'imprécision des mesures géométriques de longueur et de surface soit rendue accessible à un utilisateur, deux étapes nécessaires à l'évaluation ont été identifiées : la sélection des impacts à évaluer et le paramétrage du modèle.

Sélection des impacts à évaluer

Différentes contributions, imputables aux processus de production et aux règles de représentation, s'accumulent pour générer une erreur de mesure. Nous proposons dans le cadre de l'évaluation de modéliser ces impacts de manière séparée dans un premier temps, puis de les combiner. La question principale repose sur l'identification des processus à modéliser en fonction de la base de données à évaluer.

A priori, les impacts de la projection cartographique ou de la non-prise en compte du terrain (sauf si les données disposent d'une information altimétrique) doivent systématiquement être pris en compte pour les classes d'objets que nous souhaitons évaluer avec ce modèle (des objets de référence). Cependant, ceci n'est pas systématique en ce qui concerne l'approximation polygonale, car les objets évalués ne représentent pas toujours des entités courbes du monde réel. C'est également le cas pour les processus de production, qui ne sont pas tous mobilisés simultanément.

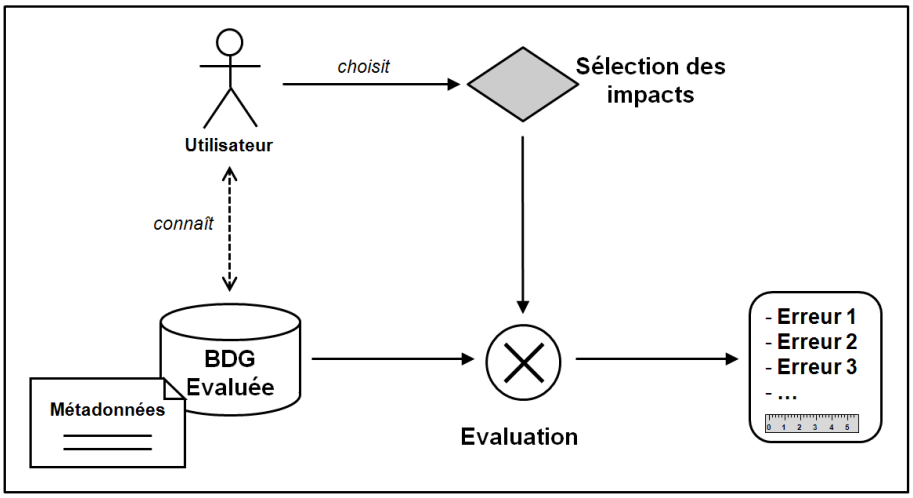


Figure 2.15 – Sélection des impacts à modéliser par l'utilisateur

Etant donné la difficulté de pouvoir identifier automatiquement les processus de production mobilisés pour la saisie d'une base de données géographiques, nous avons donc fait le choix dans un premier temps de laisser libre l'utilisateur de sélectionner les impacts qu'il souhaite évaluer. Cela suppose que l'utilisateur connaisse les processus de production utilisés pour produire la base de données, mais cette information est généralement renseignée dans la partie "Généalogie" des métadonnées (cf. figure 2.15). Il est envisageable, à terme, d'automatiser la recherche des processus impliqués dans les métadonnées, mais ceci nécessite préalablement un formatage de l'information relative au processus de production concerné.

Une fois les impacts à modéliser sélectionnés, une seconde étape est nécessaire afin de permettre l'évaluation de l'imprécision des mesures géométriques, à savoir l'acquisition des informations nécessaires au paramétrage du modèle.

Paramétrage du modèle

Le modèle développé dans le cadre de cette thèse repose sur un principe d'évaluation sans données de référence. Cela implique qu'à défaut de réaliser des comparaisons de longueur et de surface, le modèle doit s'appuyer sur des informations spécifiques à la modélisation de chaque impact sélectionné par l'utilisateur. Or, étant donné la diversité des processus pouvant être potentiellement impliqués dans l'erreur finale de mesure, différents types d'informations doivent être mobilisés. On peut les distinguer en deux catégories : les données externes et les connaissances sur les données évaluées.

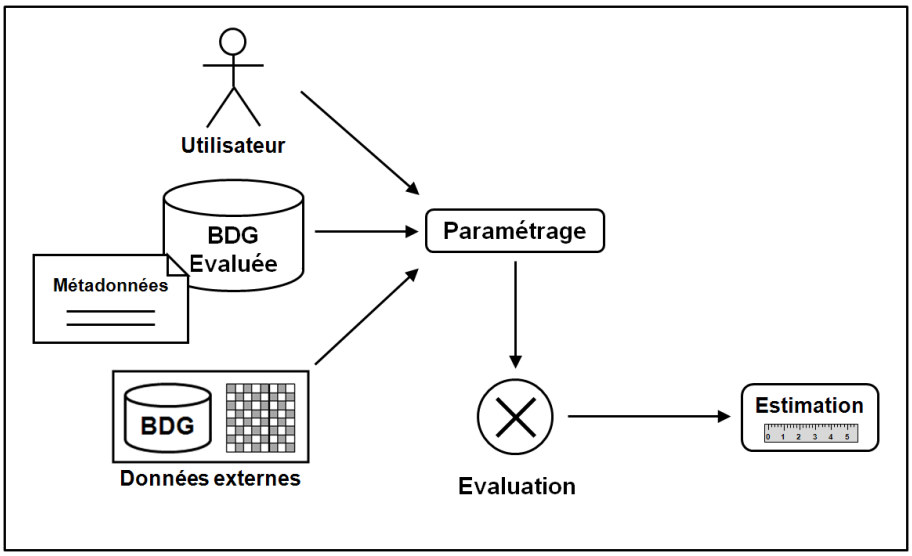


Figure 2.16 – Données externes et connaissances nécessaires au paramétrage du modèle

Données externes Différents types de données externes s'avèrent nécessaires pour permettre le bon fonctionnement de l'évaluation. A titre d'exemple, pour estimer l'impact de la non-prise en compte du terrain, si la base de données utilisée ne dispose pas d'information altimétrique, il est indispensable de disposer d'un modèle numérique de terrain disponible et de l'intégrer (ce qui est aujourd'hui possible sur l'ensemble de la surface terrestre). Nous verrons par la suite que ces données externes ne se limitent pas qu'aux modèles numériques de terrain.

Connaissances sur les données Différentes informations, que l'on peut qualifier de connaissances sur les données, sont également nécessaires pour paramétrer le modèle, et ainsi assurer son bon fonctionnement. Par exemple, la modélisation de l'impact de l'erreur de pointé suppose de connaître l'échelle de saisie moyenne. Si l'utilisateur ne la connait pas, il est nécessaire d'envisager le développement de méthodes permettant de l'estimer. Nous considérons donc que les connaissances sur les données peuvent être issues de trois sources principales :
– de l'utilisateur, à travers ses connaissances de la base de données,
– des métadonnées, qui peuvent fournir des informations utiles au paramétrage,
– de l'exploration des données, en développant des méthodes de qualification adaptées.

L'ensemble des données externes et des connaissances nécessaires au paramétrage du modèle seront précisées plus tard dans cette thèse. Une fois ces informations disponibles, le paramétrage doit pouvoir être réalisé via une interface homme-machine facile d'accès.

2.2.2.2 Interface homme-machine

Le modèle d'évaluation ainsi conçu doit pouvoir être utilisé facilement par un utilisateur de données géographiques. La meilleure solution réside dans une intégration du modèle sous forme d'extension (ou *plug-in*) au sein d'une interface SIG classique, comme illustré dans la figure 2.17.

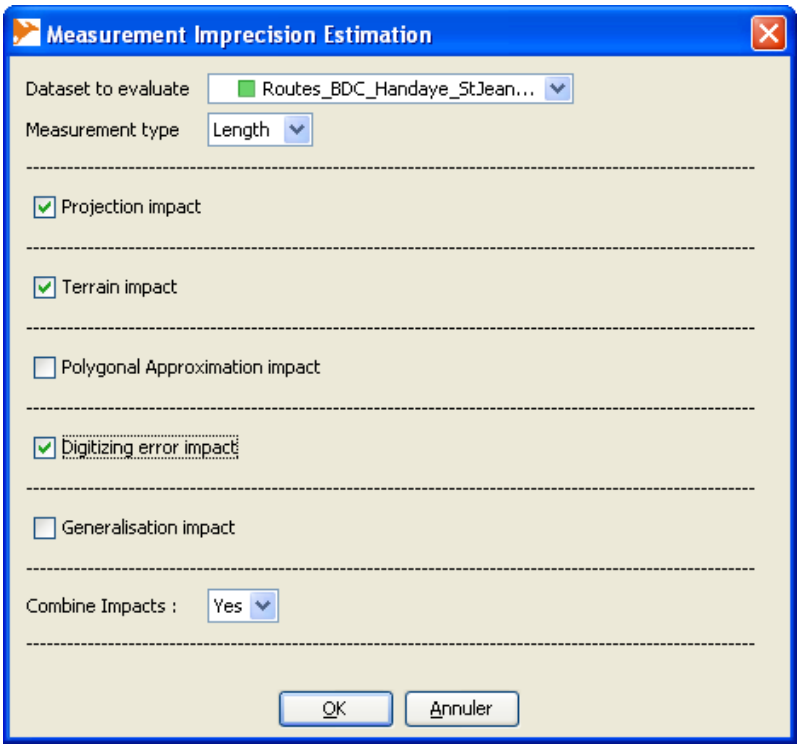


Figure 2.17 – Un modèle d'évaluation intégré dans une interface SIG

L'interface homme-machine proposée dans cette figure est pour le moment vide. Cette interface doit permettre, une fois complétée, de faciliter la sélection des différents impacts sur les mesures géométriques à évaluer, ainsi que leur paramétrage.

En mode opérationnel, l'évaluation de l'imprécision des mesures géométriques est donc prévue en trois temps :
- Etape 1 : Sélection des objets à évaluer, par l'opérateur ou un traitement automatique,
- Etape 2 : Sélection des impacts à modéliser, paramétrage des fonctions et intégration des données externes,
- Etape 3 : Lancement de l'évaluation et communication des résultats.

Cette section a permis de poser les bases du fonctionnement du futur modèle d'estimation de l'imprécision des mesures géométriques, en précisant le principe général d'évaluation et les interfaces homme-machine envisagées. La section suivante permettra de présenter l'approche utilisée pour sa modélisation.

Synthèse Section Le fonctionnement du modèle est envisagé sous la forme d'extension à une interface SIG classique, afin de faciliter l'interactivité homme-machine. Le modèle ainsi développé doit permettre à l'utilisateur de sélectionner librement les impacts à évaluer, et d'intégrer aisément les informations nécessaires à son paramétrage.

2.2.3 Modélisation par approche orientée-objet

Les grandes lignes du fonctionnement du modèle d'estimation étant à présent définies, nous allons exposer dans cette section les principes de modélisation utilisés pour sa conception.

L'approche Orientée Objet Une approche par modélisation orientée-objet sera mise en œuvre dans le cadre de la conception du modèle, en utilisant le langage UML (*Unified Modeling Language*), proposé par Booch *et al.* (2000) pour sa formalisation. Les principes généraux de la modélisation orientée-objet sont présentés en Annexe C.

Dans le cadre du développement du modèle d'estimation, le choix de l'approche orientée-objet s'est imposé puisque cette approche permet de modéliser assez fidèlement les différents concepts touchant à l'imprécision des mesures géométriques. Nous présenterons dans cette section ces différents concepts, intégrés au sein du modèle général d'estimation.

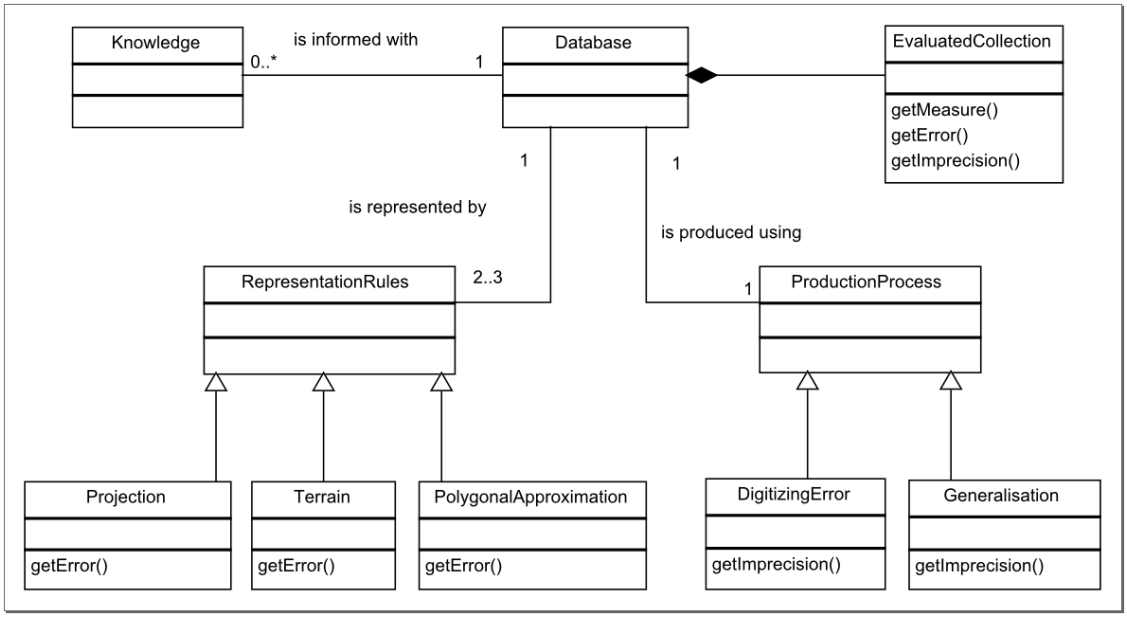


Figure 2.18 – Diagramme de classes de haut niveau du modèle d'estimation de l'imprécision des mesures géométriques

Concepts Nous avons jusqu'à présent évoqué un ensemble de phénomènes impactant les mesures de longueur et de surface calculées à partir de la géométrie des objets d'une base de données géographiques. Ces différents phénomènes peuvent être classés en deux catégories principales : les phénomènes relatifs aux processus de production et ceux relatifs aux règles de représentation.

Uniquement deux processus de production seront implémentés au sein du modèle d'estimation, à savoir les impacts de l'erreur de pointé et de la généralisation cartographique. Ces impacts se traduisent par une imprécision de mesure. Pour les règles de représentation, trois impacts ont été identifiés : projection cartographique, non-prise en compte du terrain, et approximation polygonale des courbes.

Ces différentes sources d'imprécision affectent les mesures géométriques calculées à partir d'un ensemble ou d'une sélection d'objets d'une base de données géographiques, dont on cherchera à acquérir un ensemble de connaissances utiles pour l'évaluation. L'évaluation se traduit par une estimation de l'imprécision totale de la mesure géométrique effectuée, calculée à partir des impacts sélectionnés pour chaque processus de production ou règle de représentation associés au jeu de données.

La figure 2.18 illustre l'organisation générale des classes d'objets de haut niveau représentant l'ensemble des concepts impliqués dans l'estimation de l'imprécision des mesures géométriques. Ce principe de modélisation avait été proposé dans Girres (2011b).

La suite de ce travail réside dans le développement de méthodes appropriées afin d'estimer l'impact des différentes règles de représentation et processus de production impliqués dans l'erreur finale de mesure. Ces méthodes constituent les fonctions du modèle et seront développées aux chapitres 3 et 4. La problématique de la combinaison des impacts, permettant de proposer à l'utilisateur une estimation globale de l'imprécision de mesure géométrique impliquée par les différents processus concernés, sera quant à elle présentée au chapitre 5.

Synthèse Section L'approche orientée-objet est utilisée pour le développement du modèle d'évaluation. Cette approche permet de modéliser assez fidèlement les différents concepts touchant à l'imprécision des mesures géométriques.

Conclusion du chapitre

Ce chapitre a tout d'abord permis de dresser un constat relatif aux moyens disponibles afin d'évaluer et de communiquer l'information sur l'imprécision géométrique des données vectorielles.
Pour un utilisateur, le principal moyen disponible aujourd'hui demeure les métadonnées, qui en plus de leur complexité, ne proposent que d'exprimer une information relative à l'exactitude de position des objets. On voit donc bien que l'ensemble des besoins potentiels d'un utilisateur de données géographiques en termes d'information d'imprécision géométrique ne sont pas pris en compte, et en particulier l'information concernant l'imprécision des mesures de longueur et de surface réalisées à partir de la géométrie des objets. Bien que certains modèles aient été développés en recherche ces dernières années afin d'estimer cette imprécision, ils présentent un certain nombre de limites, tant en termes de fiabilité quant à la modélisation de l'impact des processus de production impliqués, qu'à la prise en compte des règles de représentation des bases de données géographiques.

Ce constat nous a donc permis de positionner ce travail de thèse, sur la conception d'un modèle d'estimation de l'imprécision des mesures géométriques de longueur et de surface. Le modèle proposé est pensé comme un outil d'aide à la décision, développé dans une optique de réponse à des questions d'évaluation de la qualité externe. Son objectif réside dans l'estimation par un utilisateur de l'imprécision des mesures géométriques. Bien que paraissant simple, répondre à cette question suppose la prise en compte de processus complexes et hétérogènes. Nous chercherons donc dans le cadre du développement du modèle d'estimer les impacts de processus de production et de règles de représentation, puis de les combiner, afin de fournir à l'utilisateur une estimation globale.

Le modèle proposé doit être facile à prendre en main pour un utilisateur non-expert, et impliquer peu de pré-requis pour son paramétrage. Les principes régissant la modélisation et le fonctionnement de cet outil d'aide à la décision ont donc été exprimés dans ce sens.

Chapitre 3

Méthodes d'estimation de l'impact des règles de représentation

Introduction

Nous avons vu précédemment que les mesures géométriques de longueur et de surface sont affectées par différents impacts relevant :
- du processus de production des objets géographiques,
- du processus d'abstraction, définissant les règles de représentation des entités du monde réel sous forme d'objets.

Dans le cadre de l'élaboration d'un modèle permettant à un utilisateur d'estimer l'imprécision de mesures géométriques, nous avons évoqué la nécessité de prendre en compte les impacts de tous les processus pouvant potentiellement affecter ces mesures. Ainsi, nous allons présenter dans ce chapitre les différentes méthodes mises en œuvre afin de modéliser les impacts des règles de représentation sur les mesures géométriques de longueur et de surface. Nous présenterons successivement les méthodes développées afin de modéliser l'impact :
- de la projection cartographique,
- de la non-prise en compte du terrain,
- de l'approximation polygonale des courbes.

Nous avons bien conscience que dans la pratique, ces impacts s'accumulent à ceux des processus de production. Cependant, nous traiterons ces deux processus de manière indépendante. Les méthodes permettant de modéliser les impacts des processus de production seront par conséquent présentées au chapitre 4.

Différentes approches seront utilisées afin de modéliser les impacts des règles de représentation sur les mesures géométriques. Des méthodes d'enrichissement de la géométrie utilisant des sources de données externes, avec application de mesures adaptées, seront utilisées afin d'évaluer les impacts de la projection cartographique et de la non-prise en compte du terrain. Enfin, des méthodes d'interpolation seront exploitées afin de modéliser l'impact de l'approximation polygonale des courbes.

3.1 Projection cartographique

Le processus d'abstraction, permettant la représentation des entités du monde réel sous forme d'objets dans une base de données géographiques suppose, comme nous l'avons vu, un grand nombre de simplifications. On peut considérer que la représentation de la surface terrestre sur une carte en est une. En effet, par commodité, les objets géographiques sont représentés dans un plan, que ce soit pour des représentations cartographiques ou en visualisation par ordinateur, dans les logiciels SIG notamment. L'hypothèse sous-jacente est que localement la terre est plane (ce qui ne veut pas dire que les objets n'ont pas de volume). Pourtant, nous le savons bien, la surface terrestre présente une forme courbe irrégulière, et sa représentation dans un plan suppose la mise en œuvre de processus d'approximation, à l'aide d'un ellipsoïde, et de déformation, par le biais de projections cartographiques. Les déformations impliquées par ces projections cartographiques impactent la géométrie des objets, et par conséquent les mesures géométriques réalisées en aval.

Comme nous l'avons déjà évoqué auparavant, les modèles d'erreurs développés en recherche ne prennent généralement pas en compte l'impact de la projection cartographique sur les mesures géométriques. Ceci peut se comprendre à grande échelle, lorsque ces impacts sur les mesures géométriques ne sont que peu conséquents. Cependant, à petite voire moyenne échelle, ces impacts peuvent se révéler importants, générant des erreurs de mesures sensibles. Il convient donc dans ce cas de pouvoir les modéliser.

L'objectif de cette section consiste donc à proposer des méthodes permettant de quantifier l'impact de la projection cartographique sur les mesures géométriques de longueur ou de surface, afin d'être intégrées au sein du modèle général d'estimation. Pour cela, quelques rappels sur la représentation cartographique de la surface terrestre et ses impacts seront tout d'abord proposés, avant de présenter les méthodes mises en œuvre afin de quantifier ces impacts sur les mesures géométriques de longueur et de surface. Enfin, une expérimentation sera menée pour mettre en application les méthodes développées.

3.1.1 La représentation de la Terre et ses impacts

Cette section a pour objectif de présenter un certain nombre de généralités sur la représentation de la surface terrestre, et plus particulièrement sur le processus de projection cartographique, permettant de la représenter dans un plan. Ce préalable est indispensable afin de mieux comprendre l'impact que cette représentation plane peut avoir sur les mesures géométriques de longueur et de surface.

3.1.1.1 De la Terre à la carte : principes généraux

Représenter la surface terrestre sur un plan n'est pas un tâche triviale. En effet, la terre présente une forme irrégulière, soumise de surcroit à des dynamiques importantes. Ainsi, la représentation d'une "figure de la Terre" dans un plan suppose d'entreprendre un certain nombre de simplifications. Nous allons présenter dans cette section les principales formes d'approximation de la surface terrestre utilisées par les cartographes, permettant le calcul des coordonnées des lieux des objets géographiques. Nous nous intéresserons plus particulièrement aux coordonnées planes, impliquant la mise en œuvre de projections cartographiques.

Représentation de la surface terrestre

La détermination d'une figure représentative de la surface de la terre constitue l'une des tâches classiques de la discipline géodésique. Comme nous l'avons déjà précisé, la surface terrestre présente une forme irrégulière, et plusieurs modèles permettant de la représenter peuvent être utilisés (comme le Géoïde, l'Ellipsoïde ou encore le Topoïde), leur pertinence variant selon la discipline étudiée. L'un de ces modèles de représentation fait davantage office de figure de la Terre que les autres, il s'agit du Géoïde.

Le geoïde Le géoïde correspond à une surface équipotentielle du champ de gravité terrestre, c'est à dire où le potentiel de pesanteur est constant. Afin d'approximer au mieux la surface de la terre, le géoïde est défini selon un potentiel de gravité proche du niveau moyen des mers. Ce modèle de représentation offre l'avantage d'associer des propriétés physiques à une forme géométrique représentant la surface terrestre. Ainsi, le géoïde sert généralement de référence pour les mesures précises d'altitude, ainsi que dans de nombreuses applications, comme l'hydrologie ou l'aéronautique. Cependant, le géoïde n'est pas une surface régulière (cf. figure 3.1), du fait de l'inégale répartition des masses à l'intérieur et à la surface de la terre. Ainsi, dans de nombreuses disciplines (comme la cartographie), une représentation mathématique permettant d'approximer la surface terrestre lui est souvent préférée, l'ellipsoïde de révolution.

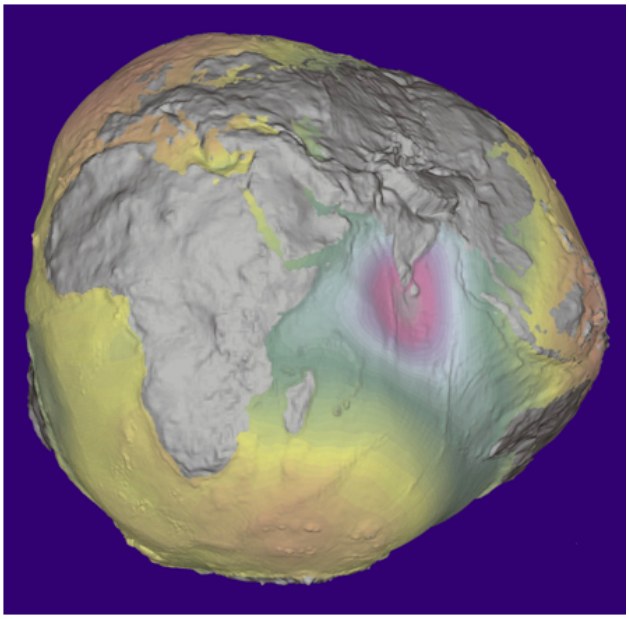


Figure 3.1 – Représentation exagérée du géoïde terrestre

L'ellipsoïde L'ellipsoïde de révolution est un modèle mathématique se présentant comme une sphère aplatie au niveau de ses pôles. Il est défini par la taille de son demi-grand axe et de son demi-petit axe (cf. figure 3.2). A l'inverse du géoïde, l'ellipsoïde présente une forme régulière, ce qui constitue une approximation conséquente de la surface terrestre, mais beaucoup plus facilement manipulable.

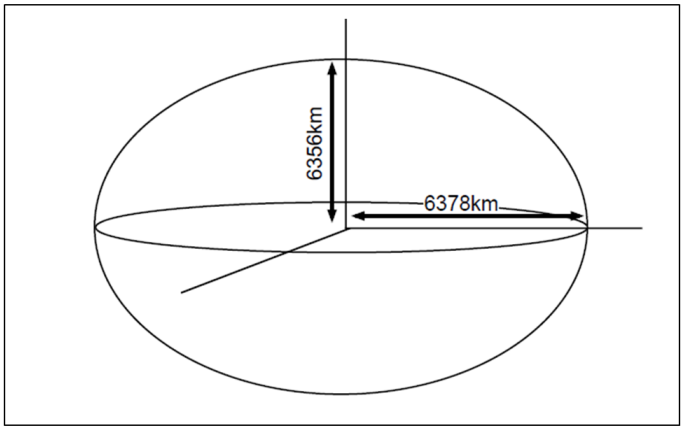


Figure 3.2 – Un ellipsoïde défini par son demi-grand axe et son demi-petit axe

Pour représenter la surface terrestre, un grand nombre d'ellipsoïdes sont en usage, comme les ellipsoïdes de Clarke 1880, Hayford 1909, ou encore IAG-GRS80 (cf. figure 3.3). Le choix de l'ellipsoïde est généralement guidé de manière à ce que celui-ci approxime du mieux possible le géoïde terrestre. Par exemple en France, l'ellipsoïde Clarke 1880 a longtemps été utilisé, puisqu'il permettait de bien approximer localement le territoire métropolitain. Il a été remplacé en 2000 par l'ellipsoïde IAG-GRS80, afin d'être compatible avec le système GPS. Enfin, on différencie les ellipsoïdes globaux, définis afin de représenter l'ensemble de la surface terrestre (comme l'IAG-GRS80), des ellipsoïdes locaux, définis afin de représenter au mieux une région d'intérêt.

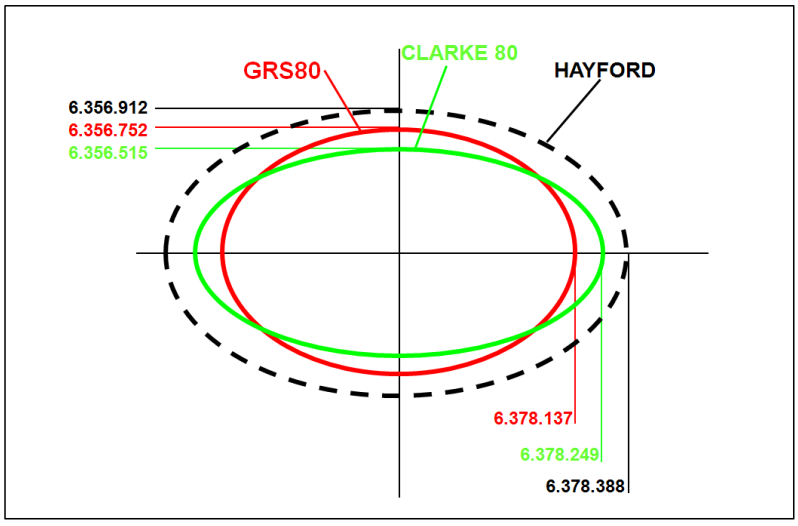


Figure 3.3 – Dimensions des ellipsoïdes Clarke 1880, Hayford 1909, et IAG-GRS80

Ainsi, le choix d'un ellipsoïde permettant d'approximer du mieux possible la surface terrestre constitue un préalable indispensable au calcul des coordonnées, qui permettront la saisie de la géométrie des objets géographiques, ainsi que les mesures géométriques de longueur ou de surface.

Calcul de coordonnées

Pour pouvoir calculer les coordonnées des points constituant la géométrie des objets géographiques, on utilise généralement des coordonnées dites planes (qui supposent que la surface de la terre soit localement plane, mais pas les objets localisés sur cette surface). Le calcul de coordonnées planes suppose d'associer un système de référence terrestre à l'ellipsoïde dans un premier temps, puis une projection cartographique dans une deuxième temps.

Système de référence Un système de référence terrestre est un repère cartésien tridimensionnel (O, x, y, z), défini par la position de son centre O et l'orientation de ses trois axes orthonormés Ox, Oy et Oz (cf. figure 3.4). Chaque point est ici localisé par ses **coordonnées cartésiennes**, c'est-à-dire un triplet de coordonnées (X,Y,Z). Historiquement, de nombreux systèmes de référence terrestres ont été crées en fonction des besoins locaux.

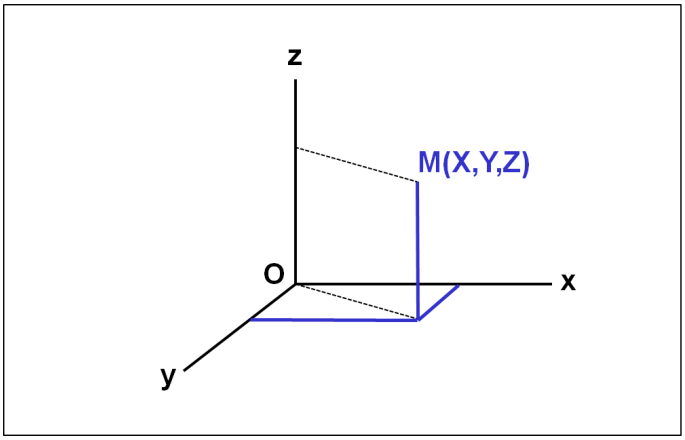


Figure 3.4 – Calcul de coordonnées cartésiennes dans un système de référence terrestre

Le **datum géodésique** est défini en associant un ellipsoïde au système de référence terrestre. Pour les systèmes globaux (ou géocentriques), le datum géodésique est constitué des paramètres de l'ellipsoïde, de la position du centre de l'ellipsoïde par rapport au centre de masse de la terre, de l'orientation des axes de l'ellipsoïde, et du méridien d'origine. Pour les systèmes locaux, les mêmes paramètres sont utilisés, en y ajoutant le point fondamental, c'est-à-dire le point où l'ellipsoïde tangente le géoïde.

En France Métropolitaine, le système de référence géodésique officiel est le RGF93, basé sur l'ellipsoïde IAG-GRS80, avec comme méridien d'origine le méridien de Greenwich. Il a succédé en 2001 au système NTF (Nouvelle Triangulation de la France), basé sur l'ellipsoïde Clarke 1880-IGN, avec comme point fondamental la croix du Panthéon, et le méridien de Paris comme méridien d'origine.

L'association d'un ellipsoïde à un système de référence rend possible le calcul des **coordonnées géographiques** de tout point de la surface terrestre. Les coordonnées géographiques sont exprimées en valeurs angulaires par la latitude ϕ et la longitude λ, ainsi que par une hauteur ellipsoïdale h (cf. figure 3.5).

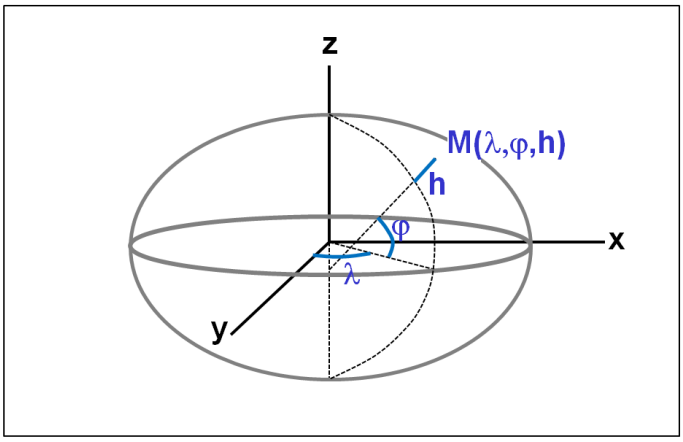


Figure 3.5 – Principe de calcul de coordonnées géographiques

Afin de passer à une représentation plane de la surface terrestre, et ainsi permettre le calcul des **coordonnées planes**, il convient d'associer au système de référence et à l'ellipsoïde une projection cartographique. En effet, il est mathématiquement beaucoup plus simple de calculer une distance sur un plan que sur une surface mathématique courbe, même si celle-ci est régulière. Nous allons donc exposer par la suite les principes généraux qui régissent la projection de tout ou partie de l'ellipsoïde de révolution dans un plan. Les différents types de projections cartographiques et leurs impacts y seront notamment exposés.

Projection cartographique

La représentation de toute ou partie de la surface terrestre sur une carte nécessite l'adoption d'une transformation permettant de passer de l'ellipsoïde à un plan. D'un point de vue mathématique, il s'agit donc de déterminer un couple de fonction (f, g) permettant de passer des coordonnées géographiques (λ, ϕ) aux coordonnées planes (E, N), comme présenté dans l'équation 3.1.

$$E = f(\lambda, \phi), N = g(\lambda, \phi) \tag{3.1}$$

Différentes projections cartographiques peuvent donc être utilisées afin de représenter un même ellipsoïde sur un plan. Ces projections peuvent être classifiées selon le type de déformations et selon le type de projection.

Types de déformations L'ellipsoïde n'est pas une surface développable. Cela signifie que la projection de l'ellipsoïde sur une surface plane le déforme nécessairement (Snyder, 1987). D'un point de vue mathématique, il n'est donc pas possible de construire une projection qui conserve à la fois les angles et les surfaces. Il est par contre possible de construire des projections :
- qui conservent les angles : les **projections conformes**,
- qui conservent les surfaces : les **projections équivalentes**,
- qui ne conservent ni les angles ni les surfaces : les **projections aphylactites**.

Pour visualiser ces déformations, l'indicateur de Tissot est généralement utilisé. Cet indicateur est construit à partir d'un cercle tracé sur la surface terrestre, dont on étudie l'homologue dans le plan suite à la projection. Dans le cas d'une projection équivalente, le cercle projeté est transformé en ellipse de même surface que le cercle initial. Dans le cas d'une projection conforme, le cercle projeté demeure un cercle, de surface différente. Ce dernier cas de figure est illustré dans la figure 3.6 avec l'exemple de la projection cylindrique conforme de Mercator, à laquelle est associée un indicateur de Tissot.

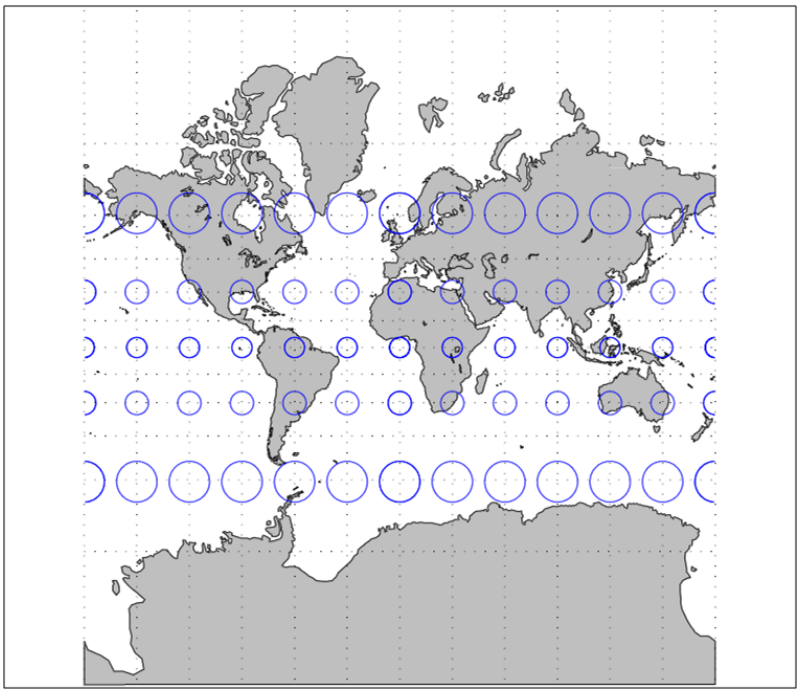


Figure 3.6 – Projection cylindrique conforme de Marcator et indicateur de Tissot, d'après Sillard (2000)

Types de projections Pour pouvoir passer à une représentation plane, l'ellipsoïde doit être projeté sur une surface développable, c'est à dire une surface qui peut être étalée sans déformation dans le plan. Pour cela, on utilise généralement trois formes mathématiques courantes qui répondent à ce critère, à savoir le plan, le cylindre et le cône, ce qui donne lieu aux trois types principaux de projections suivants :

- les **projections cylindriques**, où l'ellipsoïde est projeté sur un cylindre (tangent en un cercle ou sécant en deux cercles),
- les **projections coniques**, où l'ellipsoïde est projeté sur un cône (tangent en un cercle ou sécant en deux cercles),
- les **projections azimutales**, où l'ellipsoïde est projeté sur un plan (tangent en un point ou sécant en un cercle).

Il existe également des projections qui ne peuvent être classées dans aucun de ces types. Celles-ci sont appelées projections uniques.

En France Métropolitaine, la projection officielle est la projection conique conforme Lambert93 (cf. figure 3.7). Elle est associée au système géodésique français RGF93 et est sécante à l'ellipsoïde IAG-GRS80 aux parallèles 44 °N et 49 °N. Pour minimiser localement les erreurs impliquées par la projection Lambert93, les projections Lambert Coniques Conformes 9 zones ont également été créées. La projection Lambert93 a remplacé les projections coniques conformes Lambert quatre zones (et la projection Lambert II étendu) associées au système de référence NTF, et sécantes à l'ellipsoïde Clarke 1880-IGN.

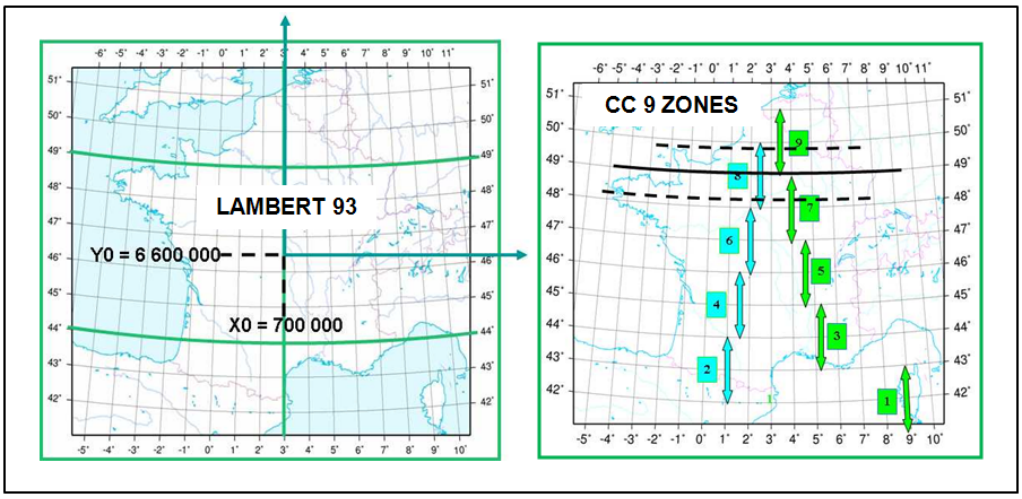


Figure 3.7 – Projections Lambert 93 et Coniques Conformes 9 zones

Comme nous pouvons le voir, le choix de la projection utilisée afin de représenter la surface de la Terre sur une carte génère des déformations qui impactent soit les angles, soit les surfaces, soit les deux. Dans tous les cas de figures, les projections ne préservent jamais les distances. Cette observation s'avère donc fondamentale dans le cadre de ce travail de thèse, d'autant plus qu'en France, les projections coniques conformes sont utilisées par convention. Ces projections ne préservent donc ni les mesures de longueur, ni les mesures de surface. Cependant ces projections conservent les angles, ce qui demeure très pratique, puisque de nombreux processus de saisie sont basés sur des mesures d'angles. Nous allons voir dans la section suivante comment estimer l'impact de la projection cartographique sur les mesures géométriques.

3.1.1.2 Impact sur les mesures géométriques

La géométrie des objets vectoriels est généralement saisie dans un système de projection bien défini. Chaque point composant la géométrie de l'objet dispose donc de coordonnées planimétriques E et N calculées dans un plan. Ces coordonnées sont donc affectées des altérations impliquées par la projection de l'ellipsoïde sur un plan, impactant par conséquent les mesures géométriques calculées en aval. En effet, une mesure de longueur calculée à l'aide de la distance euclidienne diffère de sa longueur calculée sur l'ellipsoïde (cf. figure 3.8). Dans la pratique, l'utilisateur ne s'en soucie guère, puisque localement l'impact de la projection cartographique sur des mesures géométriques est généralement très faible, d'autant plus que le système de projection est choisi afin de minimiser ces altérations. Cependant, à petite voire moyenne échelle, l'impact du système de projection peut s'avérer conséquent.

Il convient donc de mettre en place les méthodes permettant d'estimer cet impact sur les mesures géométriques.

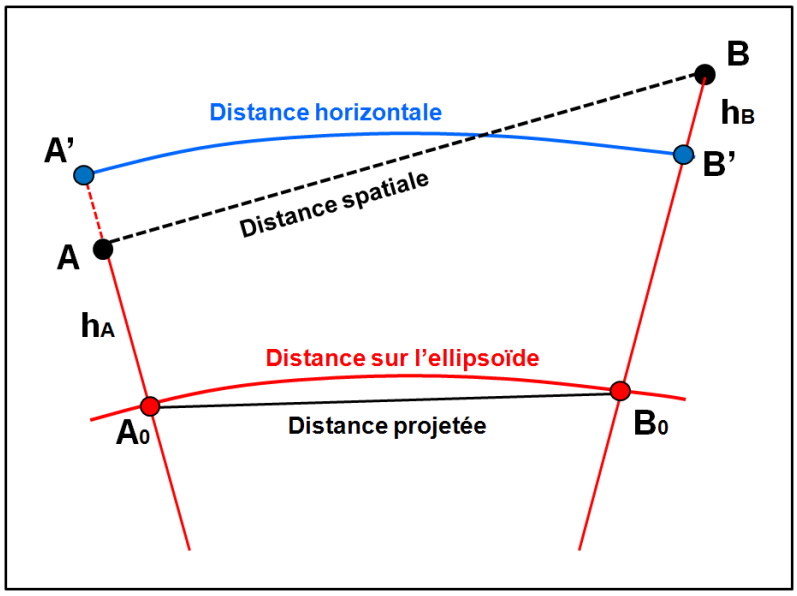


Figure 3.8 – Principe de calcul de distances sur ellipsoïde entre deux points A et B

La formalisation de l'impact de la projection cartographique sur les mesures géométriques proposée dans cette section est issue des travaux de Botton (2010), Bouteloup (2006) et Rüeger (1996).

Impact sur les mesures de longueur

Altération linéaire La notion d'altération linéaire permet d'exprimer l'impact d'un système de projection sur une mesure de longueur calculée dans un plan. L'altération linéaire e_L est calculée à l'aide du **module linéaire** m_L qui correspond au rapport entre une distance D_P calculée en projection et son homologue D_E sur l'ellipsoïde, comme exposé dans les équations 3.2 et 3.3.

$$m_L = \frac{D_P}{D_E} \tag{3.2}$$

$$e_L = m_L - 1 \tag{3.3}$$

Les systèmes de projection sont généralement choisis afin de minimiser l'altération linéaire, et ainsi permettre de réaliser des mesures de longueur les moins altérées possibles. Par exemple, la figure 3.9 présente l'altération linéaire en fonction de la latitude pour les différentes projections Lambert. On voit que pour la projection Lambert93, l'altération linéaire est nulle sur les parallèles automécoïques (longitudes 44°N et 49°N), et importante aux extrémités Nord et Sud du territoire. Cependant, l'altération linéaire de la projection Lambert93 est systématiquement inférieure à celle de la projection Lambert II étendu, qui pouvait atteindre 5 mètres par kilomètre dans le sud de la Corse (c'est-à-dire à la longitude 41°N).

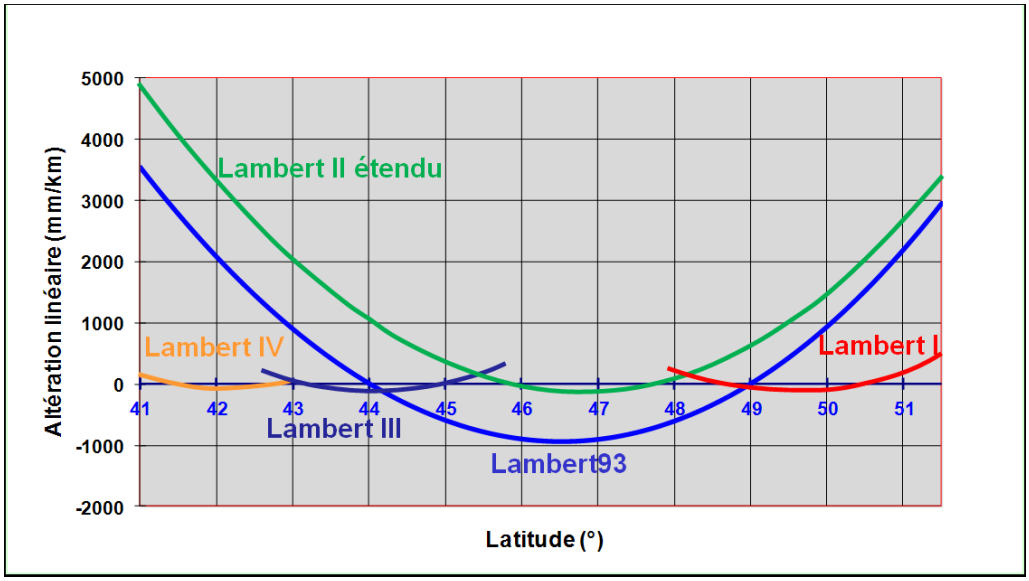


Figure 3.9 – Altération linéaire des projections Lambert I, II étendu, III, IV, et Lambert93 selon la latitude (source : IGN)

Distance sur ellipsoïde Etant donné que l'altération linéaire peut être connue en tous points d'un objet projeté, il est possible d'opérer à des calculs de longueur sur l'ellipsoïde de référence. Par exemple, pour un segment $[AB]$, de point médian C, dont on connait les valeurs d'altération linéaire e_A, e_B et e_C, l'altération linéaire moyenne e_{AB} est tout d'abord calculée, conformément à l'équation 3.4.

$$e_{AB} = \frac{e_A + 2e_C + e_B}{4} \tag{3.4}$$

La distance sur ellipsoïde D_E du segment $[AB]$ est calculée par application du module linéaire à la longueur dans le plan D_P (cf. équation 3.5).

$$D_E = \frac{D_P}{(1 + e_{AB})} \tag{3.5}$$

Bien que l'altération linéaire calculée à l'aide de cette méthode ne constitue qu'une moyenne des altérations linéaires de ses extrémités et de son centre, cette méthode permet malgré tout d'approximer une mesure de longueur sur l'ellipsoïde de référence, et ainsi d'estimer l'erreur de longueur impliquée par la projection cartographique. En reproduisant cette méthode de calcul sur l'ensemble des segments composant la géométrie d'une polyligne, il est possible de calculer sa mesure géométrique de longueur sur l'ellipsoïde.

Ainsi, pour une polyligne L composée des points L_1, $L_2 ... L_n$, nous proposons de calculer la longueur géométrique sur ellipsoïde $D_E(L)$ par le biais de l'équation 3.6.

$$D_E(L) = \sum_{i=1}^{n} D_E(L_i) \tag{3.6}$$

Distance à l'horizon Nous avons pour l'instant présenté une méthode permettant de calculer la longueur d'une polyligne à la surface de l'ellipsoïde. Cependant, lorsque l'altitude des points composant la géométrie d'un objet vectoriel est connue, il n'est plus pertinent de calculer une longueur à la surface de l'ellipsoïde, mais plutôt à une hauteur au dessus de l'ellipsoïde. Cette distance est appelée distance à l'horizon ou distance horizontale. Pour calculer cette distance, nous considérons ici que l'altitude zéro des points composant la géométrie de l'objet est équivalente à la surface de l'ellipsoïde. La figure 3.8 illustre la différence entre la distance sur ellipsoïde D_0 calculée entre les points A et B, et la distance horizontale D_H.

Pour calculer la distance horizontale, l'altitude du point A, du point B ou l'altitude moyenne entre les points A et B peuvent être utilisées. L'équation 3.7 formalise le calcul de la distance horizontale D_H moyenne entre les points A et B, d'altitudes h_A et h_B à partir d'une distance sur ellipsoïde D_0 et d'une valeur $R = 6.378$ kilomètres, correspondant au rayon moyen de la Terre approximée à une sphère.

$$D_H = D_0 * (1 + \frac{h_A + h_B}{2R}) \tag{3.7}$$

Ainsi, lorsque l'altitude des points composant les segments L_1, $L_2 ... L_n$ d'une polyligne L est connue, nous proposons de calculer sa distance horizontale $D_H(L)$, par le biais de l'équation 3.8.

$$D_H(L) = \sum_{i=1}^{n} D_H(L_i) \tag{3.8}$$

Nous venons de présenter comment procéder au calcul de longueur sur, et à une hauteur au dessus de l'ellipsoïde de référence, afin d'estimer l'impact de la projection cartographique sur les mesures de longueur. Nous allons maintenant présenter comment estimer cet impact sur les mesures de surface.

Impact sur les mesures de surface

Altération surfacique L'altération surfacique est utilisée pour exprimer l'impact d'un système de projection sur une mesure de surface calculée dans un plan. L'altération surfacique e_S est calculée à l'aide du **module aréolaire** m_S qui n'est autre que le carré du module linéaire m_L, comme exprimé dans les équations 3.9 et 3.10.

$$m_S = m_L^2 \tag{3.9}$$

$$e_S = 2 * e_L + e_L^2 \tag{3.10}$$

L'exploitation de l'altération surfacique n'est utile que dans le cas des projections conformes (ou encore aphylactites), puisque dans le cas des projections équivalentes, les mesures de surface sont préservées. Cela signifie que pour les projections équivalentes, en tous points :
– l'altération linéaire est nulle,
– le module aréolaire est égal à 1.

Surface sur ellipsoïde Pour calculer une surface sur ellipsoïde, le principe est identique à celui présenté précédemment pour les mesures géométriques de longueur. Ainsi, pour un polygone de surface S_P dans le plan, et dont on connait l'altération surfacique e_{Si} en chaque point de la géométrie, une valeur moyenne d'altération surfacique e_S est tout d'abord calculée (cf. équation 3.11).

$$e_S = \frac{1}{n} \sum_{i=1}^{n} e_{Si} \tag{3.11}$$

Une autre méthode pour calculer l'altération surfacique moyenne du polygone consiste à récupérer la valeur d'altération surfacique de son point centroïde. Une fois l'altération surfacique moyenne du polygone calculée, il est ensuite possible de calculer sa surface sur ellipsoïde S_E, par application du module surfacique (cf. équation 3.12).

$$S_E = \frac{S_P}{(1 + e_S)} \tag{3.12}$$

Tout comme dans le cadre du calcul des mesures de longueur sur ellipsoïde, le calcul des mesures de surface sur ellipsoïde se base sur une moyenne des altérations surfaciques de chacun des points constituant la géométrie du polygone. Bien que cette moyenne ne constitue qu'une approximation, elle s'avère également suffisante afin de produire une estimation de l'erreur de mesure de surface impliquée par la projection cartographique. De la même manière, si l'altitude des points composant la géométrie d'un objet polygonal est connue, il est également possible de procéder à des mesures de surface au dessus de l'ellipsoïde.

Synthèse Section Nous avons montré dans cette section que la représentation de la surface terrestre sur un plan implique la réalisation d'approximations (en représentant la terre par un ellipsoïde) et de déformations (en projetant cet ellipsoïde sur un plan), qui impactent les mesures calculées à partir de la géométrie des objets vectoriels. Le principe de calcul des mesures géométriques de longueur et de surface sur l'ellipsoïde de référence ont été présentées, permettant ainsi d'estimer l'erreur impliquée par la projection cartographique.

3.1.2 Méthodes pour le calcul de mesures géométriques sur ellipsoïde

L'impact de la projection cartographique sur les mesures géométriques de longueur et de surface a été exposé dans la section précédente. Nous avons pu voir que les altérations varient en fonction du système de projection utilisé, mais que celles-ci peuvent être définies aisément par le biais de fonctions mathématiques. La connaissance de l'altération linéaire ou surfacique en chacun des points de la géométrie d'un objet vectoriel permet la réalisation de mesures de longueur ou de surface sur l'ellipsoïde de référence. Nous allons donc présenter dans cette section comment mettre en œuvre un processus permettant de calculer ces mesures géométriques sur ellipsoïde, et ainsi estimer l'erreur impliquée par la projection cartographique utilisée.

Différentes méthodes peuvent être utilisées pour estimer cet impact. Vincenty (1975) propose par exemple des formules permettant de calculer directement des distances sur ellipsoïde. Cependant, puisque nous cherchons à élaborer un modèle permettant d'estimer l'erreur impliquée par la projection cartographique sur les mesures de longueur, mais également de surface, nous proposons une méthode d'enrichissement de la géométrie des objets basée sur l'extraction d'informations à partir de grilles d'altération.

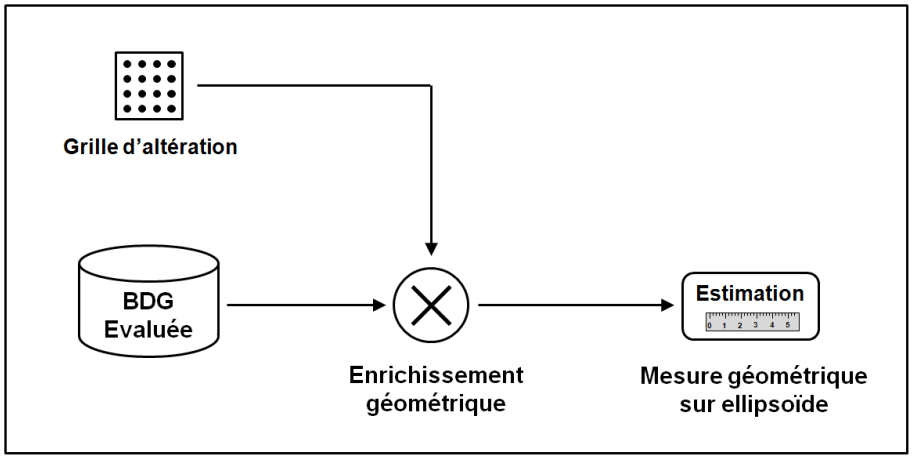


Figure 3.10 – Principe d'estimation de l'impact de la projection cartographique

Nous présenterons dans cette section les méthodes permettant de créer des grilles d'altération, puis de les exploiter afin d'enrichir la géométrie des objets évalués, et ainsi calculer des mesures sur ellipsoïde.

3.1.2.1 Création des grilles d'altération

Le préalable nécessaire à la réalisation de mesures géométriques sur ellipsoïde réside dans la connaissance de l'altération linéaire ou surfacique en chacun des points composant la géométrie des objets étudiés. L'élaboration de grilles d'altération est prévue à cet effet. Nous entendons ici par grille un carroyage régulier de points (vectoriels) renseignés d'une valeur d'altération. Nous allons voir dans cette section comment élaborer ces grilles d'altération. Nous nous appuierons pour cela sur le logiciel Circé, développé par l'Institut Géographique National (IGN-SGN, 2010).

Pour des raisons pratiques, les grilles d'altération ne sont réalisées que sur le territoire français métropolitain, mais le principe général demeure universel et peut être reproduit dans d'autres zones, à condition de disposer des paramètres de l'ellipsoïde et de la projection.

Création d'un carroyage régulier La première étape réside dans la création d'un carroyage régulier de points qui constitueront la future grille d'altération. Ce carroyage est réalisé en coordonnées géographiques (λ, ϕ) dans le système de référence RGF93. Le carroyage est compris entre les longitudes 5,5°O et 8,5°E et les latitudes 42°N et 51,5°N afin de couvrir l'ensemble du territoire métropolitain. Un pas régulier de 0,05°a été défini entre chaque point.

Conversion en coordonnées projetées La seconde étape consiste à convertir le carroyage produit en coordonnées géographiques vers des coordonnées projetées. Deux systèmes de projection "cibles" ont été sélectionnés :
- la projection Lambert93 (Conique Conforme), qui est le système de projection officiel en France métropolitaine (cf. figure 3.11)
- la projection UTM 31N (Cylindrique Conforme), qui est un système de projection utilisé au niveau mondial, en particulier dans les récepteurs GPS. Le fuseau 31 Nord étant le fuseau central en France métropolitaine.

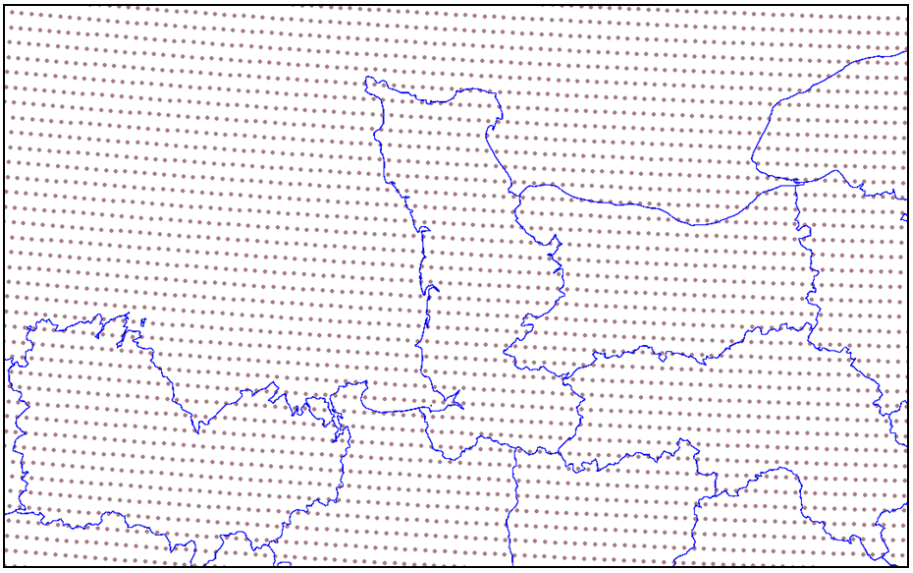


Figure 3.11 – Projection du corroyage régulier en Lambert93

Calcul de l'altération linéaire Le logiciel Circé, développé par l'Institut Géographique National, offre la possibilité de réaliser des conversions de coordonnées géographiques vers des coordonnées projetées, en fournissant pour chaque couple de coordonnées converti, son altération linéaire exprimées en mm/km. L'opération a donc été réalisée pour chacun des deux systèmes de projection sélectionnés, permettant ainsi l'élaboration de deux canevas de points renseignés de leur altération linéaire.

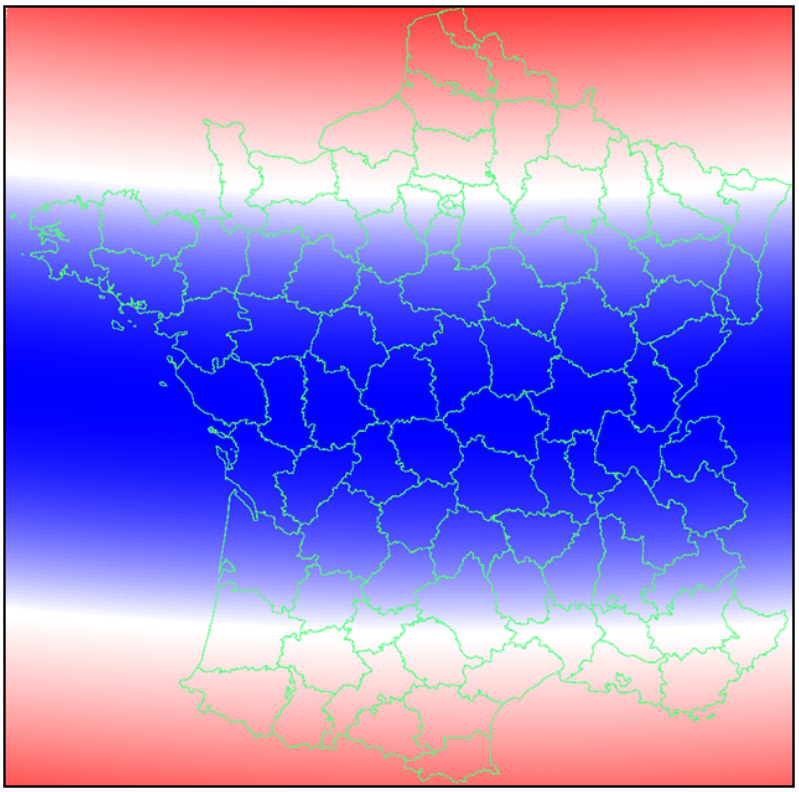


Figure 3.12 – L'altération linéaire de la projection Lambert93 (positive en rouge et négative en bleu)

La figure 3.12 représente l'altération linéaire de la projection Lambert93 sur le territoire français. Plus la couleur du carroyage vire vers le rouge, plus l'altération est positive (nord et sud de la France), tandis que plus la couleur du carroyage vire vers le bleu, plus l'altération linéaire est négative (c'est-à-dire vers 42,5°N).

La connaissance de l'altération linéaire sur chacun des points de la grille permet également de produire les deux grilles d'altération surfacique en utilisant la conversion exprimée dans l'équation 3.10.

3.1.2.2 Mesures sur ellipsoïde par enrichissement de la géométrie

Une fois les grilles d'altération produites, il est possible d'estimer l'altération linéaire ou surfacique en tous points, et ainsi procéder à des mesures de longueur et de surface sur ellipsoïde. Cependant, pour pouvoir enrichir la géométrie des objets vectoriels d'une information d'altération, il est préalablement nécessaire d'opérer à un suréchantillonnage des géométries.

Suréchantillonnage de la géométrie Le suréchantillonnage de la géométrie des objets évalués concerne principalement les **objets surfaciques**. En effet, comme expliqué dans l'équation 3.11, une valeur moyenne d'altération surfacique e_S est utilisée afin de réaliser des mesures géométriques de surface sur ellipsoïde.

Cependant, lorsque le polygone évalué présente une grande superficie, exploiter une valeur moyenne d'altération surfacique n'est pas une solution satisfaisante, puisque la valeur d'altération ne suit pas une fonction linéaire dans l'espace. Nous proposons donc de suréchantillonner la géométrie du polygone P à partir de la grille d'altération linéaire (et via une triangulation de Delaunay), en le transformant en un ensemble de polygones P_i, comme illustré dans la figure 3.13.

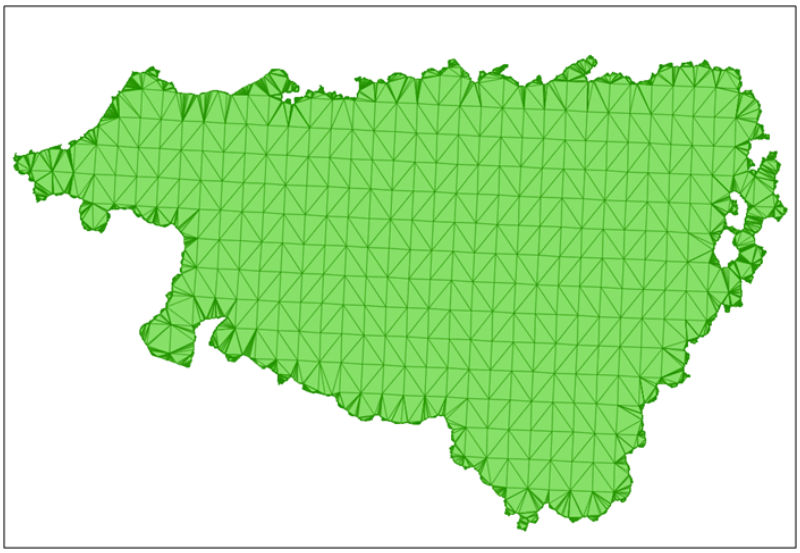


Figure 3.13 – Suréchantillonnage de la limite départementale des Pyrénées-Atlantiques à partir de la grille d'altération

Les mesures géométriques de surface sur ellipsoïde sont par conséquent réalisées sur chaque polygone P_i (par connaissance de leur altération surfacique moyenne e_{Si}), plutôt que directement sur le polygone P. La surface sur ellipsoïde $S_E(P)$ du polygone P est ainsi déterminée par la somme des surfaces sur ellipsoïde $S_E(P_i)$ des polygones P_i, comme exprimé dans l'équation 3.13.

$$S_E(P) = \sum_{i=1}^{n} S_E(P_i) \tag{3.13}$$

Pour les **objets linéaires**, comme expliqué dans l'équation 3.4, il est nécessaire de connaître l'altération linéaire en chaque point de la géométrie d'une polyligne, mais également au milieu de chaque segment inscrit entre deux points successifs. La géométrie des objets linéaires est également suréchantillonnée à l'aide de ces points.

Affectation de l'altération Etant donné que les grilles d'altération ne sont pas continues (puisque ce sont des semis de points), une méthode d'interpolation spatiale par pondération inverse à la distance, ou IDW (*Inverse Distance Weighting*), est utilisée afin de calculer l'altération en tous points dont on cherche à l'affecter.

Ainsi, pour chaque point dont on cherche à estimer l'altération linéaire (ou surfacique), les quatre points du semi les plus proches sont sélectionnés, et la valeur d'altération est déduite par interpolation IDW (cf. figure 3.14).

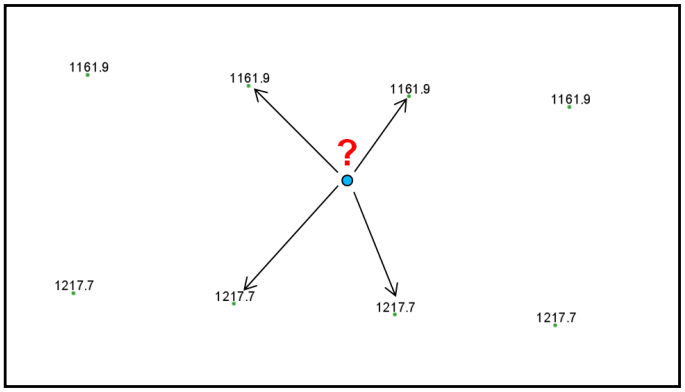


Figure 3.14 – Recherche des quatre points les plus proches dans la grille d'altération

En utilisant cette méthode, chaque point composant la géométrie des objets vectoriels peut ainsi être enrichi d'une valeur d'altération linéaire (ou surfacique, mais le principe est le même). Une fois la géométrie suréchantillonnée enrichie des informations d'altération (linéaire ou surfacique), il est à présent possible de procéder aux mesures géométriques de longueur et de surface sur ellipsoïde. La section suivante permettra de mettre en application ces méthodes d'estimation de l'impact de la projection cartographique sur les mesures géométriques de longueur et de surface.

Synthèse Section L'estimation de l'impact de la projection cartographique suppose la réalisation de mesures géométriques de longueur et de surface sur ellipsoïde. La réalisation de ces mesures sur ellipsoïde suppose de disposer d'une information d'altération (linéaire ou surfacique) sur les objets évalués. Pour cela, nous avons proposé dans cette section une méthode d'enrichissement de la géométrie des objets évalués par une information d'altération extraite de grilles d'altération préalablement créées.

3.1.3 Application du modèle d'estimation

Cette section va nous permettre de mettre en application le modèle d'estimation de l'impact de la projection cartographique, à partir des méthodes de calcul de mesures géométriques sur ellipsoïde présentées précédemment. L'expérimentation est réalisée dans le département des Pyrénées-Atlantiques, qui présente, en Lambert93, une altération linéaire positive sur l'ensemble de son territoire. Cela signifie que la projection cartographique a pour impact de sur-estimer systématiquement les mesures de longueur et de surface dans ce département. L'objectif de cette expérimentation est de quantifier cette sur-estimation.

Méthodologie Pour estimer l'impact de la projection cartographique, des objets extraits de la BDTOPO du département des Pyrénées-Atlantiques ont été exploités (cf. figure 3.15). Il s'agit ici :
– de la limite départementale, d'une surface projetée totale de 7692,21 km^2.
– du réseau routier primaire, d'une longueur projetée totale de 988,01 km.
Pour estimer l'impact de la projection cartographique, la grille d'altération linéaire de la projection Lambert93, élaborée dans la section précédente, est exploitée.

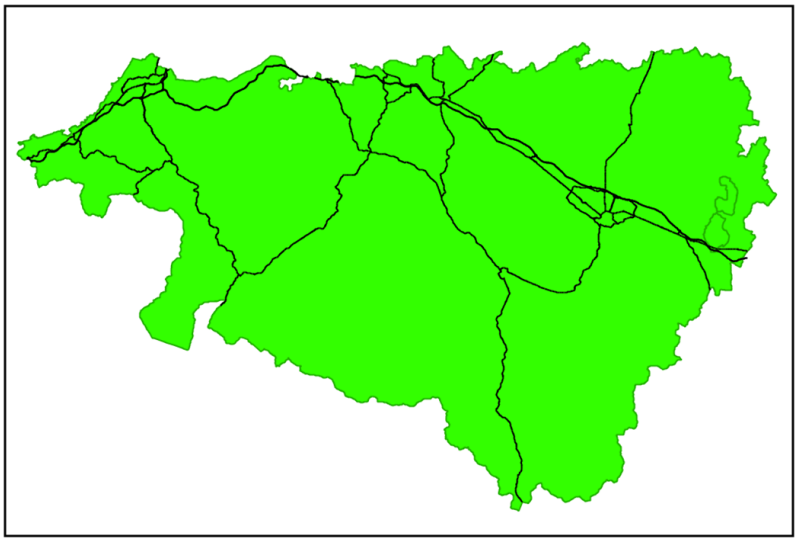


Figure 3.15 – Réseau routier primaire et limite administrative du département des Pyrénées-Atlantiques

Résultats La mise en œuvre du modèle d'estimation nous permet d'obtenir les résultats suivants :
– une surface sur ellipsoïde de 7682,34 km^2 pour le département des Pyrénées Atlantiques, soit une erreur de 9,86 km^2, équivalente à 0,12 % de la surface projetée,
– une longueur sur ellipsoïde du réseau routier de 987,48 km, soit une erreur de 533,8 mètres, équivalente à 0,05 % de la longueur projetée.
Le modèle proposé permet donc d'estimer aisément la sur-estimation impliquée par la projection Lambert93 dans cette partie du territoire métropolitain. Ce modèle sera donc intégré au sein du modèle général d'estimation de l'imprécision des mesures géométriques de longueur et de surface.

Conclusion de la section

Nous avons vu dans cette section que le processus de représentation de toute ou partie de la surface terrestre dans un plan suppose la réalisation :
– d'approximations de la surface terrestre sous forme d'un ellipsoïde,
– de déformations, induites par la projection cartographique de l'ellipsoïde dans un plan.
Ces processus de représentation impactent directement la géométrie des objets géographiques vectoriels, et par conséquent les mesures géométriques de longueur ou de surface calculées.

Cette section nous a permis de formaliser l'impact de la projection cartographique sur les mesures géométriques, et également de mettre en œuvre des méthodes permettant d'estimer automatiquement cet impact. L'intégration de ces méthodes d'estimation constituent la première pierre à l'élaboration d'un modèle général d'estimation de l'imprécision des mesures géométriques.

Nous n'avons ici évalué que l'impact de la projection cartographique, permettant de différencier les mesures géométriques calculées sur un plan de celles calculées sur un ellipsoïde. Cette estimation ne porte véritablement d'intérêt que pour les objets de grandes dimensions, dans le cadre de représentations à petite voire moyenne échelle. En effet, pour les objets de petites dimensions, l'impact de la projection cartographique sur les mesures géométriques est généralement peu significatif. Les impacts d'autres règles de représentation viennent de plus s'accumuler dans l'erreur globale de mesure, avec en premier lieu l'impact de la non-prise en compte du terrain, qui sera abordé dans la section suivante.

3.2 Prise en compte du terrain

La non-prise en compte du terrain constitue l'une des principales sources d'erreurs sur les mesures géométriques. En effet, pour diverses raisons, les mesures de longueur ou de surface sont souvent réalisées en deux dimensions, ce qui a pour impact de sous-estimer systématiquement les mesures calculées. Nous rappellerons tout d'abord les diverses raisons expliquant pourquoi le terrain n'est pas toujours intégré dans les mesures géométriques, et les solutions qui peuvent être envisagées afin de le prendre en compte. Nous présenterons ensuite les méthodes mises en œuvre afin d'enrichir la géométrie des objets vectoriels et ainsi procéder au calcul de longueurs et de surfaces en 2D5, c'est-à-dire en prenant en compte le terrain. Nous terminerons par des expérimentations permettant de valider la méthode proposée.

3.2.1 Modélisation de l'information altimétrique et mesures géométriques

Avant de détailler les solutions à mettre en œuvre afin de prendre en compte l'impact du terrain sur les mesures géométriques de longueur et de surface, il convient tout d'abord de rappeler les principales raisons limitant ces mesures à des calculs en deux dimensions.

3.2.1.1 Des mesures géométriques souvent limitées en 2D

Plusieurs raisons expliquent pourquoi aujourd'hui les mesures géométriques sont généralement réalisées en deux dimensions dans la majorité des logiciels SIG, ayant comme conséquence de les sous-estimer systématiquement. Ces raisons sont principalement liées à deux facteurs :

- la **modélisation géométrique**, qui n'intègre pas, ou mal, la dimension altimétrique,
- les **algorithmes de mesures géométriques** proposés dans les logiciels, qui sont souvent limités aux calculs en deux dimensions.

Problèmes de modélisation géométrique

Différents processus de production peuvent être mobilisés afin de procéder à la construction des objets géographiques d'une base de données. En fonction des besoins spécifiés, et des moyens financiers mobilisés afin de saisir la base de données, la modélisation géométrique des objets vectoriels intègrera, ou non l'altitude. Par exemple, dans la BDTOPO, les tronçons de route (et bien d'autres classes d'objets) sont saisis sous forme d'objets linéaires tridimensionnels, c'est-à-dire que chaque sommet constituant la géométrie d'un objet est renseigné dans ses coordonnées x, y et z (cf. figure 3.16). A contrario, dans la BDCARTO, les tronçons de route ne sont saisis qu'en deux dimensions, le z n'étant pas renseigné.

```
0  LINESTRING (
1      403082.7 6266910.9 89.6,
2      403096.5 6266909.9 89.9,
3      403103 6266909.5 89.9,
4      403111.2 6266908.9 89.9,
       403129.4 6266907.6 90
   )
```

Figure 3.16 – Renseignement de l'altitude des tronçons de route de la BDTOPO en coordonnées tridimentionnelles

Le renseignement de l'altitude dans une base de données géographiques demeure en effet un processus relativement coûteux, puisqu'il n'est assuré que par certaines méthodes de restitution exigeant des moyens importants, comme les relevés topographiques ou GNSS, les relevés Lidar, ou encore les relevés photogrammétriques. Lorsque les spécifications ne l'exigent pas, les altitudes ne sont donc pas saisies, ce qui génère une sous-estimation des mesures géométriques réalisées en aval.

Cependant, même si l'altitude d'une géométrie est renseignée, certains formats de fichiers vectoriels ne gérant pas les géométries tridimensionnelles, il peut arriver que les informations d'altitude soient stockées dans la table attributaire, empêchant la réalisation de mesures géométriques en trois dimensions. C'était par exemple le cas de la classe d'objets "tronçon de route" de la BDTOPO version 1 au format *shapefile*, pour laquelle l'altitude était renseignée dans deux champs de la table attributaire "Z-INI" et "Z-FIN" correspondant respectivement à l'altitude des deux extrémités d'un tronçon de route (cf. figure 3.17). Cependant, dans les nouvelles versions du format *shapefile*, les tronçons de route sont représentés sous forme de linéaires tridimensionnels, où chaque sommet d'une géométrie est renseigné en x, y et z.

Z_INI	Z_FIN
82.8	83.0
258.9	259.0
166.6	166.1
299.7	299.6
230.5	231.4
158.6	158.6
1139.3	1139.2
1311.8	1312.3
169.8	169.9
813.7	813.8

Figure 3.17 – Renseignement de l'altitude des tronçons de route de la BDTOPO version 1, au format *shapefile*

Bien que ce problème soit aujourd'hui résolu pour les objets linéaires, la modélisation de l'altitude des objets surfaciques est quant à elle beaucoup plus délicate à renseigner afin de permettre des mesures géométriques de surface en 2D5.

En effet, le renseignement tridimensionnel de chaque sommet d'une géométrie surfacique ne suffit pas à assurer des mesures de surface en 2D5 puisque cette modélisation ne prend pas en compte les altitudes inscrites à l'intérieur du polygone. On peut citer à ce titre l'exemple de la classe "Cimetières" de la BDTOPO, modélisée géométriquement sous forme d'objets surfaciques tridimensionnels. Le contour de chaque objet de cette classe correspond à l'enceinte du cimetière, définie soit par la hauteur du mur, du toit ou du sol suivant les cas de figure (IGN, 2011c).

Bien que l'altitude soit renseignée en chaque sommet du contour d'un cimetière, il n'est pas envisageable de réaliser une mesure géométrique de surface en 2D5, puisque les altitudes situées à l'intérieur du polygone ne sont pas prises en compte (cf. figure 3.18). D'une certaine manière, l'intérieur du polygone n'est pas représenté, ce qui n'est pas rédhibitoire pour les surfaces planes, mais ce n'est pas toujours forcément le cas.

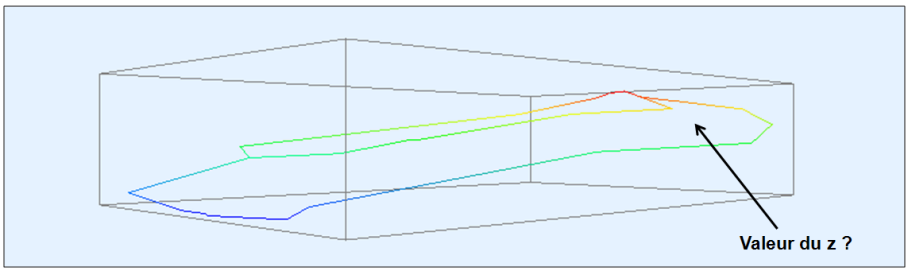


Figure 3.18 – Vue tridimensionnelle de deux polygones de la classe "Cimetières" de la BDTOPO (sur la commune d'Hendaye)

Nous voyons bien à travers ces exemples qu'une modélisation géométrique adaptée de l'information altimétrique des objets vectoriels doit être entreprise afin de réaliser des mesures géométriques de longueur ou de surface en 2D5.

Limites des algorithmes de mesures géométriques

En règle générale, la majorité des logiciels SIG traditionnels ne proposent que des fonctionnalités permettant la réalisation de mesures géométriques en deux dimensions, sans prendre en compte la dimension altimétrique. Les mesures de longueur sont généralement réalisées par application de l'abscisse curviligne en 2D (cf. équation 1.1) et les mesures de surface sont réalisées par le calcul de l'aire du polygone (cf. équation 1.2).

Depuis quelques années, avec l'avènement des objets géographiques en trois dimensions, des fonctionnalités de calculs de longueur et de surface en 2D5 sont peu à peu implémentées dans les logiciels SIG. Cependant, si le calcul d'un abscisse curviligne en 2D5 ne pose pas de problèmes, le calcul d'une surface en 2D5 est beaucoup plus coûteux en termes de préparation de données (puisqu'il est nécessaire de disposer d'un MNT) et de temps de calcul.

3.2.1.2 Proposition pour des mesures géométriques en 2D5

La prise en compte du terrain dans les mesures géométriques de longueur et de surface réalisées à partir de la géométrie d'objets vectoriels suppose d'assurer les deux conditions suivantes :
- la modélisation correcte de l'altitude sur les géométries des objets linéaires et surfaciques,
- l'implémentation d'algorithmes de mesures géométriques de longueur et de surface en 2D5.

Etant donné que de nombreuses bases de données vectorielles ne disposent pas d'altitude renseignée (dans le cas où le z n'est pas saisi) ou correctement renseignée (dans le cas où seuls les points du contour des objets surfaciques disposent d'un z saisi), il est donc nécessaire de mettre en place un système permettant d'enrichir correctement d'une information altimétrique les objets qui en sont dépourvus.

Pour enrichir l'altitude de ces géométries vectorielles, l'intégration de données altimétriques externes, provenant de modèles numériques de terrain, peut être mise en œuvre. Dans le cadre du développement du modèle, nous privilégions d'exploiter prioritairement des modèles numériques de terrain (ou d'élévation) disponibles gratuitement (par téléchargement sur Internet), afin de rendre cette application accessible au plus grand nombre. Bien évidemment, nous encourageons l'utilisation de modèles numériques de terrain (plus précis) lorsque ceux-ci existent ou sont disponibles.

La modélisation de l'impact du terrain sur les mesures géométriques de longueur et de surface que nous proposons dans cette section se décline en plusieurs étapes (cf. figure 3.19) :

1. Pré-traitement du Modèle Numérique de Terrain,
2. Enrichissement de la géométrie des objets vectoriels d'informations altimétriques,
3. Calcul des mesures géométriques (de longueur ou de surface) en 2D5.

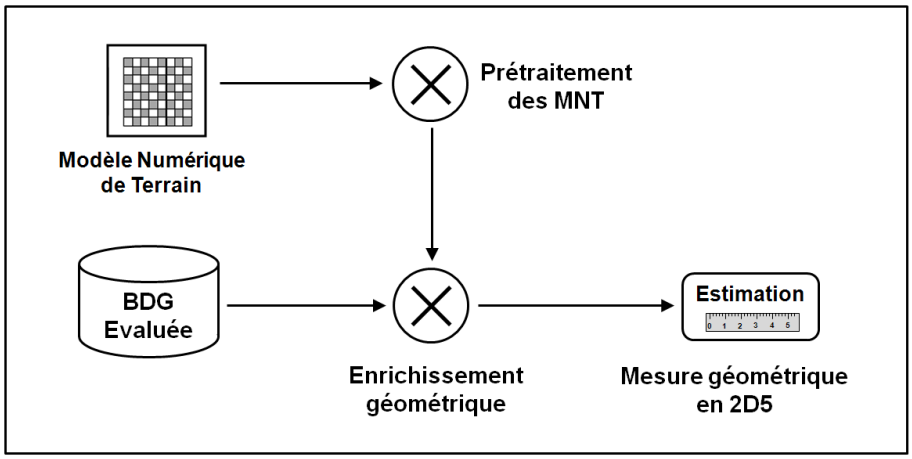


Figure 3.19 – Enrichissement de la géométrie des objets évalués pour le calcul de mesures géométriques en 2D5

La section suivante aura pour objectif de détailler les différentes étapes permettant d'estimer l'impact de la non-prise en compte du terrain sur les mesures géométriques.

Synthèse Section Nous avons vu dans cette section que pour différentes raisons, liées à la modélisation de l'information altimétrique ou à la disponibilité d'algorithmes appropriés, les mesures géométriques de longueur, et surtout de surface, sont généralement limitées en 2D. Ceci a pour conséquence de les sous-estimer systématiquement. Nous proposons donc des méthodes d'enrichissement de la géométrie des objets évalués, ainsi que des algorithmes de mesures de longueur et de surface en 2D5, afin d'évaluer l'impact de la non-prise en compte du terrain sur les mesures géométriques.

3.2.2 Méthodes d'enrichissement géométrique pour des mesures en 2D5

Cette section a pour objectif de présenter les méthodes permettant d'enrichir les géométries linéaires et surfaciques d'une information altimétrique, afin de procéder à des mesures géométriques de longueur et de surface en 2D5. Pour chacun des deux types de primitives géométriques concernées (linéaire ou surfacique), les méthodes d'enrichissement ainsi que les algorithmes de calcul des mesures géométriques seront présentés. L'enrichissement de la géométrie est basé sur l'exploitation de modèles numériques de terrain, faisant préalablement l'objet de pré-traitements en vue de leur intégration.

Pré-traitement des MNT Les Modèles Numériques de Terrain que nous allons utiliser sont généralement structurés sous forme d'un maillage carré où chaque maille (ou pixel) porte comme valeur l'information d'altitude. Afin d'être exploités en vue d'enrichir la géométrie des objets vectoriels, ces modèles numériques de terrain doivent subir deux étapes de pré-traitement afin d'être transformés en maillage triangulaire régulier (cf. figure 3.20). Ces deux étapes sont les suivantes :

1. le modèle numérique de terrain est converti de son format d'origine (par exemple le format *GeoTiff*) au format *ASCII Grid*, permettant de remplacer chaque maille par un point localisé en son centre, conservant l'information d'altitude.
2. le modèle numérique de terrain converti est ensuite transformé en maillage triangulaire régulier, après réalisation d'une triangulation de Delaunay, où chaque sommet du triangle conserve une information d'altitude.

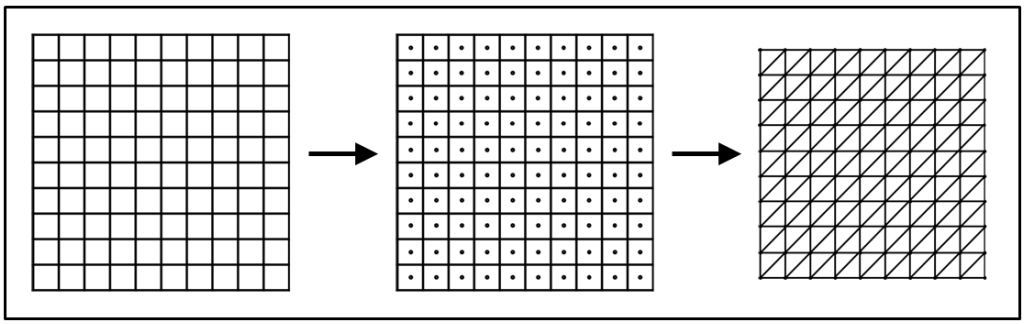


Figure 3.20 – Etapes de transformation d'un modèle numérique de terrain d'un maillage carré vers un réseau de triangles réguliers

Cette transformation de la modélisation du modèle numérique de terrain en un réseau de triangles réguliers se justifie par sa plus grande facilité d'exploitation avec des données vectorielles. Le réseau de triangles ainsi créé constitue lui aussi une base de données vectorielles (constituée de sommets, d'arrêtes et de faces), qu'il est possible d'intersecter avec d'autres objets vectoriels. Ceci nous sera nécessaire afin d'enrichir la géométrie des objets vectoriels d'une information altimétrique.

Nous allons maintenant présenter dans les deux sections suivantes les méthodes d'enrichissement des géométries linéaires et surfaciques d'une information altimétrique, ainsi que les algorithmes de mesures de longueur et de surface en 2D5.

3.2.2.1 Primitives linéaires

Pour les géométries linéaires, nous présenterons dans cette section la méthode d'enrichissement de la géométrie d'une information altimétrique, puis l'algorithme permettant de réaliser des mesures géométriques de longueur en 2D5.

Enrichissement de la géométrie

Pour les objets ne disposant que de coordonnées bidimensionnelles (x,y) en chaque sommet de leur géométrie, l'objectif de cette méthode consiste à extraire les informations altimétriques contenues dans le modèle numérique de terrain, et ainsi affecter un z aux coordonnées de chacun des sommets. La méthode d'enrichissement géométrique que nous proposons ici est basée sur une intersection entre le modèle numérique de terrain (représenté sous forme de triangles réguliers) et la géométrie de l'objet linéaire à enrichir. Deux étapes successives sont nécessaires à l'enrichissement de la géométrie : le suréchantillonnage de la géométrie de l'objet, puis l'affectation des altitudes.

Suréchantillonnage de la géométrie Afin de mettre en cohérence l'objet vectoriel linéaire dont on souhaite enrichir la géométrie d'informations altimétriques extraites d'un modèle numérique de terrain, un suréchantillonnage de la géométrie de l'objet est appliqué. Le suréchantillonnage est effectué en ajoutant à la géométrie un nouveau sommet à chaque intersection avec une arrête du réseau de triangles réguliers (cf. figure 3.21).

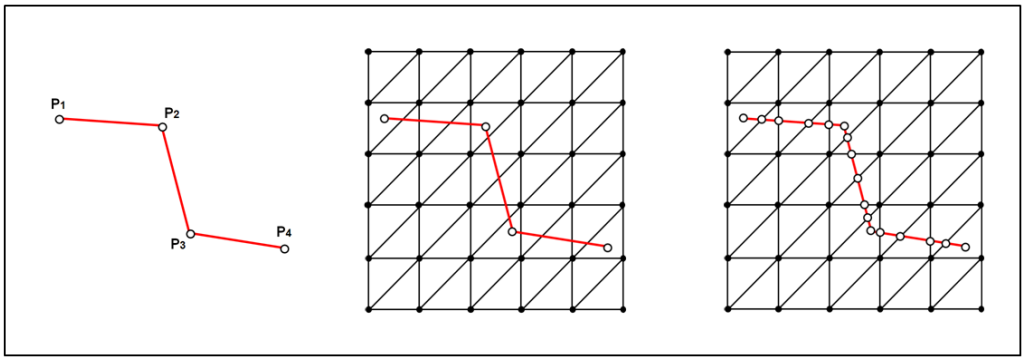


Figure 3.21 – Principe de suréchantillonnage d'une géométrie linéaire par intersection avec un réseau de triangles réguliers

Affectation des altitudes Chacun des points composant la géométrie suréchantillonnée de l'objet linéaire intersecte donc une face du réseau de triangles réguliers, dont on connait l'altitude en chaque sommet. L'affectation de l'altitude z à chacun des points de la géométrie suréchantillonnée est réalisée par application d'une **équation cartésienne de plan** entre un vecteur normal (calculé à partir du triangle intersecté du MNT) et le point, sous la forme :

$$ax + by + cz + d = 0 \tag{3.14}$$

Une fois l'objet linéaire suréchantillonné et enrichi d'informations altimétriques, il est possible d'opérer à des mesures géométriques de longueur en 2D5.

Mesures de longueur en 2D5

Pour prendre en compte le terrain dans les mesures géométriques de longueur, le calcul de l'abscisse curviligne est utilisé en intégrant l'altitude z (cf. équation 3.15). Ainsi, dans un espace à trois dimensions, la longueur L_Z d'une polyligne est calculée comme suit :

$$L_Z = \sum_{i=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2} \tag{3.15}$$

La section suivante permettra de présenter la méthode d'enrichissement géométrique et de calcul de mesures en 2D5 appliquée au géométriques surfaciques.

3.2.2.2 Primitives surfaciques

Pour les objets surfaciques, les questions d'enrichissement de la géométrie et de mesures de surface en 2D5 sont plus complexes que pour les objets linéaires. Nous les décrivons ci-dessous.

Enrichissement de la géométrie

Pour pouvoir réaliser des mesures géométriques de surface en 2D5, les objets surfaciques doivent tout d'abord disposer de coordonnées tridimensionnelles (x,y,z) en chacun des sommets du contour de leur géométrie. Pour enrichir ce contour d'une information altimétrique, la méthode utilisée est identique à celle présentée précédemment pour l'enrichissement de la géométrie des objets linéaires.

Cependant pour réaliser des mesures géométriques de surface en 2D5, les objets surfaciques doivent également disposer de coordonnées tridimensionnelles à l'intérieur du polygone, afin de prendre en compte les variations du terrain inscrites à l'intérieur de la géométrie. Nous allons donc présenter dans cette section une méthode permettant d'enrichir l'intérieur de la géométrie d'un polygone d'une information altimétrique.

Suréchantillonnage de la géométrie Afin de mettre en cohérence la géométrie de l'objet surfacique avec le modèle numérique de terrain, il est tout d'abord nécessaire de procéder à un suréchantillonnage de la géométrie. L'opération de suréchantillonnage du contour est identique à celle des objets linéaires, c'est-à-dire qu'à chaque intersection entre le contour du polygone avec une arrête du réseau de triangles réguliers, un nouveau sommet est ajouté à la géométrie.

Etant donné que les objets surfaciques doivent également tenir compte des informations altimétriques inscrites à l'intérieur du polygone, la géométrie initiale est également enrichie :
– des triangles entièrement inclus dans le polygone,
– des portions de triangles incluses à l'intérieur du contour du polygone.
Suite à cette opération, le polygone initial correspond désormais à un ensemble de polygones, triangulaires ou non, comme illustré dans la figure 3.22.

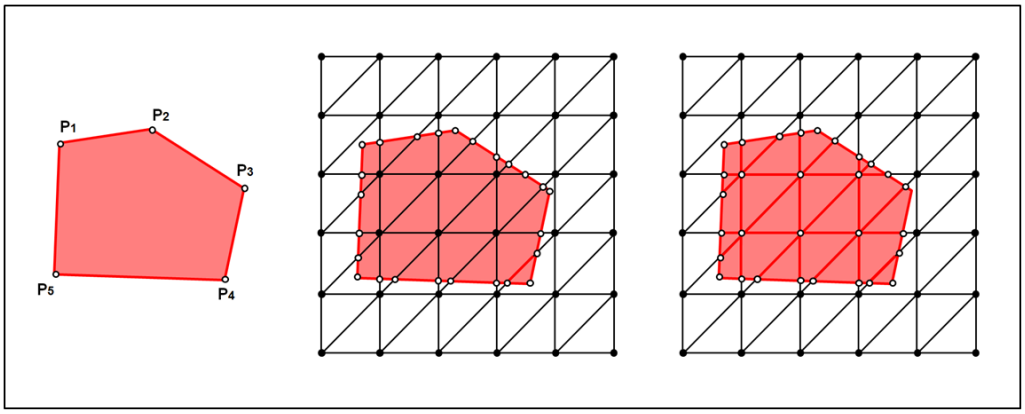


Figure 3.22 – Principe de suréchantillonnage d'une géométrie surfacique par intersection avec un réseau de triangles réguliers

Affectation des altitudes Le principe d'affectation des altitudes est identique à celui réalisé pour les objets linéaires. L'affectation de l'altitude z en chaque point de la géométrie surfacique suréchantillonnée est réalisée par application d'une équation cartésienne de plan.

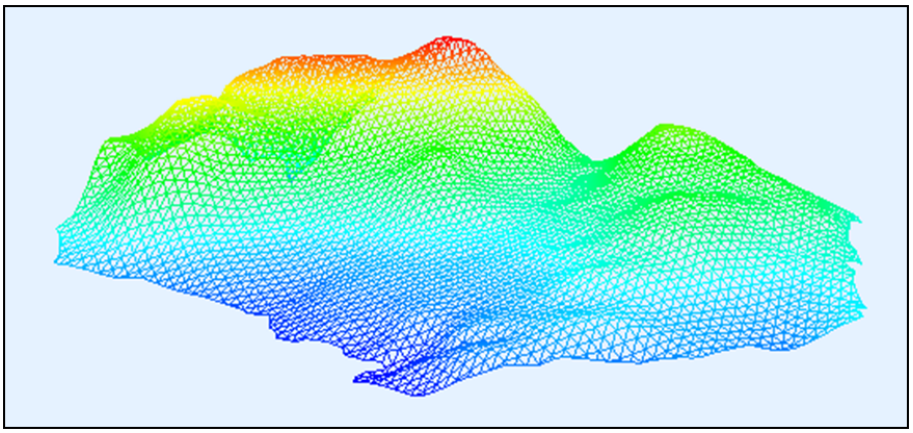


Figure 3.23 – Représentation exagérée en trois dimensions de la géométrie enrichie d'une limite communale

Le polygone initial surechantillonné est à présent enrichi d'informations altimétriques à la fois sur le contour et à l'intérieur de sa géométrie, comme l'illustre la figure 3.23 avec une limite communale.

Mesure géométrique de surface en 2D5

Pour réaliser des mesures géométriques de surface en 2D5, la stratégie mise en œuvre est différente de celle utilisée pour les mesures de longueur, où le calcul d'un abscisse curviligne est réalisé. Le calcul de mesures de surface en 2D5 est réalisable à partir de l'aire des triangles.

Cependant, dans le cas des géométries surfaciques suréchantillonnées, nous disposons effectivement de triangles entiers, correspondant aux triangles du MNT entièrement inscrits dans le contour de l'objet surfacique, mais également de polygones, correspondant aux portions de triangles incluses (et donc découpées) dans le contour de l'objet surfacique (cf. figure 3.24). Suivant ces deux cas de figure, deux calculs de surface sont mis en œuvre.

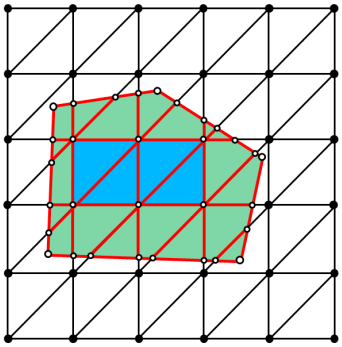


Figure 3.24 – Triangles entièrement inclus (en bleu) et portions de triangles incluses (en vert) dans le contour d'un objet surfacique

Aire des triangles entiers Lorsqu'un triangle est entièrement inclus dans le contour de l'objet surfacique, sa surface est calculée directement. Chaque triangle (aux coordonnées tridimensionnelles) est ici défini par deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$. L'aire $A_Z(T)$ du triangle T est donc égale à la norme euclidienne du produit vectoriel des deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$, divisée par deux (cf. équation 3.16).

$$A_Z(T) = \frac{|| \overrightarrow{u} \wedge \overrightarrow{v} ||}{2} \tag{3.16}$$

Aire des portions de triangles Le calcul de l'aire des polygones correspondant aux portions de triangles incluses au sein du contour de l'objet surfacique ne peut pas être réalisé directement en 2D5. Pour la calculer, une méthode basée sur le taux de recouvrement de la portion de triangle par rapport au triangle entier est utilisée.

Pour illustrer cette méthode, considérons un polygone P correspondant à une portion du triangle T. En deux dimensions, le calcul des aires $A(P)$ du polygone P et $A(T)$ du triangle T ne pose pas de problèmes. Il est donc possible de calculer le taux de recouvrement r de P par rapport à T en deux dimensions, conformément à l'équation 3.17.

$$r = \frac{A(P)}{A(T)} \tag{3.17}$$

En trois dimensions, l'aire $A_Z(P)$ du polygone P est calculée par application du taux de recouvrement r sur l'aire $A_Z(T)$ du triangle T, comme exposé dans l'équation 3.18.

$$A_Z(P) = r * A_Z(T) \tag{3.18}$$

Surface totale d'un polygone en 2D5 A partir de ces deux méthodes, il est possible de procéder au calcul d'une mesure de surface S_Z en 2D5 sur une géométrie surfacique enrichie. Ce calcul correspond à la somme des aires $A_Z(T)$ des triangles entièrement inscrits dans le contour de l'objet surfacique avec les aires $A_Z(P)$ des polygones correspondant aux portions de triangles incluses dans le contour de l'objet surfacique (cf. équation 3.19).

$$S_Z = \sum A_Z(P) + \sum A_Z(T) \tag{3.19}$$

Nous venons de présenter des méthodes permettant d'enrichir la géométrie d'objets dépourvus d'information altimétrique, et ainsi réaliser des mesures géométriques en 2D5 afin d'estimer l'impact de la non-prise en compte du terrain. Ces méthodes présentent cependant certaines limites, que nous allons présenter dans la section suivante.

3.2.2.3 Limites des méthodes proposées

Les méthodes proposées offrent une possibilité simple et efficace pour réaliser des mesures géométriques de longueur et de surface en 2D5, et ainsi estimer l'impact de la non-prise en compte du terrain. Cependant, ces méthodes présentent un certain nombre de limites, principalement liées à la modélisation de l'altitude, à sa représentation sous forme de triangles réguliers, ainsi qu'à la qualité intrinsèque des modèles numériques de terrain.

Modélisation du terrain La première limite relative aux méthodes proposées dans cette section fait référence à la modélisation du terrain sous forme de triangles réguliers. En effet, dans la réalité, le relief n'est pas modélisé sous forme de mailles carrées ou de triangles réguliers. Le monde réel est beaucoup plus complexe et ce type de représentation ne constitue qu'une simplification du terrain tel qu'il existe réellement. De plus, ces données altimétriques portent une imprécision, puisque elles sont généralement calculées par le biais de méthodes d'interpolation, permettant de déterminer des valeurs d'altitudes là où il n'en existe pas dans les données sources, comme des courbes de niveau par exemple. Ces questions relatives à l'imprécision des modèles numériques de terrain ont fait l'objet de nombreuses contributions (voir par exemple Shortridge (2001), Fisher et Tate (2006) ou Oksanen (2006)).

MNT et MNE La seconde limite à prendre en compte fait référence à l'altitude réellement représentée dans le modèle numérique. Il est en effet nécessaire de différencier les modèles numériques de terrain (MNT) des modèles numériques d'élévation (MNE), comme illustré dans la figure 3.25. Les modèles numériques de terrain représentent l'altitude au niveau du terrain (à proprement parler), tandis que les modèles numériques d'élévation représentent l'altitude au niveau du terrain (si celui-ci n'est pas recouvert), mais également l'altitude des éléments du sur-sol (comme la canopée d'une forêt ou le toit des bâtiments).

Des différences d'altitudes importantes peuvent donc être observées entre MNT et MNE. L'utilisation de MNE aura donc pour conséquence de dégrader les mesures géométriques de longueur ou de surface des objets enrichis d'une information altimétrique.

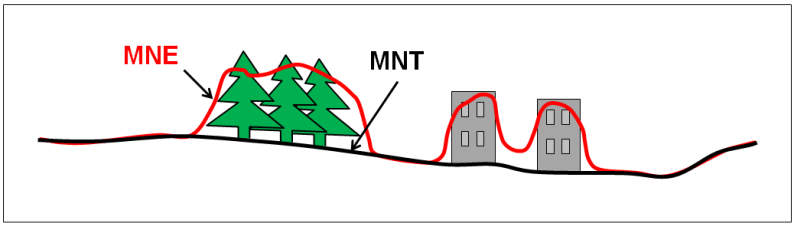


Figure 3.25 – Différence entre un MNT (en noir) et un MNE (en rouge)

Il est donc préférable d'exploiter uniquement des modèles numériques de terrain dans le cadre de ce travail de thèse. Cependant, ces modèles numériques de terrain ne sont pas toujours disponibles facilement (voire gratuitement) sur l'ensemble de la surface terrestre. En effet, le processus de production d'un modèle numérique de terrain est plus long et complexe que celui d'un modèle numérique d'élévation.

Malgré ces limites relatives à la modélisation de l'altitude dans les modèles numériques d'élévation, ils offrent une source d'information altimétrique intéressante pour notre travail, puisque différentes sources sont disponibles gratuitement. On peut citer notamment le SRTM (*Shuttle Radar Topography Mission*) produit par la NASA (Farr *et al.*, 2007), ou encore l'ASTER GDEM (*Advanced Spaceborne Thermal Emission and Reflection Radiometer*), fruit d'une collaboration METI/NASA (voir par exemple Hirt *et al.* (2010)). Une expérimentation portant sur la comparaison des impacts entre MNT et MNE sur les mesures géométriques en 2D5 sera entreprise dans la section suivante.

Qualité du modèle La dernière limite observée traite de la qualité intrinsèque du modèle altimétrique utilisé. En effet, les caractéristiques des modèles numériques exploités pour enrichir les géométries des objets vectoriels auront un impact direct sur les mesures géométriques réalisées par la suite. On peut noter à ce titre l'impact de la **résolution spatiale**, qui fait référence à la taille du pixel dans la grille d'altitude. Plus la résolution spatiale du modèle numérique est faible, plus la taille du pixel est grande. Par conséquent la modélisation du terrain est plus imprécise, ce qui se répercute sur la géométrie des objets enrichis. Une étude de l'impact de la résolution spatiale sur les mesures de longueur et de surface sera également proposée dans la section suivante.

Nous voyons bien à travers ces quelques arguments, que les stratégies de modélisation du terrain impactent directement l'enrichissement des objets vectoriels d'une information altimétrique, et par conséquent les mesures géométriques de longueur ou de surface en 2D5 calculées en aval. Il conviendra donc de tenir compte de ces différents impacts dans la section suivante, qui permettra de mettre en application les méthodes d'enrichissement et de mesures géométriques proposées.

Synthèse Section Des méthodes d'enrichissement de la géométrie d'un information altimétrique et des algorithmes de calcul de mesures géométriques de longueur et de surface sont proposés afin d'estimer l'impact de la non-prise en compte du terrain. L'évaluation par le biais de ces méthodes présente cependant quelques limites, impliquées notamment par les moyens d'acquisition et de modélisation du terrain, en particulier si la source altimétrique est un MNT ou MNE.

3.2.3 Application du modèle d'estimation

Les expérimentations qui sont menées dans cette section ont pour objectif d'étudier la sensibilité de l'erreur entre des mesures géométriques réalisées en 2D et 2D5, suite à un enrichissement de la géométrie des objets vectoriels. Ces expérimentations chercherons également à prendre en compte :
– l'impact de la résolution du modèle numérique de terrain sur l'erreur de mesure,
– l'impact d'un modèle numérique d'élévation sur l'erreur de mesure.

La librairie GeOxygene (Grosso *et al.*, 2012) a été utilisée pour implémenter les méthodes d'estimation de l'impact de la non-prise en compte du terrain proposées précédemment, en exploitant notamment les fonctionnalités apportées par la bibliothèque GeOxygene3D (Brasebin, 2009).

Bases de données utilisées Pour mener ces expérimentations, des données localisées dans le département de l'Isère sont utilisées. Ceci se justifie par le fait que ces données sont situées dans une zone de fortes pentes, ce qui nous permet d'étudier facilement l'impact de la non-prise en compte du terrain sur des mesures géométriques.
Les données utilisées sont extraites des bases de données suivantes :
– la BDCARTO (IGN, 2011b), base de données cartographiques que nous cherchons à enrichir d'informations altimétriques,
– la BDTOPO (IGN, 2011c), base de données topographiques de référence, où l'altitude est renseignée dans certaines classes d'objets,
– la BDALTI (IGN, 2011a), modèle numérique de terrain, produit à différentes résolutions spatiales,
– le SRTM (Farr *et al.*, 2007), modèle numérique d'élévation, produit à une résolution spatiale de trois secondes d'arc.
La figure 3.26 présente le MNT BDALTI représenté à différentes résolutions spatiales, ainsi que le MNE SRTM.

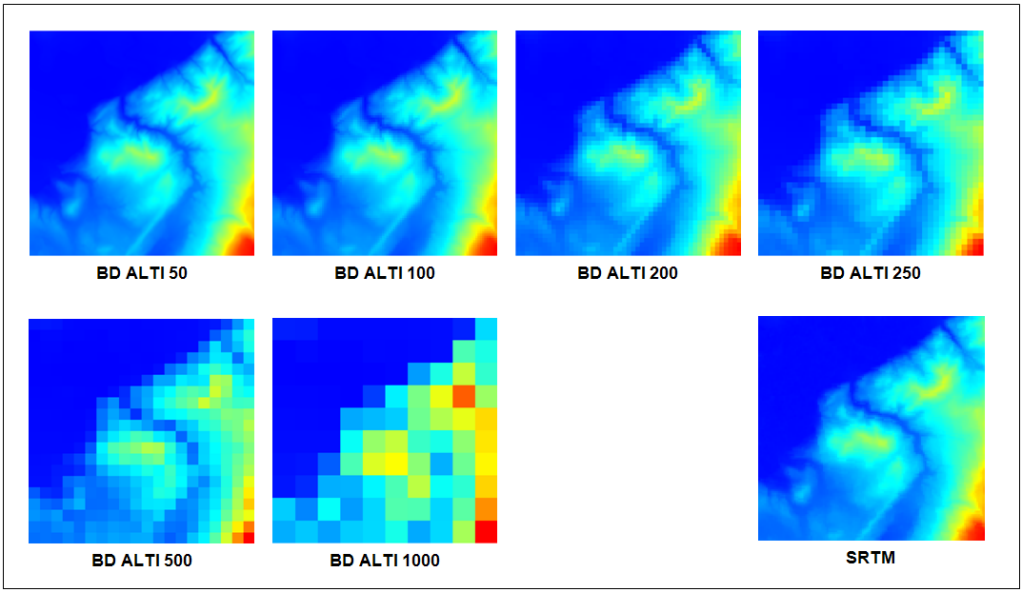


Figure 3.26 – Les MNT et MNE utilisés pour l'expérimentation

Nous chercherons dans les deux sections suivantes à étudier la sensibilité de l'erreur de mesure géométrique en comparant les différentes sources de données altimétriques, selon la résolution du MNT ou en utilisant un MNE.
Les expérimentations sont menées sur le réseau routier et la limite communale de la commune d'Herbeys, extraits de la BDCARTO (cf. figure 3.27).

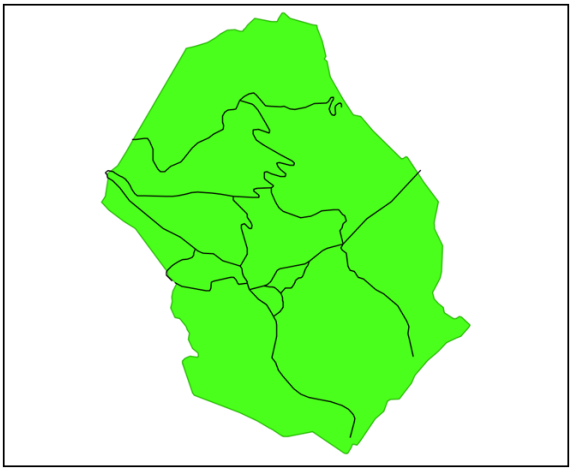


Figure 3.27 – Limite communale et réseau routier de la commune d'Herbeys

3.2.3.1 Impact de la résolution spatiale

Pour étudier l'impact de la résolution spatiale d'un modèle numérique de terrain sur les mesures géométriques de longueur et de surface, une première expérimentation est proposée en comparant les mesures calculées suite à un enrichissement réalisé à l'aide du MNT BDALTI produit à différentes résolutions spatiales (50, 100, 200, 250, 500 et 1000 mètres). Les objets à enrichir sont issus de la BDCARTO. Pour les primitives linéaires, une extraction du réseau routier est utilisée, tandis que pour les primitives surfaciques, une limite communale est exploitée.

Primitives linéaires Le graphique de la figure 3.28 présente les erreurs de longueur, suite aux comparaisons entre la longueur du réseau routier extrait de la BDCARTO (dont la longueur en deux dimensions est de 17,46 km) et les longueurs 2D5 suite à un enrichissement de la géométrie avec le MNT BDALTI à différentes résolutions spatiales. L'erreur correspond à la différence entre la longueur 2D5 et la longueur 2D.

On constate que la résolution spatiale du modèle numérique de terrain impacte les mesures de longueur calculées à partir de la géométrie enrichie. En effet, dans cet exemple, plus la résolution est faible, plus l'erreur de longueur est importante, sauf pour la BDALTI échantillonnée à 1000 mètres (qui n'est pas assez résolue pour des objets de la BDCARTO). A titre de comparaison, la longueur en 2D5 de ce même réseau routier extrait de la BDTOPO est de 17,74 km. Cela signifie que malgré l'enrichissement de géométrie, et la prise en compte du terrain dans la mesure de longueur, l'impact d'autres processus doit être pris en compte afin d'expliquer l'erreur de longueur.

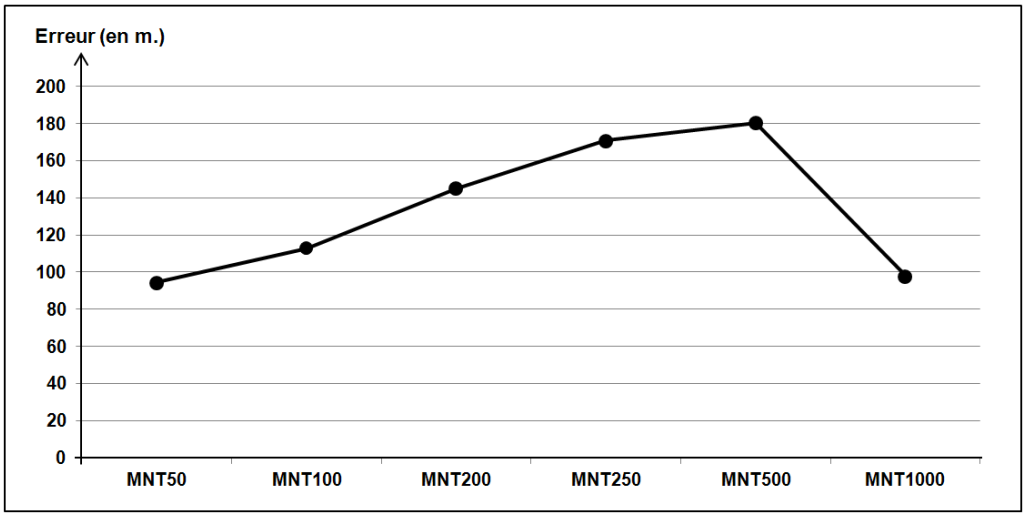


Figure 3.28 – Erreurs de longueur du réseau routier de la commune d'Herbeys entre mesures en 2D et en 2D5

Primitives surfaciques Le graphique de la figure 3.29 présente les erreurs de surface, suite aux comparaisons entre la surface 2D de la commune de Herbeys (dont la surface en 2D est de 7.92 km^2) et les surfaces 2D5 suite à un enrichissement de la géométrie avec le MNT BDALTI à différentes résolutions spatiales.
Contrairement à l'étude des longueurs avec l'exemple d'une route de montagne, la surface calculée en 2D5 a plutôt tendance à décroitre avec la baisse de la résolution spatiale du modèle numérique de terrain. On peut considérer que l'erreur de surface demeure relativement similaire pour les modèles numériques de terrain à haute et moyenne résolution spatiale (BDALTI à 50, 100, 200, voire 250 m), ce qui n'est pas le cas de ceux à faible résolution spatiale (BDALTI à 500 et 1000 m).

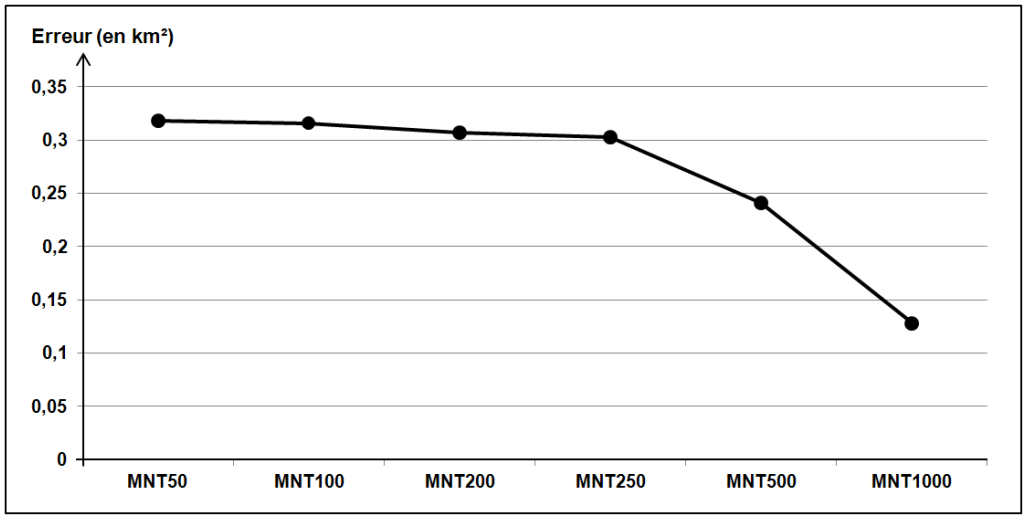


Figure 3.29 – Erreurs de surface de la limite administrative de la commune d'Herbeys entre mesures en 2D et en 2D5

3.2.3.2 Comparaison de l'impact entre MNT et MNE

Afin d'étudier l'impact des modèles numériques d'élévation sur les mesures géométriques de longueur et de surface, par comparaison avec l'étude précédente réalisée avec la BDALTI à différentes résolutions, le MNT SRTM (Farr *et al.*, 2007) a été utilisé. Ce modèle numérique d'élévation dispose d'une résolution spatiale au sol de 3 secondes d'arc, ce qui correspond dans notre zone à une taille de pixel au sol d'environ 80 mètres. Le SRTM a été utilisé afin d'enrichir la géométrie des objets linéaires et surfaciques de la BDCARTO exploités dans l'expérimentation précédente.

Suite à l'enrichissement géométrique, les mesures géométriques en 2D5 obtenues sont les suivantes :
– longueur du réseau routier : 17,59 km, soit une erreur de longueur de 129,97 m,
– superficie de la commune : 8.28 km^2, soit une erreur de surface de 0,35 km^2.

Ces résultats montrent que l'enrichissement géométrique à l'aide d'une information altimétrique extraite d'un MNE permet d'estimer des erreurs de longueur et de surface globalement identiques à celles calculées avec une information altimétrique extraite de MNT. Si pour l'étude du réseau routier, l'erreur constatée est proche de celle calculée avec le MNT à 100 mètres de résolution, on constate malgré tout une surestimation de la surface de la limite communale avec le MNE, par comparaison avec les MNT, peu importe la résolution utilisée. Malgré tout, ces résultats montrent que l'utilisation d'un MNE comme le SRTM permet d'estimer une approximation réaliste de l'erreur impliquée par la non-prise en compte du terrain.

Cette observation révèle toute son importance pour la suite de ce travail et en particulier pour le fonctionnement du modèle général d'estimation. En effet, il est souhaité que le modèle développé dans le cadre de cette thèse soit accessible au plus grand nombre. Il est donc nécessaire de privilégier l'exploitation de données externes facilement accessibles. Or, en ce qui concerne les informations altimétriques, les modèles numériques d'élévation sont beaucoup plus facilement disponibles (en termes de couverture et de gratuité) que les modèles numériques de terrain, qui sont généralement payants, car plus complexes à produire.
Ainsi, l'exploitation de modèles numériques d'élévation ne doit être réalisée que si aucun modèle numérique de terrain n'est disponible. Il serait en effet dommageable de ne pas exploiter des modèles numériques d'élévation dans le cas où seules ces données sont disponibles, puisque même si elles ont tendance à sur-estimer l'erreur de mesure impactée par la non-prise en compte du terrain, elles offrent malgré tout une approximation assez réaliste.

Synthèse Section L'expérimentation réalisée au cours de cette section a permis de quantifier l'impact de la résolution spatiale d'un MNT sur l'erreur impliquée par la non-prise en compte du terrain, à travers l'exemple d'un réseau routier et d'une limite communale localisés en zone de montagne. Les résultats ont été comparés avec ceux obtenus par application du MNE SRTM, montrant que l'utilisation de ce dernier permet d'approximer de manière réaliste l'impact de la non-prise en compte du terrain.

Conclusion de la section

Nous avons vu dans cette section que la non-prise en compte du terrain a pour impact de sous-estimer systématiquement les mesures géométriques de longueur et de surface calculées à partir de la géométrie des objets vectoriels. Cette non-prise en compte du terrain dans les mesures géométriques s'explique pour différentes raisons, principalement liées à la modélisation de l'information altimétrique dans la géométrie des objets vectoriels, ou encore à l'absence d'algorithmes adaptés dans les logiciels pour permettre des mesures géométriques en 2D5.

Cette section a donc permis de présenter des méthodes permettant :
- d'enrichir (correctement) la géométrie des objets vectoriels d'une information altimétrique lorsque ceux-ci n'en disposent pas,
- de réaliser des mesures géométriques de longueur et de surface en 2D5.

Les expérimentations réalisées dans cette section ont permis de montrer que l'erreur de mesure impliquée par la non-prise en compte du terrain est sensible à la modélisation de l'altitude dans les modèles numériques. Cette erreur de mesure dépend notamment de la résolution spatiale, ou encore du type de modèle numérique exploité (MNT ou MNE).

Ainsi, après avoir étudié l'impact de la projection cartographique, nous avons pu voir à travers cette section consacrée à l'impact de la non-prise en compte du terrain, comment le processus d'abstraction du monde réel affecte les mesures géométriques de longueur et de surface par une représentation d'objets vectoriels dans un plan. Dans le cadre de ces deux sections, nous n'avons pas pour l'instant remis en cause la géométrie initiale des objets vectoriels, mais nous sommes contentés de l'enrichir d'information permettant d'estimer l'impact de ces deux règles de représentation. La section suivante intègrera cette notion de remise en cause de la géométrie initiale, en étudiant l'impact de l'approximation polygonale des courbes.

3.3 Approximation polygonale des courbes

Le processus d'abstraction implique de nombreuses simplifications afin de représenter les entités du monde réel sous forme d'objets dans une base de données géographiques vectorielles. Nous avons vu que ces simplifications se manifestent notamment en termes de niveau de détail géométrique. Celui-ci doit satisfaire des exigences relatives à l'échelle de représentation attendue, de volume de stockage, ou encore à la facilité de manipulation des objets (comme le temps de traitement par exemple). La représentation des entités courbes du monde réel est particulièrement conditionnée par ces exigences.

En effet, dans une base de données vectorielles, la géométrie des objets linéaires ou surfaciques est généralement constituée d'une succession de points, reliés entre eux par des segments de droites, appelés plus classiquement polylignes (ou lignes polygonales). Or, lorsque les entités du monde réel sont des courbes, la représentation vectorielle sous forme de ligne polygonale permet uniquement d'approximer la courbe réelle. C'est ce que nous appelons ici l'approximation polygonale des courbes, dont le degré varie en fonction du niveau de détail de la base de données. Cette représentation des courbes dans les bases de données vectorielles impacte les mesures géométriques de longueur, puisque la longueur de la ligne polygonale sous-estime systématiquement la longueur de la courbe (en partant du principe que les points d'une polyligne sont localisés exactement sur la courbe). Dans ce contexte, et dans le but d'élaborer un modèle général d'estimation de l'imprécision des mesures géométriques, il est nécessaire de mettre en œuvre des méthodes permettant d'évaluer l'impact de l'approximation polygonale, puisque de nombreuses entités du monde réel présentent des géométries courbes, comme les routes par exemple.

Ainsi, nous présenterons tout d'abord le principe général de l'approximation polygonale des courbes et son impact sur les mesures géométriques, puis proposerons des méthodes afin d'estimer cet impact sur les mesures de longueur, et terminerons par une mise en application du modèle.

3.3.1 L'approximation polygonale et son impact

Nous allons présenter dans cette section le principe d'approximation polygonale, et sa particularité dans le cadre des bases de données géographiques vectorielles. Nous évoquerons ensuite son impact sur les mesures géométriques de longueur et identifierons des solutions permettant de l'estimer.

3.3.1.1 L'approximation polygonale des courbes : principe général

Définition générale Garnesson et Giraudon (1992) définissent l'approximation polygonale comme l'*opération qui consiste à transformer une chaine de points connexes en une suite de segments de droite*. L'approximation polygonale est une étape classique en vision par ordinateur, si les caractéristiques géométriques des objets doivent être prises en compte. L'objectif d'une opération d'approximation polygonale est de réduire la quantité d'information (et donc le volume des fichiers) afin de ne conserver que le nécessaire pour la compréhension d'un phénomène ou son traitement. De nombreuses méthodes permettent de procéder à une opération d'approximation polygonale, comme l'algorithme Dunham (1986), ou encore celui proposé par Wall et Danielsson (1984).

La figure 3.30 permet d'illustrer l'algorithme de Wall et Danielsson (1984). Cet algorithme a pour objectif de créer une ligne polygonale à partir d'une courbe, en minimisant les aires inscrites entre elles.

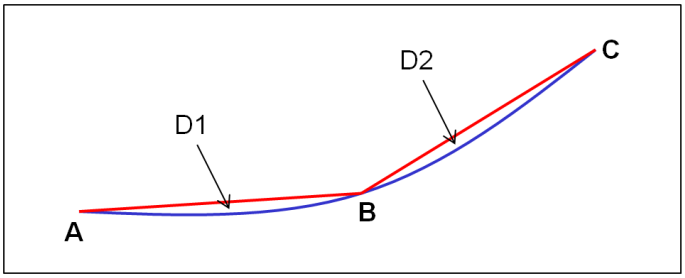


Figure 3.30 – Illustration de l'algorithme de Wall et Danielsson (1984)

Particularité des objets géographiques Les objets géographiques vectoriels sont eux, par définition, affectés par ce principe d'approximation polygonale. En effet, les primitives linéaires et surfaciques, sont, dans l'extrême majorité des cas, construites à partir de points reliés entre eux par des segments de droites. Toutefois, la géométrie des objets géographiques vectoriels peut également être construite à partir de points reliés par des courbes plus complexes, comme des courbes de Bézier ou des splines (Bonin, 2002), qui permettraient de représenter de manière plus réaliste les entités courbes. Cependant ceci n'est généralement pas le cas pour différentes raisons. Comme nous l'avons exposé, la modélisation géométrique des objets vectoriels est réalisée classiquement sous forme de lignes polygonales afin de répondre à des exigences en termes de capacité de stockage, et de facilité de traitement. De plus, la majorité des logiciels SIG ne prend pas en compte la modélisation de la géométrie des objets vectoriels sous forme de courbes polynomiales, au contraire des logiciels de DAO. Ainsi, la géométrie des objets géographiques vectoriels est systématiquement affectée par l'approximation polygonale, qui, comme nous allons le voir, impacte les mesures géométriques, en particulier lorsque l'entité représentée est une courbe (cf. figure 3.31).

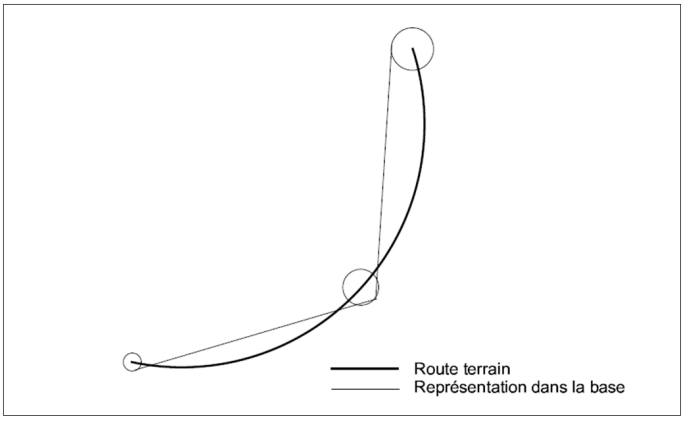


Figure 3.31 – Une route courbe du monde réel représentée par une ligne polygonale dans une base de données, d'après Bonin (2002)

Impact sur les mesures géométriques L'approximation polygonale de la géométrie des objets vectoriels ne pose pas de problèmes lorsque les entités du monde réel à représenter sont rectilignes, ou présentent une angularité bien marquée. Les mesures géométriques calculées seront généralement bien préservées. Par contre, lorsque les entités du monde réel à représenter sont courbes, l'approximation polygonale sous-estime systématiquement les longueurs calculées, puisque par définition, la corde (le segment de droite) est plus courte que l'arc (la courbe du monde réel).

Ainsi, l'approximation polygonale impacte directement les mesures de longueur calculées à partir de la géométrie des objets vectoriels. Ceci peut également être le cas pour les mesures de surface, mais nous ne les aborderons pas dans cette section. L'élaboration d'un modèle d'estimation de l'imprécision des mesures géométriques suppose donc de mettre en œuvre des méthodes permettant de modéliser l'impact de l'approximation polygonale, lorsque les objets évalués représentent des entités courbes du monde réel.

3.3.1.2 Degré d'approximation polygonale

Dans les bases de données géographiques vectorielles, l'opération d'approximation polygonale des entités courbes intervient lors du processus de production de la géométrie des objets géographiques, soit en saisie initiale, soit en post-traitement. C'est généralement l'opérateur de saisie qui définit lui-même le degré d'approximation polygonale lors de la saisie des objets (dans le cadre de relevés photogrammétriques ou lors de la digitalisation manuelle de cartes ou de plans). Cependant, certaines méthodes automatiques peuvent également être utilisées, comme l'algorithme de Douglas et Peucker (1973), utilisé afin de simplifier une géométrie (en éliminant des points de la ligne initiale) tout en conservant sa forme générale. L'application de cet algorithme suppose l'existence d'une géométrie à simplifier, provenant de relevés existants. On notera que l'application excessive de l'algorithme de Douglas-Peucker peut faire perdre la forme initiale de l'objet à représenter, comme cela a été souligné dans de nombreuses contributions (voir par exemple Visvalingam et Whyatt (1991) et Mustière (2001)).

Granularité et degré d'approximation polygonale Le degré d'approximation polygonale des objets géographiques est étroitement conditionné par le niveau de détail géométrique de la base de données. En effet, pour satisfaire des exigences en termes de représentation, de volume de stockage, ou encore de facilité de traitement, la géométrie des objets vectoriels subit de nombreuses simplifications, et ceci est particulièrement le cas pour la modélisation géométrique des courbes.

En effet, le niveau de détail géométrique peut être qualifié par la granularité des objets géographiques, c'est-à-dire la taille des plus petites formes géométriques présentes dans la base de données (Ruas, 2002; Vauglin, 2002). La granularité se manifeste sur les distances entre points successifs d'une géométrie. Ces distances inter-points ont généralement tendance à diminuer lorsque le niveau de détail de la base de données augmente. Dans le cas de la représentation d'entités courbes, la granularité conditionne directement le réalisme de la modélisation par une ligne polygonale, puisque plus la granularité est importante, plus la distance entre points successifs diminue, et plus la polyligne approxime la courbe (en partant du principe que chacun des points est correctement positionné sur la courbe).

La figure 3.32 illustre l'impact du degré d'approximation polygonale d'un virage représenté par la BDTOPO et la BDCARTO. Outre les problèmes d'exactitude de position, on voit bien avec cette illustration que la distance entre points successifs impacte directement le réalisme de la modélisation du virage, et par conséquent la longueur du tronçon de route.

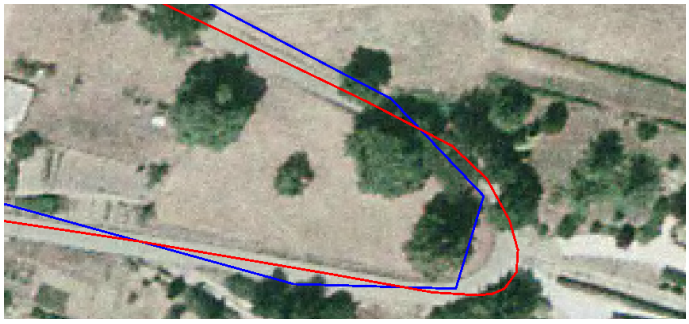


Figure 3.32 – Impact de la granularité sur la représentation d'un virage par la BDTOPO (en rouge) et la BDCARTO (en bleu)

La granularité des objets géographiques se révèle donc être un indicateur pertinent pour caractériser le degré d'approximation polygonale des entités courbes. Comme nous avons pu le voir, ce degré d'approximation polygonale influence directement le réalisme de la représentation des courbes, mais également la justesse de la mesure de longueur.

3.3.1.3 Solutions envisagées pour estimer son impact

Plusieurs contributions ont proposé des méthodes permettant de modéliser les erreurs de positionnement impliquées par la représentation des entités courbes (voir par exemple Tong *et al.* (2003) et Tong et Shi (2010)). Il est cependant délicat, à partir d'une géométrie existante, d'estimer l'impact de l'approximation polygonale d'une courbe sur les mesures géométriques, étant donné que le degré d'approximation varie en fonction du niveau de détail recherché, et de la forme des objets. Nous proposons donc une méthode d'estimation qui prend en compte ces deux caractéristiques de la géométrie des objets évalués.

Principe général Pour estimer l'impact de l'approximation polygonale, notre démarche consiste à mettre en œuvre des méthodes permettant de reconstruire une géométrie courbe réaliste à partir d'objets approximés sous forme de ligne polygonale (cf. figure 3.33).

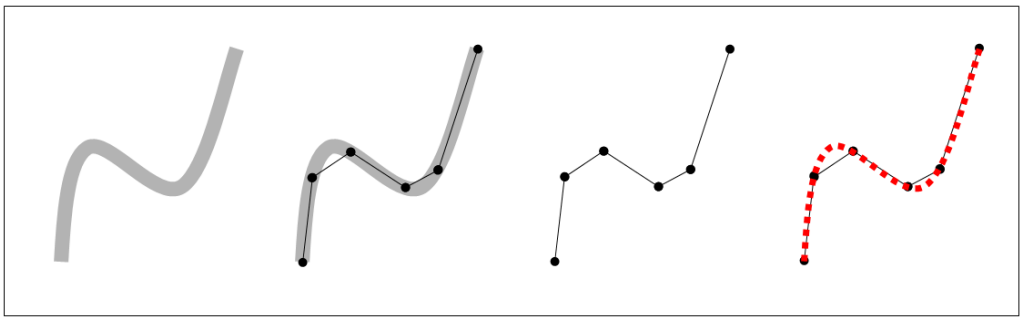


Figure 3.33 – Une entité courbe du monde réel (en gris), représentée sous forme de ligne polygonale (en noir), sur laquelle est reconstruite une géométrie courbe réaliste (en rouge)

Pour cela, des méthodes d'interpolation polynomiale, comme les courbes de Bézier ou les splines, peuvent être exploitées. En utilisant cette approche, il nous sera possible d'estimer l'erreur de mesure géométrique conditionnée par l'approximation polygonale.

Problèmes rencontrés La mise en œuvre de cette méthode d'évaluation se heurte à deux difficultés principales. La première réside dans la recherche de méthodes d'interpolation numériques facilement implémentables et paramétrables afin de permettre la reconstruction de géométries courbes réalistes. La seconde difficulté liée à la reconstruction de géométries courbes réalistes fait référence aux objets, voire aux portions d'objets auxquelles appliquer ces méthodes. En effet, tous les objets d'une base de données ne représentent pas des entités courbes du monde réel, et par conséquent ne supposent pas une reconstruction de géométrie courbe. Par exemple, une même route peut être constituée de zones de virages et de zones à caractère rectiligne. Pour réaliser une reconstruction réaliste, nous devons chercher à appliquer une méthode d'interpolation uniquement sur les zones de virages, et non sur les zones rectilignes.

Nous proposerons donc dans la section suivante des méthodes permettant de reconstruire des géométries courbes réalistes, ainsi que des méthodes de sélection des objets sur lesquels les appliquer (cf. figure 3.34).

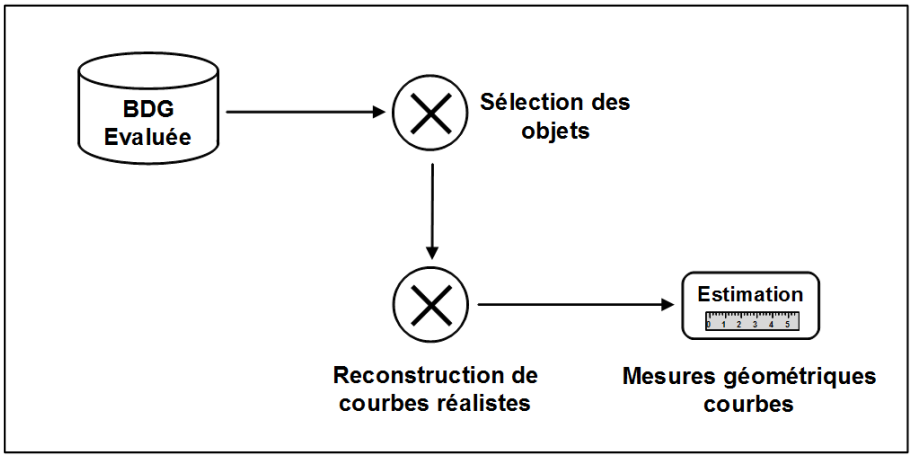


Figure 3.34 – Principe d'estimation de l'impact de l'approximation polygonale des courbes

Cette étude sur l'impact de l'approximation polygonale des courbes sur les mesures géométriques ne sera ici traitée que sur les primitives géométriques linéaires, et par conséquent sur les mesures de longueur. Pour l'illustrer, nous utiliserons des géométries extraites de la classe d'objets des tronçons de route.

Synthèse Section La représentation des entités courbes du monde réel dans une base de données géographiques implique une opération d'approximation polygonale, qui a pour conséquence de sous-estimer systématiquement les longueurs mesurées. Le degré d'approximation polygonale varie notamment en fonction du niveau de détail de la base de données. Afin d'estimer l'impact de l'approximation polygonale sur les mesures géométriques de longueur, nous proposons une méthode basée sur la reconstruction des géométries courbes, par le biais de méthodes d'interpolation polynomiale.

3.3.2 Proposition de méthodes de modélisation réaliste des courbes

Nous présenterons dans cette section les méthodes utilisées pour estimer l'impact de l'approximation polygonale, à partir d'une reconstruction de géométries courbes. Nous chercherons notamment à préserver au maximum le réalisme des objets évalués. Ainsi, nous présenterons tout d'abord les méthodes d'interpolation polynomiale classiquement utilisées pour reconstruire une géométrie courbe à partir d'une ligne polygonale. Ensuite, afin de préserver le réalisme des courbes reconstruites, une méthode permettant d'éviter les exagérations locales sera proposée. Enfin, étant donné que tous les objets ne représentent pas systématiquement des entités courbes du monde réel, nous proposerons une méthode permettant de détecter les objets sinueux dans les tronçons de route, afin d'estimer l'impact de l'approximation polygonale uniquement sur les objets pertinents.

3.3.2.1 Méthodes d'interpolation polynomiale

Deux méthodes d'interpolation seront présentées successivement dans cette section, les courbes de Bézier, et les splines cubiques d'Hermite. Bien que d'autres études sur la modélisation des géométries courbes aient été proposées (Affholder, 2002), notre choix s'est porté sur ces deux méthodes d'interpolation polynomiale puisqu'elles sont couramment utilisées et demeurent facilement implémentables, tout en fournissant des résultats performants.

Les courbes de Bézier

Les courbes de Bézier (du nom de l'ingénieur français Paul Bézier) sont des courbes polynomiales paramétriques définies à l'aide de points de contrôle. Pour définir une courbe de Bézier cubique, quatre points P_0, P_1, P_2 et P_3 sont nécessaires. La courbe se trace en partant du point P_0 pour arriver au point P_3 en suivant la direction définie par les points P_1 et P_2, comme illustré dans la figure 3.35.

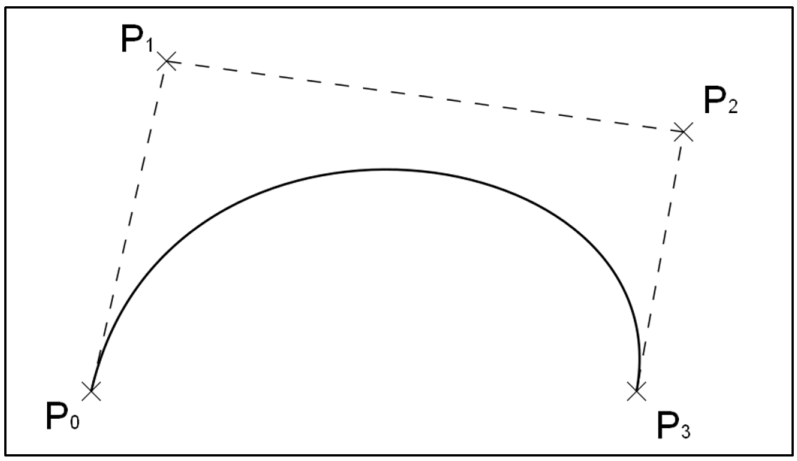


Figure 3.35 – Une courbe de Bézier cubique définie par les points de contrôle P_0, P_1, P_2 et P_3

Le principal problème induit par les courbes de Bézier, tel qu'illustré dans la figure 3.35, est qu'elle ne passe pas par les points P_1 et P_2. Ces deux points sont utilisés pour la construction géométrique de la courbe de Bézier selon la méthode de Casteljau (cf. figure 3.36).

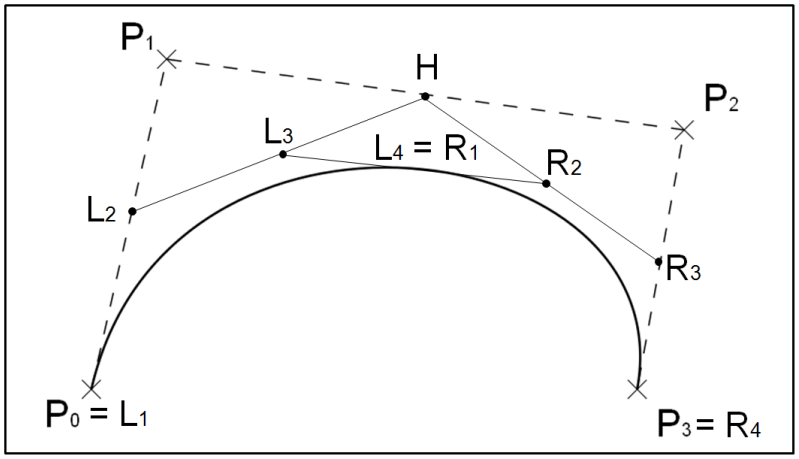


Figure 3.36 – Construction d'une courbe de Bézier cubique définie par les points de contrôle P_0, P_1, P_2 et P_3 selon la méthode de Casteljau

Afin de reconstruire des géométries courbes plus réalistes sur les objets géographiques affectés par l'approximation polygonale, l'application des courbes de Bézier cubiques offre une première solution intéressante. Cependant, pour reconstruire des courbes réalistes, il est nécessaire que la courbe interpolée passe par chacun des points de contrôle, ce qui n'est pas le cas dans les exemples proposés. Ainsi, une solution pour reconstruire des géométries courbes, en passant par tous les points de contrôle, est d'utiliser une autre méthode d'interpolation, les splines cubiques d'Hermite.

Les splines cubiques d'Hermite

L'interpolation d'Hermite (du nom du mathématicien français Charles Hermite) est une méthode d'interpolation polynomiale, qui, contrairement aux courbes cubiques de Bézier, passe par chacun des points de contrôle, en fonction des tangentes qui leur sont associées, comme illustré en figure 3.37.

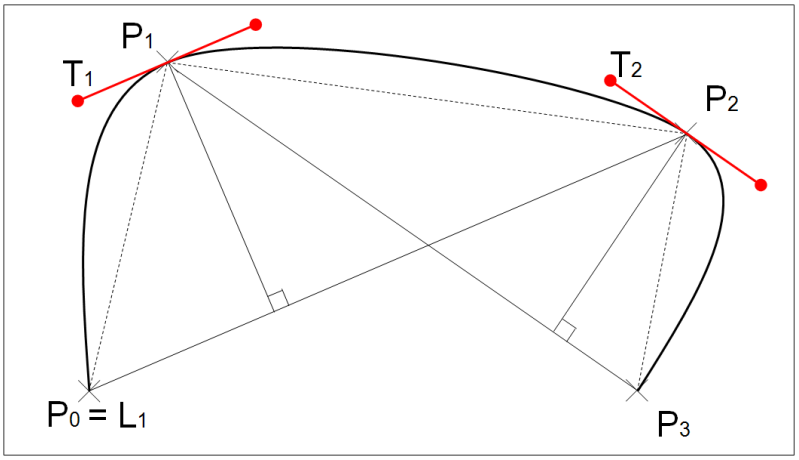


Figure 3.37 – Construction d'une interpolation par splines cubiques d'Hermite passant par les points de contrôle P_0, P_1, P_2 et P_3 selon les tangentes T_0, T_1, T_2 et T_3

Le choix du paramétrage des tangentes conditionne directement le résultat d'une interpolation par splines cubiques d'Hermite. Différentes méthodes sont proposées afin de paramétrer les tangentes, notamment la méthode proposée par Kochanek et Bartels (1984). Cette méthode permet de modifier le comportement des tangentes selon trois paramètres, compris entre 0 et 1 :
- la *tension* t, qui permet de paramétrer la longueur du vecteur tangent, et donc la tension de la courbe,
- la *continuité* c, qui permet de paramétrer la rudesse du changement entre deux vecteurs,
- le *biais* b, qui permet de paramétrer la direction du vecteur tangent.

Plusieurs niveaux de spécialisation de la méthode de Kochanek et Bartels (1984) sont proposés afin de définir le comportement des tangentes, notamment :
- les *Catmull-Rom splines*, dans le cas où le paramètre de biais est fixé à 0,
- les *Cardinal splines*, dans le cas où les paramètres de continuité et de biais sont fixés à 0
- la *Différence finie*, dans le cas où les trois paramètres sont fixés à 0 (la longueur de la tangente T_i est dans ce cas égale à la moitié de la longueur $P_{i-1}P_{i+1}$).

Le paramétrage du comportement des tangentes conditionne directement le résultat de l'interpolation, et surtout l'aspect réaliste de la courbe reconstruite à partir des points de contrôle. Il est donc nécessaire d'avoir bien conscience de l'impact de chacun de ces paramètres sur l'aspect final de l'interpolation. Cependant, étant donné qu'aucun a priori sur les paramètres de continuité et de biais entre points successifs n'est introduit dans nos interpolations, nous nous limiterons aux Cardinal splines dans les expérimentations suivantes, qui ne nécessitent qu'un paramètre de tension. La définition du paramètre de tension offrant le meilleur niveau de réalisme sur un réseau routier ne peut par contre être réalisé que par validation visuelle.

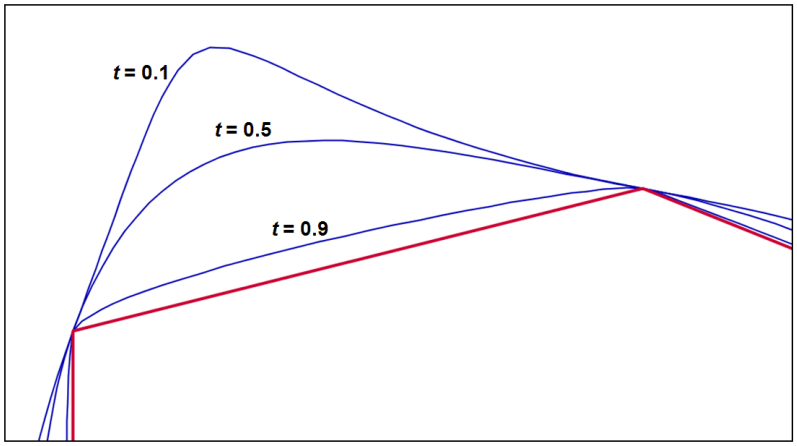


Figure 3.38 – Impact du paramètre de tension sur la reconstruction d'une courbe

La figure 3.38 illustre l'impact du paramètre de tension sur la reconstruction de la géométrie courbe d'un virage à l'aide d'une interpolation par Cardinal splines. On observe que plus le paramètre de tension est important, plus l'interpolation "colle" avec la polyligne initiale. A l'inverse, plus le paramètre de tension est faible, plus la courbe est exagérée, ce qui génère des géométries irréalistes. Dans le cadre des futures expérimentations, nous étudierons la sensibilité des mesures de longueur en fonction de ce paramètre de tension.

Ainsi, pour reconstruire des géométries courbes réalistes afin d'estimer l'impact de l'approximation polygonale, la méthode d'interpolation polynomiale offrant le meilleur compromis s'est portée sur les splines cubiques d'Hermite, en utilisant la méthode des Cardinal splines pour le paramétrage des tangentes. Cette méthode a finalement été retenue pour deux raisons :
– elle passe par l'ensemble des points de contrôle, au contraire des courbes de Bézier,
– elle ne nécessite qu'un paramètre de tension, assez facilement maitrisable.

Cependant, la mise en œuvre de reconstructions de courbes par interpolations polynomiales demeure un problème complexe, puisque tous les objets n'y sont pas systématiquement assujettis, et que dans certains cas de figure, l'utilisation d'interpolations polynomiales peut conduire à la création de courbes exagérées, ne proposant pas l'aspect réaliste que nous cherchons conserver. Ces problèmes et les solutions envisagées pour les résoudre seront présentés dans les sections suivantes.

3.3.2.2 Préservation du réalisme des courbes

La méthode d'estimation de l'impact de l'approximation polygonale que nous proposons repose sur une reconstruction réaliste des géométries courbes. Or, l'application stricte d'une reconstruction de géométrie courbe par la méthode des splines cubiques d'Hermite peut avoir pour effet de générer des courbes irréalistes. Nous avons déjà vu que l'impact du paramètre de tension devait être pris en compte afin de veiller à ne pas générer des courbes exagérées. Mais dans certains cas de figure, même avec un paramètre de tension fort, la configuration intrinsèque de l'objet peut conduire à la génération de courbes irréalistes. Aprés avoir présenté les causes de ce problème, nous proposons dans cette section une méthode permettant d'y remédier

Présentation du problème La distance entre les points successifs d'une géométrie n'est généralement pas constante au sein d'une base de données. Nous avons d'ailleurs souligné que cette distance inter-points a généralement tendance à baisser dans les zones sinueuses, afin de représenter plus fidèlement les courbures (cf. figure 3.39). A l'inverse, elle a plutôt tendance à augmenter dans les zones rectilignes, où l'opérateur peut se contenter de localiser les points aux extrémités d'une ligne droite.

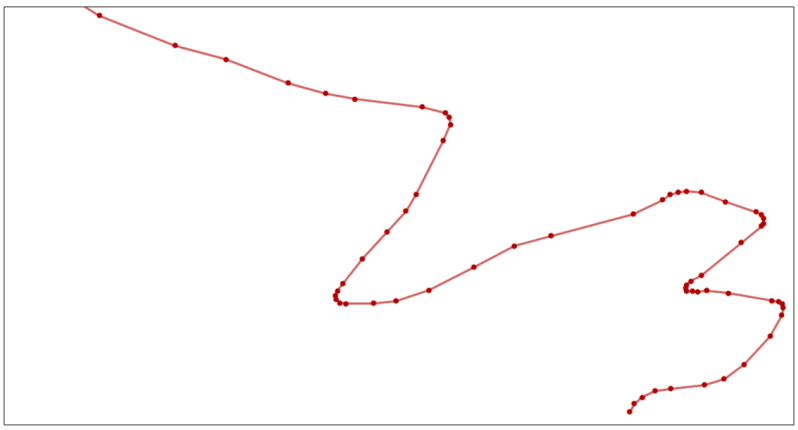


Figure 3.39 – Hétérogénéité des distances inter-points d'un réseau routier

Cette hétérogénéité des distances inter-points peut poser problème en vue de reconstruire une géométrie courbe par interpolation polynomiale. En effet, lorsque se succèdent des segments courts (dans les zones de virages par exemple) suivis de segments longs (dans une ligne droite), la reconstruction de la courbe par interpolation polynomiale risque de présenter une forme irréaliste, comme l'illustre la figure 3.40.

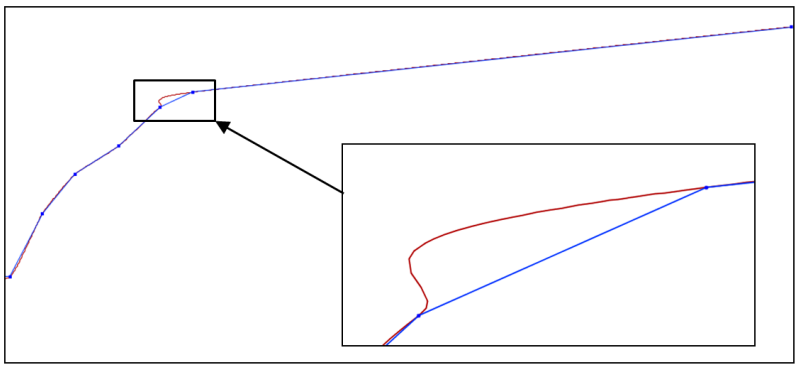


Figure 3.40 – Exemple de reconstruction de courbe irréaliste

En effet, dans ce cas de figure, la tangente associée au sommet de transition entre deux segments de longueurs hétérogènes oblige l'interpolation polynomiale à prendre une forte courbure, et ainsi une forme irréaliste. Au delà de l'irréalisme de la courbe reconstruite, impliqué par l'application "brute" d'une interpolation polynomiale sur la géométrie des objets évalués, la longueur de la géométrie courbe est également surestimée. Nous allons donc proposer une méthode permettant d'éviter ce type d'effets indésirables.

Proposition Pour remédier à ce problème d'interpolation, comme le propose Kunt (2000), une segmentation des objets par analyse de données doit être opérée préalablement, afin de reconstruire les courbes uniquement dans les sections pertinentes. Ainsi, lorsque se succèdent des segments de longueurs hétérogènes, un **suréchantillonnage de la géométrie** de l'objet permet de diviser un segment de longueur importante en une succession de segments alignés. L'application du suréchantillonnage a pour effet d'éviter à la courbe reconstruite de prendre une trop forte courbure, permettant ainsi de générer des courbes réalistes (cf. figure 3.41).

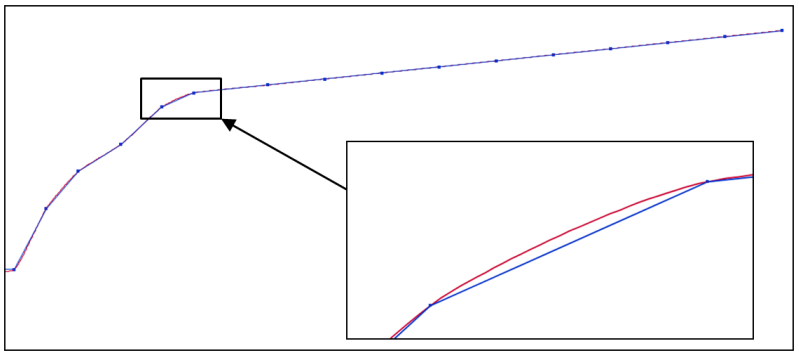


Figure 3.41 – Reconstruction d'une courbe réaliste suite au suréchantillonnage de la polyligne

Pour mettre en œuvre cette méthode, il est cependant nécessaire de définir le pas utilisé pour le suréchantillonnage. Pour cela, nous proposons une solution basée sur l'exploration des distances inter-points de la base de données évaluée.

Exploration des distances inter-points L'objectif de cette méthode est de déterminer une valeur pertinente afin de suréchantillonner les objets évalués. Comme le montre la figure 3.42, la distribution des longueurs des segments des tronçons de route de la BDCARTO est très étirée, ce qui prouve que les distances inter-points sont très hétérogènes au sein d'une base de données géographiques.

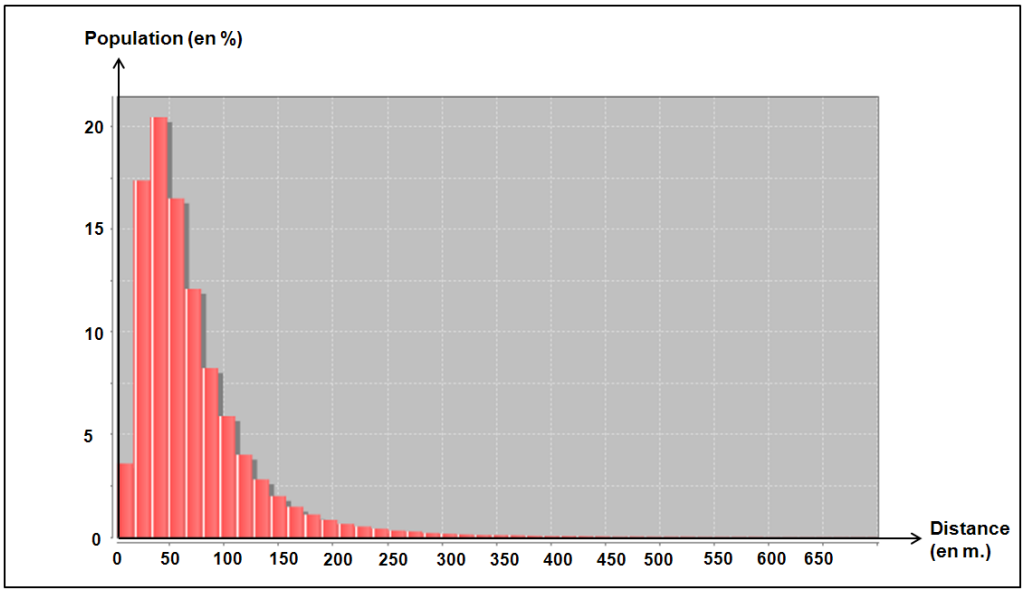


Figure 3.42 – Distribution des longueurs de segments des tronçons de route de la BDCARTO dans le département des Pyrénées-Atlantiques

Cependant, comme nous l'avons déjà souligné, nous pouvons supposer que les segments les plus courts correspondent à des sections courbes (des virages), et les plus longs à des sections rectilignes (des lignes droites). Ainsi, à la vue de cette distribution, l'exploitation de la longueur de segment médiane peut s'avérer pertinente, puisqu'elle évite de prendre en compte les valeurs extrêmes, au contraire de la longueur de segment moyenne.

Méthodologie Pour paramétrer le suréchantillonnage des objets sur lesquels appliquer une reconstruction de courbe par interpolation polynomiale, nous proposons donc une méthode en trois étapes :

1. Exploration des distances inter-points de la base de données,
2. Suréchantillonnage des objets par la longueur de segment médiane,
3. Reconstruction des courbes par interpolation polynomiale.

Cette méthodologie permettant de préserver le réalisme des courbes reconstruites sera présentée dans la cadre d'une expérimentation sur une extraction d'un réseau routier. Préalablement à cela, nous pouvons nous interroger sur l'intérêt d'appliquer une reconstruction de géométrie courbe sur l'ensemble des objets d'une base de données.

En effet, dans une base de données géographiques, et notamment dans l'exemple d'un réseau routier, tous les objets, voire certaines portions d'objets, ne représentent pas des entités courbes du monde réel. Certaines sections d'une même route peuvent par exemple présenter une forme rectiligne (en milieu urbain notamment) et d'autres une forme courbe. Or, dans le cadre de ce travail, il serait préférable d'estimer l'impact de l'approximation polygonale uniquement sur les objets représentant des entités courbes du monde réel.

Pour cela, nous proposons de détecter préalablement les objets pertinents, sur lesquels appliquer une reconstruction de géométrie courbe. Dans le cas des tronçons de route, nous pouvons considérer que les objets sinueux sont à ce titre particulièrement pertinents. Ainsi, une méthode permettant de détecter automatiquement les routes sinueuses sera proposée dans la section suivante.

3.3.2.3 Détection des objets sinueux

Au sein d'une même base de données géographiques, des objets représentant des entités courbes du monde réel, comme les virages par exemple, peuvent cohabiter avec d'autres objets ne représentant pas des entités courbes, comme des lignes droites ou des angles de rues. Afin d'être réaliste, la reconstruction de géométries courbes à l'aide des méthodes d'interpolation présentées précédemment ne doit donc être appliquée qu'aux objets le nécessitant. Ceci pose donc la question de la détection des objets représentant des entités courbes du monde réel.

En nous appuyant sur l'exemple d'un réseau routier, nous pouvons supposer que les objets présentant un caractère sinueux représentent des entités courbes du monde réel. En nous appuyant sur des algorithmes existant de détection de la sinuosité dans un réseau routier, comme celui proposé par Plazanet (1996) (cf. figure 3.43), ainsi que sur la qualification des virages (Mustière, 2001), nous proposons de mettre au point une méthode permettant de n'extraire que les objets à caractère sinueux, pour ensuite reconstruire des géométries courbes.

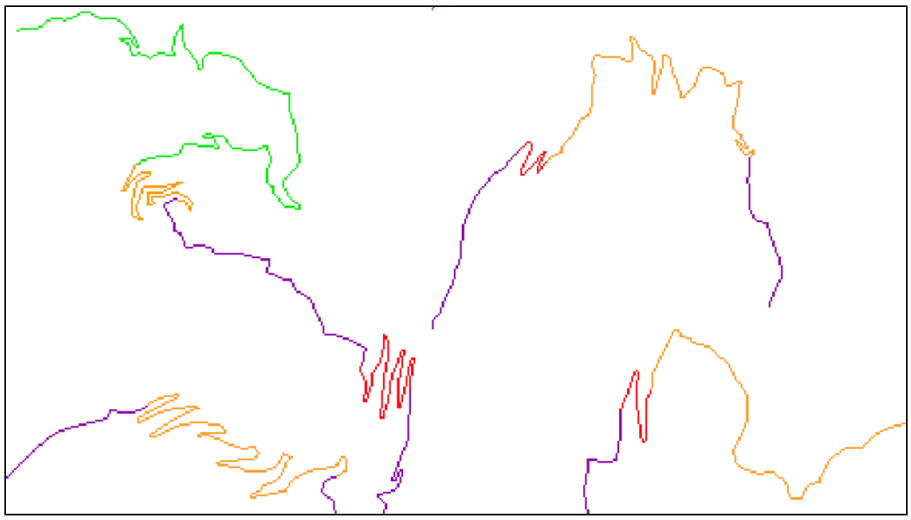


Figure 3.43 – Classification de la sinuosité de portions de routes, d'après Plazanet (1996)

Pour mettre en œuvre cette méthode de qualification de la sinuosité sur un réseau routier, la méthodologie proposée repose sur la détection préalable des points d'inflexion de la polyligne.

Points d'inflexion Les points d'inflexion sont définis comme les points de la géométrie d'une polyligne où l'orientation de la courbure change de signe. Cependant, afin de ne pas prendre en compte toutes les micro-inflexions, la polyligne doit subir préalablement un lissage, en utilisant par exemple l'algorithme du lissage gaussien.

Lissage gaussien des polylignes

Le lissage gaussien est un algorithme qui a pour but d'éliminer les détails sans importance d'une polyligne. C'est un algorithme très utilisé dans le domaine de la généralisation cartographique du linéaire, comme par exemple dans le travail de Plazanet (1996).

Le fonctionnement du lissage gaussien repose selon le principe d'une fenêtre glissante sur l'ensemble de la polyligne (Mustière, 2001). Chaque point de la polyligne est déplacé vers le barycentre calculé avec les autres points localisés dans la fenêtre, en pondérant le déplacement en fonction de la distance curviligne à chaque point. La force du lissage est déterminée par le paramètre σ, correspondant à l'écart-type de la loi gaussienne (cf. figure 3.44). Plus la valeur σ est importante, plus les points les plus éloignés sont pris en compte, et par conséquent plus le lissage est important.

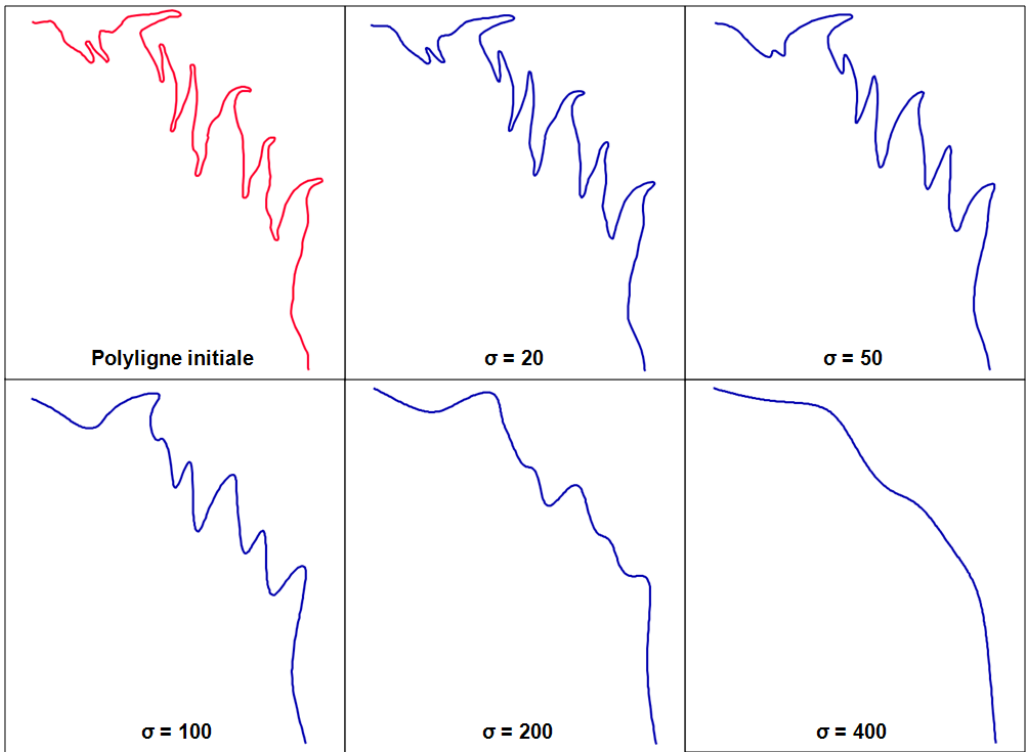


Figure 3.44 – Effet du lissage gaussien sur une polyligne pour différentes valeurs de σ

L'intérêt du lissage gaussien, à travers l'élimination des détails sans importance d'une polyligne, est que cette dernière conserve exactement tous les points de sa géométrie. Aucun point n'est éliminé, contrairement à un filtrage, comme celui de Douglas et Peucker (1973).

Ainsi, l'application d'un léger lissage est un préalable indispensable à la détection des principaux points d'inflexion d'une polyligne, et s'affranchir des micro-inflexions. Cependant, plus la force du lissage de la polyligne est importante, moins de points d'inflexion seront détectés. Il est donc important de bien calibrer le lissage afin de ne pas trop lisser la polyligne et ainsi éliminer des informations qui pourraient s'avérer utiles pour qualifier sa sinuosité.

Détection des points d'inflexion

Principe de détection Pour détecter les points d'inflexion, la polyligne lissée est parcourue en chacun de ses points, excepté ses extrémités. Pour chacun des points parcourus, le signe de l'orientation de la courbure est calculé, puis stocké dans un tableau en prenant les valeurs "G" (le virage tourne vers la gauche) si l'angle est supérieur à π, et "D" (le virage tourne vers la droite) si l'angle est inférieur à π.

A partir de ce tableau, un nouveau tableau récapitulant les séquences de points successifs ayant le même signe d'orientation de la courbure est ensuite crée. Par exemple, si la polyligne est constituée de 6 points successifs ayant la valeur "G", puis de 17 points ayant la valeur "D", le tableau de séquences prendra pour valeurs [6, 17]. Ce tableau de séquences permet d'identifier les points d'inflexion de la polyligne, en prenant le premier point de chaque séquence (cf. figure 3.45).

```
Séquence = [6,17,33,28,20,18,3,8,30,22,17,16,20,28,44,14,10,31]

ID Inflexion = [0,6,23,56,84,104,122,125,133,163,185,202,218,238,266,310,324,334,365,388]
```

Figure 3.45 – Tableaux de séquences et d'identification des points d'inflexion d'une polyligne

Cas des micro-inflexions Il est possible, selon la force du lissage utilisé, que des séquences d'orientation de courbure ne soient composées que d'un seul point. Ces séquences sont appelées **micro-inflexions**. Elles doivent être supprimées puisqu'elles ont pour effet de perturber la méthode d'analyse de la sinuosité, qui s'attache à détecter les changements d'orientation principaux.

Application de la méthode de détection La figure 3.46 représente les points d'inflexion détectés sur un tronçon de route de la BDTOPO, suite à un lissage gaussien de paramètre $\sigma = 85$ m. Les résultats sont globalement satisfaisants, bien que des cas de sur-détection et de sous-détection de points d'inflexion soient observés. Cette méthode de détection reste très dépendante de la force du lissage utilisé, qui conditionne l'ensemble du processus de qualification de la sinuosité des tronçons de route.

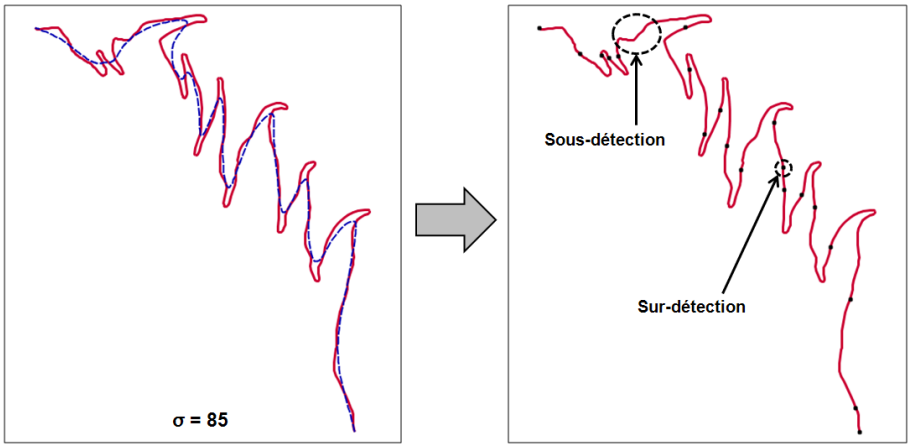


Figure 3.46 – Résultat de la détection des points d'inflexion sur une polyligne de la BDTOPO, suite à un lissage gaussien de paramètres $\sigma = 85$ m.

Une fois les points d'inflexion localisés, il est possible de procéder à la qualification de la sinuosité des tronçons de route. Pour cela, la méthode de qualification des tronçons de route proposée par Mustière (2001) est utilisée.

Qualification de la sinuosité des tronçons de route

L'étude de la sinuosité d'un réseau routier a fait l'objet de nombreux travaux dans le domaine de la généralisation cartographique (Plazanet, 1996; Mustière, 2001). En vue d'appliquer des processus appropriés de généralisation cartographique, Mustière (2001) a élaboré une base de règles permettant de classer la sinuosité des tronçons de route en quatre classes : nulle, hétérogène, virages peu sinueux et épingles à cheveux.

Cette base de règles s'appuie sur deux mesures principales extraites de la géométrie des tronçons de route :
- le rapport **base sur longueur**, qui équivaut au rapport entre la distance euclidienne entre les extrémités de la polyligne, divisée par la longueur de la polyligne (cf. figure 3.47),
- le **nombre de virages**, qui est défini par le nombre de points d'inflexion de la polyligne.

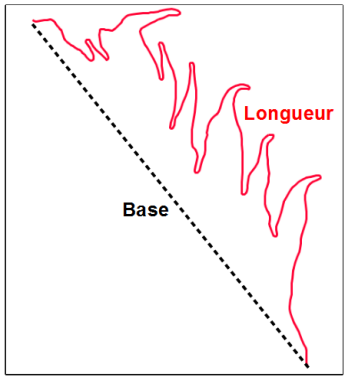


Figure 3.47 – Rapport base sur longueur

A l'aide de ces deux mesures, Mustière (2001) a élaboré une base de règles par apprentissage automatique. Le pseudo code ci-dessous présente la base de règles permettant de qualifier la sinuosité des tronçons de route. Dans ce pseudo-code *nbBends* est le nombre de virages et *baseLength* est le rapport base sur longueur.

```
1  for each road :dbRoad
2     if baseLength < 0.73 then
3          if nbBends <= 15 then
4               road.classe() = "Epingles à cheveux"
5          if nbBends > 15 then
6               road.classe() = "Hétérogène"
7          end if
8     if baseLength >= 0.73 and baseLength < 0.97 then
9          road.classe() = "Virages peu sinueux"
10    if baseLength >= 0.97 then
11         road.classe() = "Nulle"
12    end if
13 end for
```

Application de la base de règles La base de règles est ensuite exploitée afin de détecter les tronçons sinueux, considérés ici comme pertinents en vue de reconstruire une géométrie courbe par le biais de méthodes d'interpolation. On estime ici que les tronçons qualifiés d'*épingles à cheveux* et d'*hétérogènes* correspondent aux objets dont on souhaite évaluer l'impact de l'approximation polygonale des courbes.

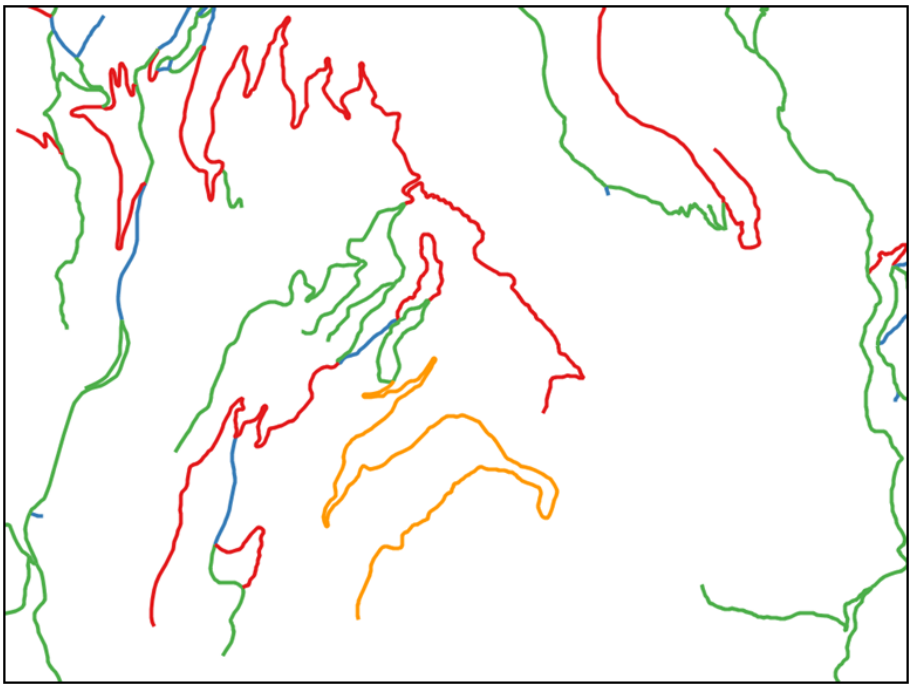


Figure 3.48 – Classification de la sinuosité de tronçons de route, selon la méthode de Mustière (2001) en *sinuosité nulle* (en bleu), *virages peu sinueux* (en vert), *sinuosité hétérogène* (en orange), et *épingles à cheveux* (en rouge)

La figure 3.48 présente un ensemble de tronçons de route dont la sinuosité est classée selon la base de règles de Mustière (2001). Ces résultats montrent que certains tronçons, dont la sinuosité est qualifiée d'*hétérogène* voire d'*épingles à cheveux* présentent en effet une forte sinuosité, mais également des portions peu sinueuses.

Nous savons que la reconstruction de courbe sur des objets ne le nécessitant pas risque de sur-estimer l'impact de l'approximation polygonale, ce qui n'est pas souhaitable. Cependant, la méthode de reconstruction des géométries courbes s'appuie sur un suréchantillonnage (basé sur une exploration des distances inter-points) qui permet d'éviter que les courbes reconstituées ne soient exagérées.

La méthode de détection de la sinuosité présentée dans cette section sera donc intégrée au sein du modèle d'estimation de l'impact de l'approximation polygonale, afin de sélectionner les objets sur lesquels appliquer la méthode de reconstruction des géométries courbes. Cependant, l'application de cette méthode demeurera optionnelle. En effet, d'autres méthodes de sélection des objets pertinents peuvent être envisagées, comme une sélection manuelle par exemple. De plus, l'ensemble de la méthode de détection des objets sinueux est conditionnée par la paramétrage du lissage gaussien, pour lequel il est difficile de déterminer une valeur optimale en fonction de la base de données utilisée. Ceci nous encourage donc à ne pas imposer cette méthode au sein du processus général d'estimation de l'impact de l'approximation polygonale des courbes. Nous préférons laisser l'utilisateur libre de l'utiliser ou non.

Synthèse Section Nous avons vu dans cette section que plusieurs méthodes d'interpolation pouvaient être utilisées afin de reconstruire des géométries courbes, et ainsi estimer l'impact de l'approximation polygonale. Notre choix s'est porté sur les splines cubiques d'Hermite, à travers la méthode des Cardinal Splines. Afin de préserver le réalisme des courbes reconstruites, un suréchantillonnage préalable des lignes polygonales est proposé. Le pas de celui-ci est déterminé suite à une exploration des distances inter-points des objets de la base de données. Enfin, une méthode de détection automatique des objets sinueux est également proposée à titre optionnel, afin de n'appliquer la reconstruction des courbes que sur une sélection d'objets pertinents.

3.3.3 Application de la méthode de reconstruction

Afin d'étudier l'intérêt de la méthode de reconstruction de courbes par interpolation polynomiale, une expérimentation est proposée à partir d'une extraction du réseau routier de la BDCARTO, sur trois communes de montagne des Pyrénées-Atlantiques : Aldudes, Banca, et Urupel (cf. figure 3.49). L'ensemble du réseau sélectionné représente une longueur totale de 191,1 kilomètres.

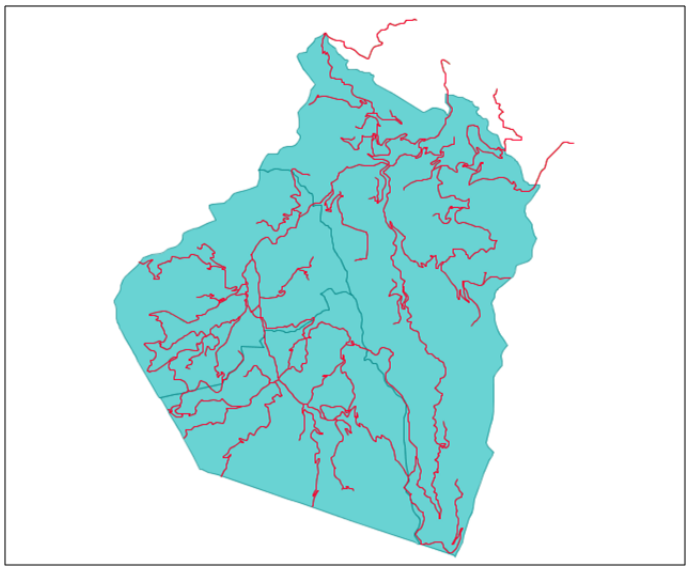


Figure 3.49 – Extraction du réseau routier de la BDCARTO sur les communes de Aldudes, Banca, et Urupel (Pyrénées-Atlantiques)

La sélection du réseau routier de ces trois communes se justifie car elles sont localisées en haute montagne, et le réseau routier y présente ici une sinuosité importante et hétérogène. Nous partons donc du postulat que l'ensemble du réseau routier nécessite une reconstruction de géométrie courbe.

Dans cette expérimentation, nous allons comparer les résultats d'interpolations polynomiales construites avec la méthode des Cardinal Splines (avec et sans suréchantillonnage préalable des polylignes), en faisant varier le paramètre de tension. Nous étudierons l'erreur de longueur en fonction de ce paramètre, mais porterons également attention au réalisme des courbes reconstruites, afin d'estimer la valeur de tension la plus appropriée.

Exploration des distances inter-points La première étape de cette expérimentation consiste à déterminer la valeur utilisée pour le suréchantillonnage des géométries. L'exploration des distances inter-points du réseau routier sélectionné présente une distribution dissymétrique, comme nous pouvons le voir dans la figure 3.50.

Les 3182 segments comptabilisés présentent des longueurs comprises entre 9 mètres et 367 mètres. La longueur médiane est d'environ 51 mètres. C'est cette valeur que nous allons retenir pour suréchantillonner les géométries par la suite.

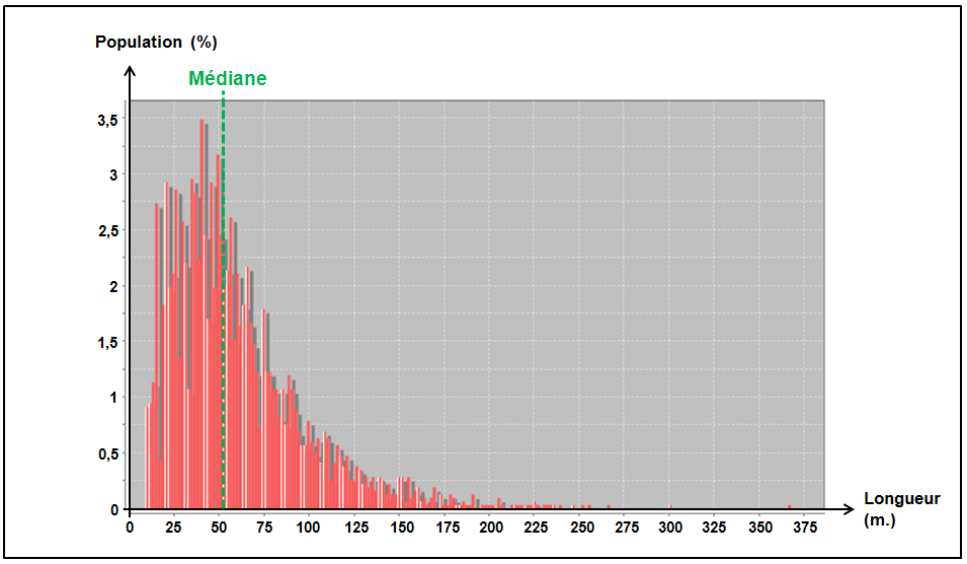


Figure 3.50 – Distribution des distances inter-points du réseau routier expérimenté

Résultats des interpolations La seconde étape de cette expérimentation consiste à reconstruire les géométries courbes par interpolation polynomiale avec la méthode des cardinal splines :
– en faisant varier le paramètre de tension de 0,1 à 0,9,
– avec et sans suréchantillonnage préalable.
Les résultats de l'expérimentation sont présentés dans la table 3.1. Ce tableau présente les différences de longueurs entre le réseau routier initial (sous forme de ligne polygonale) et les reconstructions de courbes avec et sans suréchantillonnage.

	Sans suréchantillonnage			**Avec suréchantillonnage**		
Tension	**Longueur**(km)	**Dif.**(m)	**Dif.**(%)	**Longueur**(km)	**Dif.**(m)	**Dif.**(%)
0,1	199,48	8394	4,39	196,69	5606	2,93
0,2	197,59	6506	3,4	195,52	4432	2,32
0,3	196,04	4952	2,59	194,54	3451	1,81
0,4	194,76	3674	1,92	193,71	2625	1,37
0,5	193,72	2633	1,38	193,01	1925	1,01
0,6	192,88	1793	0,94	192,42	1335	0,69
0,7	192,20	1117	0,58	191,93	843	0,44
0,8	191,67	583	0,3	191,53	442	0,23
0,9	191,28	189	0,1	191,23	142	0,07

Table 3.1 – Différences de longueurs entre la polyligne initiale et les courbes reconstruites avec ou sans suréchantillonnage

Les résultats confirment l'utilité du suréchantillonnage afin d'éviter une sur-estimation trop importante de l'impact de l'approximation polygonale, les différences de longueurs étant en moyenne plus importante de 27% sans suréchantillonnage, du fait des nombreuses reconstructions de courbes irréalistes. Egalement, ces résultats montrent la sensibilité de la différence de longueur (variant de 3% à moins de 0,1%) au paramétrage de la tension, comme le résume le graphique de la figure 3.51.

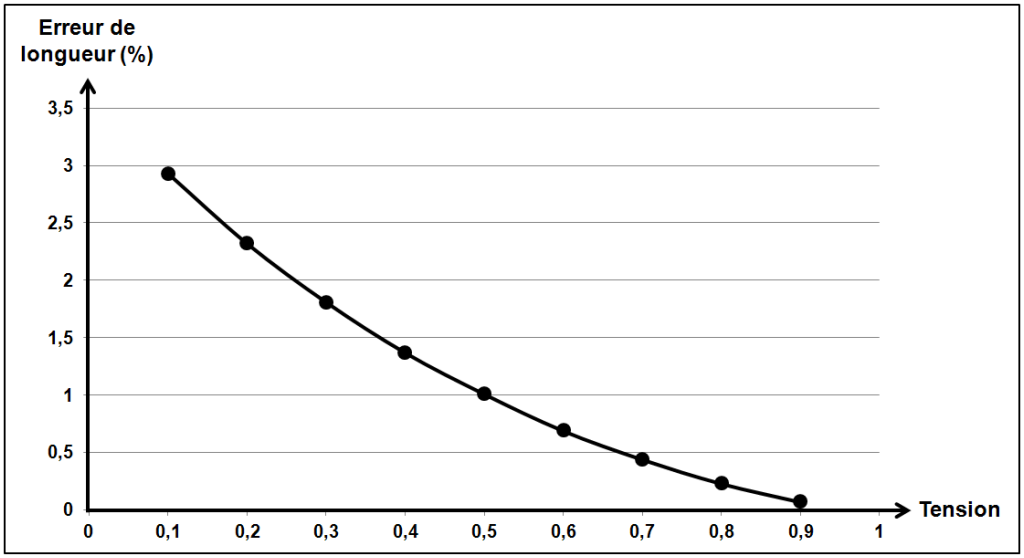


Figure 3.51 – Différence de longueur en fonction de la tension

Réalisme des courbes La dernière étape de cette expérimentation consiste à valider visuellement le réalisme des courbes reconstruites par interpolation polynomiale. Comme l'expose la figure 3.52, le paramètre de tension impacte directement le réalisme de la courbe reconstruite, et par conséquent la longueur calculée. On y voit que les interpolations polynomiales avec le paramètre de tension le plus faible sont celles qui s'éloignent le plus de la polyligne originale.

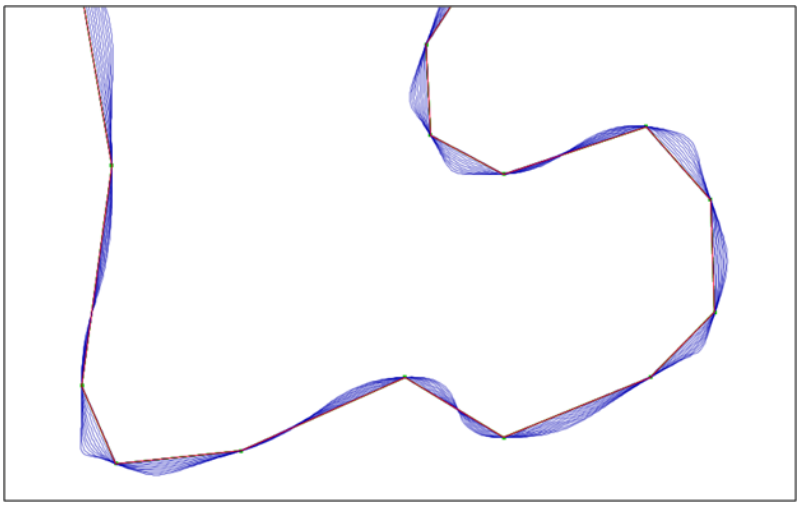


Figure 3.52 – Résultats des reconstructions de courbes avec un paramètre de tension de 0,1 à 0,9

La sélection de la valeur de tension la plus réaliste pour reconstruire les courbes par le biais d'interpolations polynomiales est réalisée par validation visuelle. Les valeurs de tension les plus faibles (inférieures à 0,5) présentent clairement des reconstructions de courbes irréalistes. Les valeurs de tension supérieures à 0,8 génèrent quant à elle des courbes trop proches de la ligne polygonale initiale.

Ainsi, après élimination, notre choix s'est porté sur une valeur de tension de 0,7, qui présente le meilleur compromis pour la reconstruction de courbes réalistes, tout en évitant les exagérations locales (qui peuvent encore être détectées à une valeur de 0,6). La figure 3.53 présente le résultat de la reconstruction avec une valeur de tension de 0,7, qui offre un résultat visuellement satisfaisant.

Figure 3.53 – Résultat de reconstruction de courbe avec un paramètre de tension de 0,7

Ceci nous permet donc, dans le cadre de cette expérimentation, d'estimer l'impact de l'approximation polygonale sur cette extraction du réseau routier de la BDCARTO, par une sous-estimation de la longueur totale de 843 mètres, soit 0,44 % de la longueur de la polyligne initiale.

Cette expérimentation réalisée sur une extraction du réseau routier de la BDCARTO en zone de montagne nous a donc permis de valider l'intérêt du suréchantillonnage pour la reconstruction de géométries courbes par interpolation polynomiale. Egalement, cette expérimentation a permis de fixer le paramètre de tension nécessaire à la méthode des Cardinal Splines à une valeur de 0,7. Ceci permet de reconstruire des courbes visuellement satisfaisantes pour estimer l'impact de l'approximation polygonale sur les mesures de longueur.

Synthèse Section L'application de la méthode de reconstruction des courbes à partir des polylignes du réseau routier de la BDCARTO a permis de mettre en valeur l'importance du suréchantillonnage sur le réalisme des courbes reconstruites. Cette expérimentation a également permis de fixer le paramètre de tension, nécessaire à la méthode des Cardinal splines à 0,7. Celui-ci offre les résultats les plus visuellement satisfaisants, et sera donc utilisé par la suite afin d'estimer l'impact de l'approximation polygonale sur les mesures de longueur.

Conclusion de la section

Dans le cadre de l'estimation de l'impact des règles de représentation sur les mesures géométriques, cette section a permis de proposer des méthodes permettant d'estimer l'impact de l'approximation polygonale des courbes sur les mesures de longueur.

Nous avons tout d'abord montré que les bases de données géographiques sont généralement impactées par l'approximation polygonale, et que le degré de celle-ci varie en fonction du niveau de détail géométrique recherché. Nous avons notamment souligné que lorsque des entités courbes du monde réel sont représentées par une ligne polygonale dans une base de données géographiques, les mesures géométriques de longueur sont systématiquement sous-estimées, en partant du principe que les points composant la polyligne sont correctement localisés. Pour estimer l'impact de l'approximation polygonale sur les mesures de longueur, nous avons donc proposé une méthode basée sur la reconstruction de géométries courbes réalistes, par le biais de méthodes d'interpolation polynomiale.

La méthode d'interpolation par splines cubiques d'Hermite a été retenue pour opérer à la reconstruction de géométries courbes. Le paramétrage des tangentes utilise la méthode des Cardinal Splines, qui ne nécessitent de définir qu'un paramètre de tension. Suite aux expérimentations réalisées sur des tronçons de route, il s'est avéré que la valeur de tension proposant le meilleur compromis est de 0,7.

Le problème de reconstruction de courbes irréalistes a également été exposé. Pour le résoudre, un suréchantillonnage de la géométrie initiale, basé sur l'exploration des distances inter-points des objets évalués a été proposé. Egalement, étant donné que tous les objets d'une base de données routière ne représentent pas des entités courbes du monde réel, une méthode de détection des objets sinueux a également été proposée, afin de limiter la reconstruction des courbes à une sélection d'objets pertinents. Cette méthode demeure cependant optionnelle dans le cadre de la mise en œuvre du modèle d'estimation, l'utilisateur conservant la possibilité de sélectionner manuellement les objets à évaluer.

La méthode de reconstruction des géométries courbes proposée dans cette section part du principe que l'ensemble des points composant la géométrie des objets vectoriels sont bien localisés. Cependant, nous savons que la localisation de ces points est soumise à une imprécision qui dépend principalement des processus de production mis en œuvre.

Conclusion du chapitre

Ce chapitre nous a permis de présenter les méthodes élaborées afin d'estimer l'impact des règles de représentation des données géographiques sur les mesure géométriques. Trois règles de représentation ont été traitées : la projection cartographique, la non-prise en compte du terrain, et l'approximation polygonale des courbes. Des méthodes d'enrichissement géométrique ont été proposées afin d'estimer ces impacts.

Nous avons montré que l'estimation de l'impact des règles de représentation implique un processus d'acquisition des connaissances sur les données évaluées (cf. figure 3.54). Ces connaissances peuvent être des données externes (grille d'altération linéaire, MNT) ou des informations extraites des données évaluées (qualification de la sinuosité).

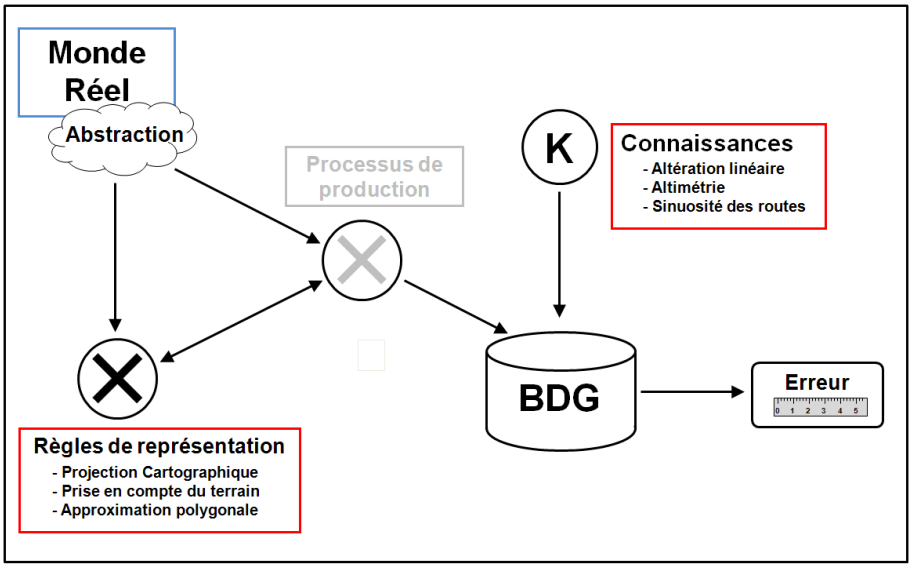


Figure 3.54 – Modélisation de l'impact des règles de représentation au sein du processus global de génération d'erreur de mesure

Pour les règles de représentation traitées au cours de ce chapitre, les impacts estimés se matérialisent sous forme d'erreur de mesure. Nous avons malgré tout bien conscience que cette erreur présente une imprécision, en particulier du fait de l'imprécision des données sources, comme c'est par exemple le cas des MNT. Cependant, la quantification formelle de cette imprécision, et ses conséquences sur les impacts estimés, n'a pu être effectuée dans le cadre de ce travail de thèse.

Après les règles de représentation, le chapitre suivant va nous permettre de proposer des méthodes permettant d'estimer l'impact des processus de production sur les mesures géométriques de longueur et de surface.

Chapitre 4

Méthodes d'estimation de l'impact des processus de production

Introduction

Le chapitre précédent nous a permis d'élaborer des méthodes permettant d'estimer l'impact des règles de représentation sur les mesures géométriques de longueur et de surface. Ces règles de représentation sont conditionnées par le processus d'abstraction du monde réel, qui spécifie comment sont modélisés géométriquement les objets d'une base de données. Les impacts que nous avons pu modéliser se manifestent, pour chacun d'entre eux, par une erreur de mesure.
Dans ce chapitre, nous allons proposer des méthodes permettant d'estimer l'impact du second processus générateur d'erreurs de mesures, à savoir le processus de production des objets géographiques d'une base de données. Nous allons voir que son impact se manifeste par une imprécision de mesure.
Nous savons que dans la pratique les impacts de ces deux processus s'accumulent. Cependant, afin d'évaluer séparément l'impact de chacune des composantes de l'erreur finale de mesure, les erreurs impliquées par les règles de représentation ne seront pas prises en compte dans cette section. Nous ne proposerons ici que des méthodes permettant d'estimer l'impact des processus de production, indépendamment des effets impliqués par leur représentation. Les méthodes de combinaison de ces différentes sources d'erreurs seront traitées en section 5.1.

Deux processus de production dont nous cherchons à modéliser l'impact sont étudiés dans ce chapitre :
– l'erreur de pointé,
– la généralisation cartographique.
Pour différentes raisons, l'étude de l'impact de l'imprécision des acquisitions GNSS n'a pas été intégré dans ce chapitre, mais quelques pistes permettant sa modélisation seront tout de même proposées. Nous verrons que deux approches sont utilisées pour modéliser l'impact de ces processus de production sur les mesures géométriques. Une approche par simulations sera tout d'abord proposée afin de modéliser l'impact de l'erreur de pointé de l'opérateur, tandis qu'une approche par tests, basée sur des comparaisons avec des données de référence, sera proposée afin de modéliser l'impact de la généralisation cartographique.

4.1 Erreur de pointé

Comme l'ont proposé de nombreuses contributions, l'imprécision de localisation de la géométrie des objets vectoriels peut être modélisée par le biais de simulations (Vauglin, 1997), basées sur des méthodes dites de Monte-Carlo, qui utilisent des procédés aléatoires. Ce type d'approche peut se révéler adapté pour modéliser les impacts de l'erreur de pointé de l'opérateur (Keefer *et al.*, 1988; Hunter et Goodchild, 1996) ou encore de l'imprécision des acquisitions GNSS (De Bruin *et al.*, 2008) sur le positionnement des objets géographiques. En exploitant les mesures de longueur ou de surface calculées à partir des réalisations d'une simulation, cette approche permet également d'estimer l'imprécision des mesures géométriques (Goodchild *et al.*, 1999). Dans cette logique, nous proposons dans cette section d'élaborer un modèle de simulation de l'erreur de pointé, permettant d'estimer son impact sur les mesures géométriques de longueur et de surface.
Nous nous focaliserons dans cette section uniquement sur la modélisation de l'impact de l'erreur de pointé de l'opérateur. En effet, bien que les approches par simulations puissent s'avérer pertinentes pour évaluer l'impact de l'imprécision des acquisitions GNSS, la difficulté de proposer un modèle reproduisant l'ensemble des sources d'erreurs (environnementales, atmosphériques, satellitaires) ne nous permet pas d'intégrer ce processus de production de manière suffisamment complète et rigoureuse dans notre processus d'évaluation.

Nous présenterons dans un premier temps les principes généraux relatifs à l'approche par simulations, basée sur un bruitage de la géométrie des objets. Nous proposerons ensuite des méthodes permettant de paramétrer l'amplitude du bruit utilisé pour reproduire l'erreur de pointé, tout en préservant le réalisme des objets simulés. Une expérimentation sera ensuite réalisée afin de calibrer le modèle, et nous terminerons par une mise en application du modèle proposé.

4.1.1 Le bruitage de la géométrie des objets vectoriels

Pour mieux comprendre le fonctionnement et les enjeux des approches par simulations, le principe général du bruitage de la géométrie des objets vectoriels sera tout d'abord exposé. Nous présenterons ensuite les problèmes classiques impliqués par ce type d'approche, auxquels nous chercherons à remédier afin de mettre en œuvre des simulations réalistes.

4.1.1.1 Principe général

L'approche par simulations part du principe que tout point composant la géométrie d'un objet vectoriel possède une imprécision relative à sa localisation. On considère généralement que cette imprécision de localisation peut être modélisée par une loi normale indépendante en x et en y, comme l'ont souligné plusieurs contributions (Bolstad *et al.*, 1990; Hunter et Goodchild, 1996). De nombreux tests réalisés sur des bases de données géographiques d'objets ponctuels tendent à confirmer ce constat (Vauglin, 1997).

La figure 4.1 illustre l'imprécision de localisation d'un point par un disque gaussien, où la probabilité de localisation du point est plus importante en son centre qu'aux extrémités. On part ici de l'hypothèse que les erreurs en chaque point sont indépendantes de celles des points voisins, supposant donc qu'il n'y a pas de corrélation d'erreur entre points successifs.

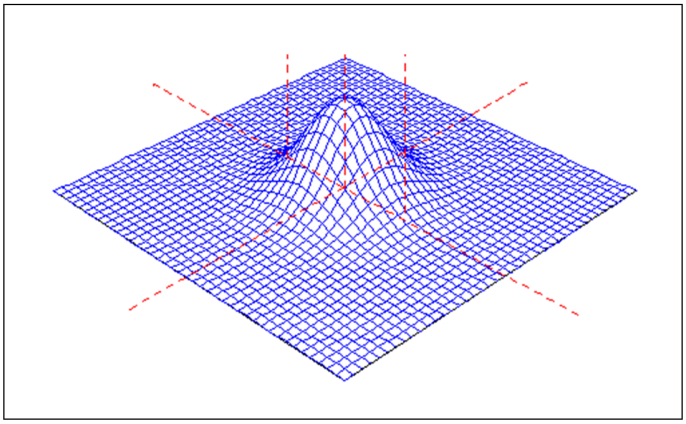


Figure 4.1 – Imprécision de localisation d'un point selon une probabilité suivant une loi normale en x et en y

Dans le cas d'objets linéaires, Vauglin (1997) propose de modéliser l'imprécision de localisation des points des géométries linéaires par une loi GES, c'est-à-dire un mélange de loi Gaussienne (permettant de modéliser l'imprécision des erreurs de pointé) et d'une loi de Laplace (permettant de modéliser les effets de l'interpolation entre points). Cependant, comme nous ne nous intéressons dans cette section qu'à l'erreur de pointé, nous n'utiliserons pas la loi GES. De plus, comme le souligne Bonin (2002), l'utilisation de la loi GES pour simuler des erreurs de localisation conduit très vite à des problèmes topologiques, qui ne sont pas réalistes.

Ainsi, à partir des hypothèses relatives à l'imprécision de localisation des points composant la géométrie des objets vectoriels, l'approche par simulation consiste à générer un grand nombre de réalisations d'objets vectoriels "bruités". Le bruit est ici entendu comme une translation des coordonnées x et y de chaque point de la géométrie simulée, en fonction des paramètres de la loi utilisée pour modéliser l'imprécision de localisation. Nous utilisons ici la loi normale, mais d'autres lois de probabilité peuvent être appliquées.

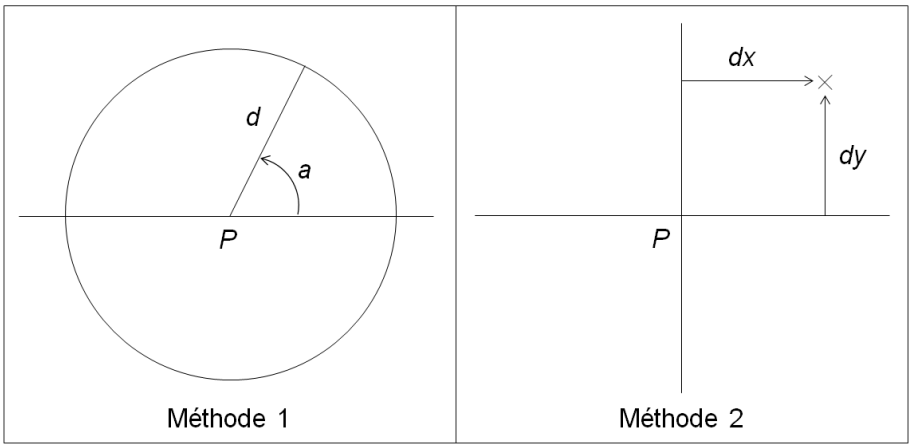


Figure 4.2 – Principe de déplacement d'un point selon deux méthodes

Méthode de bruitage Deux méthodes peuvent être utilisées pour bruiter les coordonnées des points d'une géométrie, comme illustré dans la figure 4.2 :

- dans la première méthode, les coordonnées d'un point P sont translatées selon un angle a, déterminé aléatoirement, à une distance d de sa position initiale (d étant tiré selon la loi utilisée, ici une loi normale),
- dans la seconde méthode, les coordonnées x et y d'un point P sont respectivement translatées d'une distance dx et dy, ces dernières étant tirées indépendamment selon la loi utilisée (une loi normale ici également).

L'opération de bruitage est répétée sur l'ensemble des points composant la géométrie d'un objet vectoriel. En renouvelant cette opération un grand nombre de fois, il est possible d'étudier la sensibilité de la mesure de longueur ou de surface des objets bruités, en la comparant avec la mesure initiale (cf. figure 4.3).

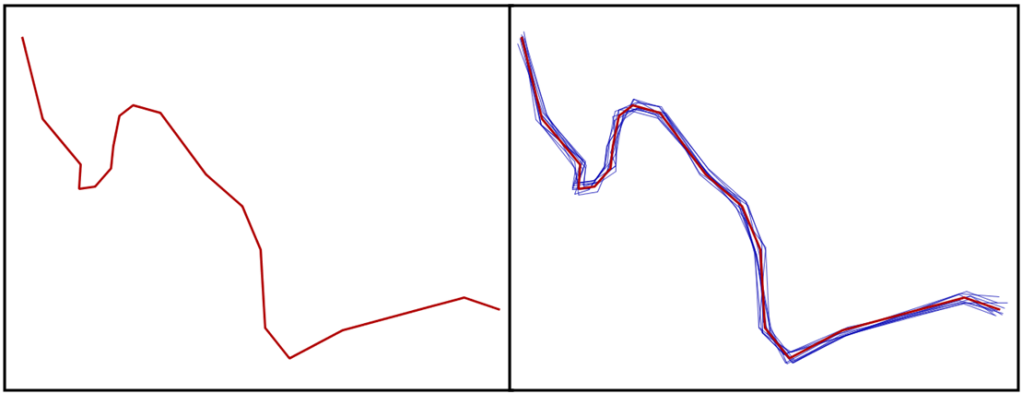


Figure 4.3 – Une polyligne représentant un tronçon de route dans la BDCARTO, et 10 réalisations de son bruitage

L'imprécision des mesures de longueur ou de surface calculées à partir des réalisations de la simulation varie en fonction du paramétrage de la loi utilisée (ici une loi normale). Ce paramétrage conditionne l'amplitude du bruit appliqué aux géométries simulées, et par conséquent la sensibilité des mesures de longueur ou de surface calculées en aval (cf. figure 4.4). Le paramétrage du bruit constitue donc un point fondamental de cette approche, puisque l'imprécision de la mesure résultante en dépendra directement.

Figure 4.4 – Effet du paramétrage de la fonction gaussienne sur le bruitage des sommets d'une polyligne

Nous venons de voir que le réalisme du bruitage des objets vectoriels dépend fortement du paramétrage du bruit exercé sur les points composant leur géométrie. Nous avons pris parti de modéliser le bruit appliqué sur chaque point à l'aide d'une probabilité d'imprécision de localisation suivant une loi normale. Ce choix ne convient donc que pour modéliser des erreurs qui suivent les mêmes hypothèses. Nous allons chercher par la suite à savoir si ce choix de modélisation est adapté pour estimer l'impact de l'erreur de pointé sur les mesures géométriques. Nous chercherons également à savoir comment paramétrer le bruit de façon réaliste. Précédemment à cela, nous allons exposer les problèmes classiques auxquels se heurtent les approches par simulations, puisqu'il nous sera nécessaire de les prendre en compte afin de préserver le réalisme des objets simulés.

4.1.1.2 Problèmes impliqués par le bruitage

Différents problèmes peuvent perturber le réalisme du bruitage d'une géométrie vectorielle, et ainsi en dégrader la pertinence afin d'étudier l'impact sur les mesures géométriques. Nous en présenterons deux : les problèmes d'ordres topologiques, et les problèmes de respect de forme.

Les problèmes topologiques

Comme l'a souligné Bonin (2002), les approches par simulations peuvent très vite conduire à des problèmes topologiques, lorsque le bruit exercé sur les points composant la géométrie de l'objet vectoriel devient important. Par exemple, après bruitage, une polyligne peut s'intersecter elle-même, ce qui n'est pas cohérent avec les propriétés de l'objet (cf. figure 4.5).

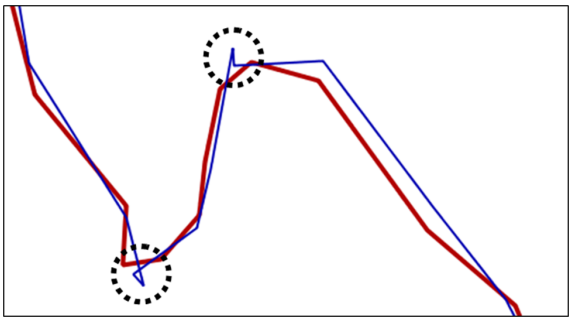


Figure 4.5 – Incohérences topologiques impliquées par un bruitage trop important

Ces incohérences topologiques ne sont pas réalistes. Elles sont généralement évitées lors de la saisie des objets par l'opérateur. Il convient donc d'élaborer des méthodes permettant de les éviter. Ainsi, Bonin (2002) et Fouqué (1999) proposent d'introduire des corrélations entre les erreurs simulées sur les points successifs d'une même polyligne. L'approche proposée par les auteurs est basée sur les variogrammes. Elle est illustrée ici par l'exemple du bruitage d'un réseau routier et se divise en plusieurs étapes :

- les nœuds du réseau (les carrefours) sont bruités indépendamment selon une loi normale $N(0,\sigma)$,
- les points intermédiaires des polylignes composant le réseau sont ensuite bruités selon une loi normale $N(0,\sigma')$, où σ' est calculé à l'aide d'un variogramme en fonction de sa distance par rapport aux extrémités de la polyligne (qui prennent les valeurs 0 et 1), c'est-à-dire les nœuds du réseau déjà bruités. (cf. figure 4.6).

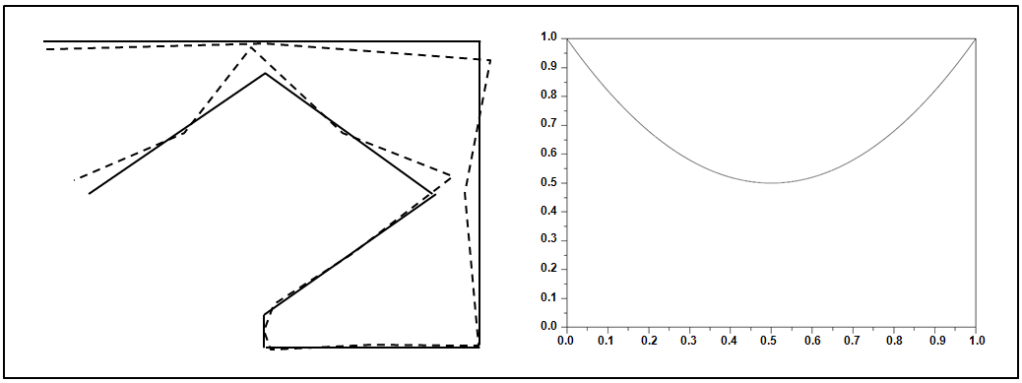


Figure 4.6 – Bruitage d'une polyligne avec corrélations (en tiretés), suivant un variogramme (à droite) déterminant la valeur de σ', d'après Bonin (2002) et Fouqué (1999)

Suite à des expérimentations, Bonin (2002) et Fouqué (1999) estiment que cette approche permet de réduire sensiblement les problèmes topologiques qui apparaissent lors de simulations sans corrélations. Cependant cette approche s'avère critiquable sur certains aspects, puisqu'afin de résoudre ces problèmes topologiques, des points intermédiaires sont intégrés au milieu de chaque segment de polyligne (Fouqué, 1999).

D'autres auteurs ont proposé des solutions afin de remédier aux problèmes topologiques impliqués par le bruitage des géométries, notamment Hunter et Goodchild (1996) qui proposent une méthode basée sur des grilles afin de remédier aux problèmes de "fracture", c'est-à-dire lorsqu'une polyligne s'intersecte elle-même. Les auteurs proposent également de densifier les géométries de certaines classes d'objets pour lesquelles cela s'avère pertinent, telles que les forêts par exemple, qui ont des géométries plus irrégulières.

Les problèmes de préservation de forme

Outre les problèmes d'incohérences topologiques évoqués précédemment, le bruitage de la géométrie d'un objet vectoriel doit être réalisé de manière réaliste pour préserver la forme générale des objets saisis.

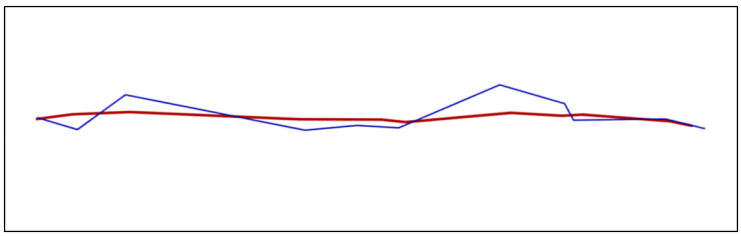


Figure 4.7 – Une polyligne au caractère rectiligne (en rouge) et une simulation (en bleu)

Par exemple, lorsqu'une polyligne composée de plusieurs points présente un caractère rectiligne, cela signifie que c'est a priori le cas dans la réalité. On peut donc supposer que l'alignement des points saisis par l'opérateur suppose la représentation d'une réalité tangible. Or, comme illustré dans la figure 4.7, la réalisation d'un bruitage strictement aléatoire sur des points (relativement) alignés peut avoir pour conséquence de faire perdre le caractère rectiligne de la polyligne.

Cet exemple de simulation n'est pas réaliste, puisqu'il ne préserve pas la forme initiale de l'objet saisi par l'opérateur. Mais il affecte également de manière irréaliste la mesure géométrique de longueur calculée à partir de l'objet simulé. En effet, l'alignement des points composant la géométrie de la polyligne initiale minimise sa longueur. Les effets du bruitage de la polyligne génèrent dans ce cas de figure une perte de son caractère rectiligne, et se traduisent par une augmentation systématique de sa longueur. Cet exemple, illustré ici par une simple polyligne, peut prendre davantage d'ampleur lorsqu'il est généralisé à un ensemble de polylignes, comme pour la simulation d'un réseau routier.

Les différents problèmes que nous venons d'évoquer mettent simplement en valeur l'idée qu'il n'est pas pertinent de modéliser par une approche purement aléatoire, des processus humains de restitution qui font intervenir des mécanismes cognitifs beaucoup plus complexes à modéliser. Le respect de la forme générale d'un objet l'illustre bien, et nous devons le prendre en compte afin de proposer une méthode réaliste de simulation, qui permette d'estimer fidèlement l'impact de l'erreur de pointé sur les mesures géométriques de longueur et de surface. Ainsi, nous présenterons dans les sections suivantes les méthodes mises en œuvre afin de paramétrer le bruit appliqué en chaque point, et celles permettant de préserver le réalisme des objets simulés.

Synthèse Section Nous avons montré dans cette section que pour estimer l'impact de l'erreur de pointé sur les mesures géométriques de longueur et de surface, des approches par simulations peuvent être employées. Après avoir présenté le principe général des approches par simulations, qui reposent sur un bruitage de la géométrie des objets, nous avons souligné les principaux problèmes, d'ordres topologique ou de préservation de forme, auxquels elles peuvent être confrontées.

4.1.2 Paramétrage de l'amplitude du bruit

Cette section a pour objectif de définir les méthodes permettant de paramétrer l'amplitude du bruit appliqué en chaque point d'une géométrie simulée. Comme nous l'avons souligné, la définition de l'amplitude du bruit est déterminante dans le cadre des approches par simulations. Nous présenterons donc dans un premier temps les méthodes permettant d'estimer l'imprécision de pointé de l'opérateur. Nous verrons que l'imprécision de pointé dépend de l'échelle de saisie des objets. Nous proposerons donc dans un deuxième temps une méthode permettant d'estimer cette échelle.

4.1.2.1 Estimation de l'imprécision de pointé

Le paramétrage de l'amplitude du bruit conditionne directement le résultat des simulations, et par conséquent l'estimation de l'imprécision des mesures géométriques. Nous avons évoqué précédemment qu'un bruit trop important peut générer des incohérences topologiques. La question du paramétrage du bruit appliqué sur les géométries simulées s'avère donc fondamentale pour assurer une estimation réaliste de l'imprécision des mesures géométriques. Nous allons présenter dans cette section comment paramétrer l'amplitude de ce bruit, à travers l'estimation de la précision de pointé.

Estimation de la précision de pointé

L'erreur de pointé est définie comme une erreur accidentelle, impliquée par l'opérateur lors de la saisie de chaque point composant la géométrie d'un objet vectoriel. On sait que cette erreur au caractère aléatoire suit un comportement gaussien et peut être modélisée par une loi normale $N(0,\sigma)$. La question relative au paramétrage de cette loi concerne donc l'estimation de la valeur de l'écart-type σ, qui définit la **précision de pointé**. Afin de paramétrer la précision de pointé, nous nous sommes inspirés des observations relevées en restitution photogrammétrique.

Précision planimétrique Chandelier (2011) estime la précision de restitution planimétrique σ_p en fonction de l'échelle moyenne $1/E$ des prises de vues et d'une valeur σ_c, conformément à l'équation 4.1.

$$\sigma_p^2 = \sigma_x^2 + \sigma_y^2 = E * \sqrt{2\sigma_c} \tag{4.1}$$

Dans des conditions standards, avec un appareil de restitution de qualité normale, la valeur de σ_c est estimée à 20 μm, ce qui signifie que $\sigma_p^2 = E * 0,2\text{mm}$.

Cette estimation de la précision de pointé planimétrique est également mentionnée par certains producteurs de données. Par exemple, dans le descriptif technique du produit Histolit-v2.0 (IGN-SHOM, 1992), le produit TC25000 (c'est à dire le trait de côte produit au 1 :25.000) est décrit avec "*une précision de pointé estimée meilleure que 0,2 mm*" et une "*précision géométrique estimée à 0,3 mm, correspondant à une précision terrain de 12m.*" Cet exemple nous montre bien que l'estimation de la précision de pointé planimétrique doit être formalisée en prenant en compte l'échelle de saisie des objets géographiques.

Précision altimétrique En ce qui concerne l'estimation de la précision de restitution altimétrique σ_z, Chandelier (2011) l'exprime en fonction de l'éloignement des sommets des prises de vue B, de la distance à l'objet H, d'une valeur σ_a (équivalente à 15 μm en conditions standards) et de l'échelle moyenne des prises de vues $1/E$, conformément à l'équation 4.2.

$$\sigma_z = \sigma_a * (\frac{H}{B}) * E \tag{4.2}$$

Dans le cadre de cette étude relative à l'impact de l'erreur de pointé sur les mesures géométriques, nous ne nous focaliserons que sur l'impact de l'imprécision planimétrique. En effet, comme l'exprime l'équation 4.2, il est nécessaire de connaître les valeurs de B et de H, afin d'étudier l'impact de la précision altimétrique. De plus, comme ceci a été précisé auparavant, de nombreuses bases de données géographiques ne sont pas saisies en coordonnées tridimensionnelles. C'est pourquoi nous ne prendrons pas en compte l'impact de la précision altimétrique sur les mesures géométriques dans ce travail.

Paramétrage du bruit

Pour paramétrer la loi normale $N(0,\sigma)$ utilisée pour le bruitage de la géométrie des objets vectoriels, nous retiendrons comme écart-type la valeur σ_p, comme proposé par Chandelier (2011). Cette estimation de l'écart-type σ_p implique de connaître l'échelle moyenne de saisie E de la base de données, qui demeure la principale inconnue du paramétrage.

Dans le cadre de la mise en œuvre du modèle d'évaluation, deux solutions peuvent être envisagées : soit l'utilisateur connait l'échelle de saisie, et est donc capable de la paramétrer, soit il ne la connait pas, et dans ce cas des méthodes d'extraction des connaissances doivent être mises en œuvre afin de l'estimer. Nous proposons dans la section suivante une méthode permettant d'estimer l'échelle de saisie moyenne des objets évalués, dans le cas où l'utilisateur ne la connait pas.

4.1.2.2 Estimation de l'échelle de saisie

La connaissance de l'échelle moyenne de saisie est un préalable nécessaire au paramétrage du modèle d'estimation de l'impact de l'erreur de pointé. Avant de présenter la méthode permettant de l'estimer, quelques définitions relatives à la notion d'échelle dans les bases de données seront proposées afin d'éclaircir notre propos.

La notion d'échelle dans les bases de données géographiques

La notion d'échelle mathématique est a priori une notion assez simple. Il s'agit du rapport entre une distance mesurée sur une carte et sa valeur sur le terrain (Ruas, 2002). Cependant, si une carte possède une échelle fixe, ce n'est pas le cas d'une base de données.

Echelle de représentation En effet, Ruas et Mustière (2005) s'accordent à dire que les bases de données n'ont pas d'échelle, mais plutôt qu'elles sont faites pour être utilisées dans une **gamme d'échelles**.

Par exemple, la BDTOPO peut être utilisée dans une gamme d'échelles allant du 1 :5.000 au 1 :50.000 (IGN, 2011c), tandis que la BDCARTO peut être utilisée dans une gamme d'échelle allant du 1 :50.000 au 1 :250.000 (IGN, 2011b). On parle donc ici de la **gamme d'échelles de représentation**. Cependant, on conçoit généralement que l'**échelle caractéristique de représentation**, c'est-à-dire l'échelle la plus appropriée pour représenter les données, est environ au 1 :10.000 pour la BDTOPO et au 1 :50.000 pour la BDCARTO. La notion de gamme d'échelle ne se limite pas à l'échelle de représentation, elle peut également être appliquée à l'échelle de saisie des objets dans la base de données.

Echelle de saisie Les bases de données géographiques sont produites à des niveaux de détail qui leur sont propres. Cela suppose la mobilisation de processus de production permettant de représenter les phénomènes du monde réel au niveau de détail souhaité, et donc dans la gamme d'échelle de représentation désirée. Cependant, des processus de production hétérogènes sont utilisés afin de saisir la géométrie des objets de la base, ce qui nous amène à penser que l'échelle de saisie n'est pas fixe. Par exemple, les objets de la BDTOPO sont principalement saisis par le biais de restitution photogrammétrique, mais également de relevés GPS de précision. De plus, un opérateur de restitution photogrammétrique change régulièrement l'échelle de visualisation afin de saisir avec davantage de précision certains détails de la géométrie des objets.

Nous voyons donc bien qu'il n'est pas possible de considérer une seule et unique échelle de saisie à une base de données, mais davantage une **gamme d'échelles de saisie**. Par exemple, l'échelle de saisie des objets de la BDTOPO va du 1 :200 au 1 :5.000 et son échelle de représentation du 1 :5.000 au 1 :50.000 . On peut en effet considérer que l'échelle inférieure de saisie correspond à l'échelle supérieure de représentation (cf. figure 4.8).

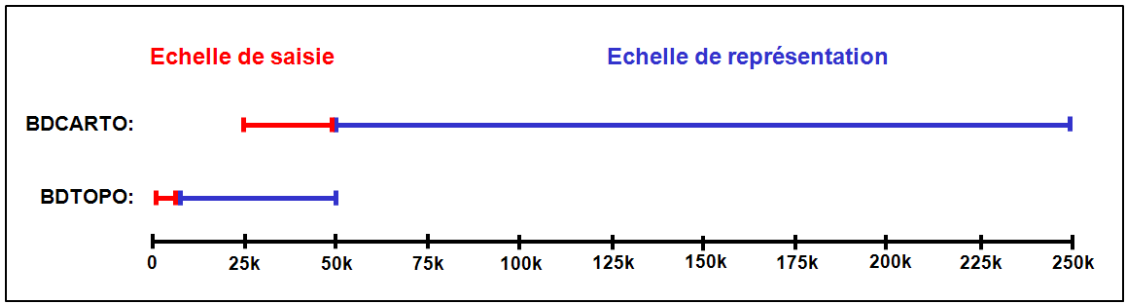


Figure 4.8 – Gammes d'échelles de saisie et de représentation de la BDTOPO et de la BDCARTO

Dans le cadre de ce travail, nous cherchons à estimer l'échelle de saisie moyenne. Par échelle de saisie moyenne nous entendons l'échelle moyenne à laquelle sont affichées les données sources utilisées pour produire la géométrie des objets. On peut parler d'échelle caractéristique de saisie, même si nous savons que dans la réalité, celle-ci n'existe pas.
Par conséquent, il est difficile pour un utilisateur de connaître cette échelle de saisie caractéristique. On sait juste a priori que celle-ci est supérieure à l'échelle limite supérieure de représentation de la base de données, puisqu'il ne serait pas correct d'utiliser une base de données à une échelle supérieure à son échelle de saisie.

Pour estimer l'échelle de saisie caractéristique, nous proposons d'exploiter les distances inter-points des objets de la base de données évaluée. Nous savons en effet qu'une base de données géographiques est saisie dans une gamme d'échelle correspondant à un niveau de détail désiré pour son utilisation (pour une gamme d'échelle de représentation correspondante). Etant donné qu'il existe un lien entre niveau de détail et granularité des objets saisis, l'exploration des distances inter-points peut offrir une source d'information intéressante.

Estimation de l'échelle de saisie moyenne par exploration des distances inter-points

Nous savons, par définition, que les plus petites distances entre points consécutifs permettent de définir la granularité d'une base de données. Or, il existe une relation entre l'échelle de saisie, le niveau de détail recherché, et la granularité des objets saisis. On peut en effet considérer que plus l'échelle de saisie des objets est importante, plus la distance entre points successifs d'une géométrie est faible, ce qui permet la restitution de nombreux détails. Ainsi, nous pensons que l'analyse des distances entre points consécutifs d'une géométrie peut nous permettre d'estimer l'échelle de saisie caractéristique d'une base de données.

Exploration des distances inter-points La granularité décrit la taille des plus petites formes géométriques présentes dans la base de données. Cependant, nous n'allons pas étudier ici les plus petites distances, mais l'ensemble des distances inter-points d'une base de données afin d'estimer son échelle de saisie. Nous savons en effet que les distances entre points successifs dans une base de données ne sont pas constantes. Par exemple, la distance entre points successifs d'une polyligne représentant une route, sera plus faible lorsque la sinuosité augmente, que lorsque celle-ci baisse.

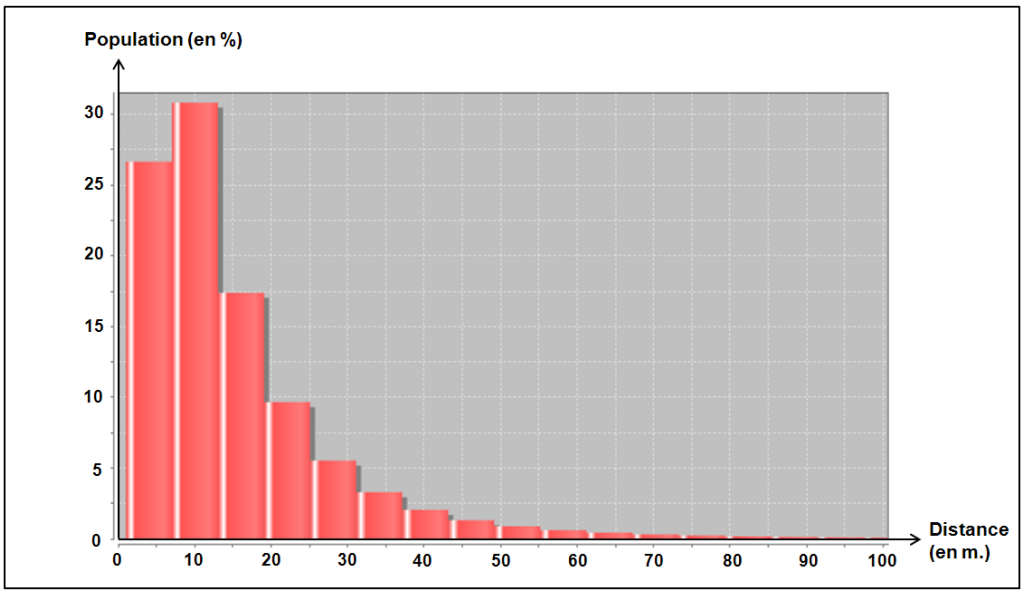


Figure 4.9 – Distribution des distances inter-points à partir d'une extraction des tronçons de route de la BDTOPO

Le graphique de la figure 4.9 illustre la distribution des distances inter-points sur l'ensemble du réseau routier de la BDTOPO du département des Pyrénées-Atlantiques. Cette distribution montre que les distances entre points successifs d'une géométrie sont hétérogènes. Dans cet exemple, la grande majorité des distances est inférieure à 20 mètres, mais certaines d'entre elles peuvent atteindre plusieurs centaines de mètres. La moyenne et la médiane des distances inter-points sont respectivement de 15,60 mètres et de 11,25 mètres.

Pour estimer l'échelle de saisie moyenne, nous proposons, à la vue de cette distribution, d'exploiter la distance inter-points médiane, étant donné que cet indicateur, au contraire de la moyenne, n'est pas perturbé par les valeurs extrêmes. En partant du principe qu'il existe une relation entre l'échelle de saisie et la distance inter-points médiane, nous proposons d'étudier cette relation sur des classes d'objets extraites de plusieurs bases de données.

Expérimentation Pour étudier la relation entre la distance inter-points médiane et l'échelle de saisie, trois classes d'objets de la BDTOPO et de la BDCARTO ont été exploitées, à savoir le réseau routier, le réseau hydrographique et les limites communales.

	Réseau routier	Réseau hydrographique	Limites communales
BDTOPO	11,26 m.	9,4 m.	12,03 m.
BDCARTO	55,32 m.	52,4 m.	65 m.

Table 4.1 – Distances inter-points médianes de trois classes d'objets de la BDTOPO et la BDCARTO

Résultats Le tableau 4.1 présente les distances inter-points médianes calculées sur les trois classes d'objets de la BDTOPO et de la BDCARTO. Ces résultats montrent qu'il existe une relation étroite entre l'échelle de saisie et la distance inter-points médiane. En effet, la distance inter-points médiane, toutes classes confondues, est en moyenne de 10,9 mètres pour la BDTOPO et de 57,7 mètres pour la BDCARTO.

Paramétrage du modèle En partant du principe que l'échelle caractéristique de représentation de la BDTOPO est au 1 :10.000 et celle de la BDCARTO au 1 :50.000, nous constatons un rapport d'environ 1 à 1.000 entre distances inter-points médianes et échelle caractéristique de représentation. Cependant, dans le cadre de l'élaboration du modèle d'estimation, nous cherchons à estimer l'échelle de saisie moyenne, et non-pas l'échelle de représentation. Or, comme nous l'avons souligné, il est délicat de formaliser une estimation de l'échelle de saisie moyenne, puisque dans la pratique, celle-ci n'existe pas.
Afin de proposer une estimation de l'échelle de saisie moyenne, et sachant que celle-ci est par définition supérieure à l'échelle de représentation, nous proposons de l'approximer à l'aide de l'équation 4.3. Dans cette équation, $1/E_S$ est l'échelle de saisie moyenne et D_M est la distance inter-points médiane.

$$E_S = \frac{D_M * 1000}{2} \tag{4.3}$$

En appliquant cette méthode aux objets expérimentés précédemment, l'échelle de saisie moyenne est estimée au 1 :5.450 pour la BDTOPO et au 1 :28.850 pour la BDCARTO.

Bien que cette estimation de l'échelle moyenne de saisie soit approximative, elle permet de fournir une information intéressante à l'utilisateur dans le cadre de la mise en œuvre du modèle de simulation de l'erreur de pointé. L'utilisateur demeure néanmoins libre d'exploiter directement cette estimation de l'échelle de saisie (ou bien de l'arrondir) afin de paramétrer le modèle de simulation.

La méthode de paramétrage de l'amplitude du bruit utilisée dans notre approche par simulations étant à présent définie, nous allons proposer dans la section suivante des méthodes permettant de simuler l'erreur de pointé de manière réaliste.

Synthèse Section Nous avons proposé dans cette section une méthode permettant de paramétrer l'amplitude du bruit appliqué à la géométrie des objets simulés, afin de modéliser l'impact de l'erreur de pointé. Le paramétrage du bruit nécessitant de connaître l'échelle de saisie moyenne des objets, une méthode d'estimation de cette échelle, basée sur l'exploitation de la distance inter-points médiane, est proposée.

4.1.3 Contraintes appliquées pour préserver le réalisme du bruitage

Comme nous l'avons évoqué précédemment, l'application stricte d'une loi normale indépendante en x et en y, sur chacun des sommets d'une géométrie n'est pas une solution réaliste. Cette solution a pour conséquences d'engendrer des problèmes topologiques et de non-préservation de la forme initiale des objets, que la saisie par un opérateur aurait permis d'éviter. Pour reproduire de façon réaliste l'impact de l'erreur de pointé sur les mesures géométriques, le modèle de simulation doit satisfaire différents types de contraintes. Deux types de contraintes seront proposées : les contraintes de respect de la forme initiale des objets, et les contraintes en termes de corrélation des erreurs entre points successifs.

4.1.3.1 Respect de la forme initiale des objets

L'application d'un bruit aléatoire peut avoir pour conséquence de ne pas préserver la forme initiale de l'objet, comme nous l'avons montré avec l'exemple d'une polyligne au caractère rectiligne (cf. figure 4.7). Au delà de l'irréalisme de la simulation, nous avons vu que cette modélisation affecte également les mesures géométriques calculées à partir de l'objet simulé. Pour simuler de manière réaliste l'erreur de pointé de l'opérateur, des contraintes relatives au respect de la forme initiale des objets peuvent être appliquées sur le bruit simulé. Nous proposons pour cela de pondérer le bruit, en prenant en compte l'angularité inscrite entre points successifs d'une géométrie ainsi que le degré des nœuds situés aux extrémités.

Pondération par l'angularité entre points successifs

Hypothèse Nous pouvons considérer que l'angle inscrit entre sommets successifs d'un objet vectoriel implique une représentation de la réalité, déterminée consciemment par l'opérateur de saisie. En effet, on peut supposer que si les sommets successifs d'un objet sont alignés, un bruitage réaliste doit préserver cet alignement. Au contraire, lorsque des changements de direction entre sommets successifs sont détectés, le risque d'erreur de pointé est plus important. Cette hypothèse nous amène donc naturellement à prendre en compte l'angularité inscrite entre sommets successifs comme contrainte sur le bruitage, ce qui doit nous permettre de remédier au problème de non-préservation de la forme initiale de l'objet.

Pour pondérer le bruit appliqué en chaque point de la géométrie d'un objet simulé, nous considèrerons que plus l'angle (en radian) inscrit entre trois sommets successifs est proche de π, plus l'amplitude de son bruit est faible. A contrario, plus l'angle est proche de 0 ou 2π, plus l'amplitude du bruit sera importante.

Modélisation Afin de mettre cette hypothèse en application, une pondération q de l'amplitude du bruit, comprise entre 0 et 1, est appliquée en chaque point P de la polyligne en fonction de l'angle (calculé en radian) formé par les points successifs P_{-1}, P, et P_{+1}. Par exemple, supposons trois points successifs A, B, et C sur une polyligne. Pour appliquer une pondération au bruit du point B, nous proposons de déterminer son angularité α_B, comprise entre 0 et 1, conformément à l'équation 4.4.

$$\alpha_B = \frac{|\ (\widehat{ABC} - \pi)\ |}{\pi} \tag{4.4}$$

Pour déterminer la pondération q_B du bruit au point B, un redressement doit être appliqué à la valeur d'angularité α_B par une fonction affine. En effet, selon notre hypothèse, plus l'angle est proche de π, plus le bruit est faible. Cela ne veut pas dire pour autant qu'il soit nul. Or, dans le cas où un angle inscrit entre trois points successifs est égal à π, une application stricte de la valeur d'angularité α_B (par une fonction affine $f(q) = \alpha$) impliquerait un bruit nul, ce qui n'est pas réaliste. De la même manière, il est rare dans une base de données, de disposer de points présentant une angularité proche de 0 ou 2π.

Il convient donc de définir quelle fonction affine $f(q) = a\alpha + b$ appliquer, afin de déterminer une pondération réaliste du bruit appliqué en chaque point de la géométrie en fonction de son angularité. Différentes fonctions, comme illustré dans la figure 4.10, seront donc expérimentées afin d'en déterminer la plus pertinente pour pondérer l'amplitude du bruit (cf. section 4.1.4).

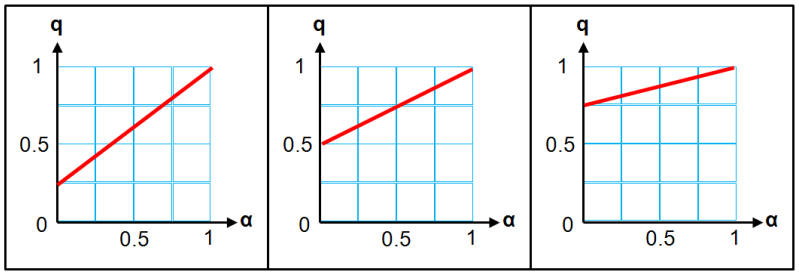


Figure 4.10 – Exemple de fonctions permettant de déterminer la pondération q du bruit gaussien en fonction de l'angularité α du point de la géométrie

Une fois la pondération q_B définie, elle est ensuite appliquée au bruit du point B, issu du tirage d'une valeur aléatoire suivant une loi normale $N(0,\sigma)$. L'opération est ensuite reproduite sur chaque point de la géométrie de l'objet. Ainsi, pour respecter la forme générale de la géométrie d'un objet, et en particulier son caractère rectiligne lorsque c'est la cas, cette pondération nous permet de n'appliquer qu'un bruit de faible amplitude aux points où l'angle inscrit se rapproche de π.

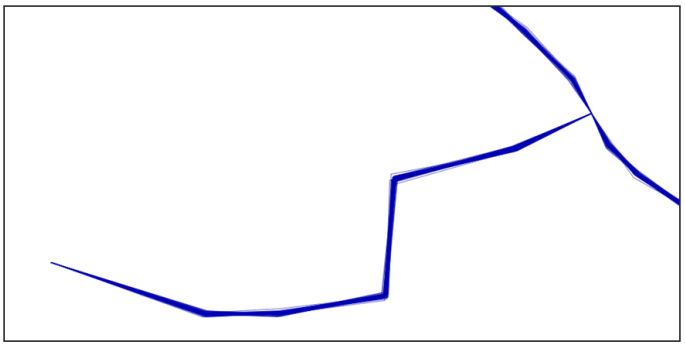


Figure 4.11 – Bruitage d'une polyligne en prenant en compte l'angularité

La figure 4.11 illustre le résultat d'une simulation sur des tronçons de route pondérée par l'angle inscrit entre les sommets successifs d'une polyligne. On remarque bien que plus l'angularité est forte, plus l'amplitude du bruit est importante.

Cependant, cette figure montre qu'aucun bruit n'est appliqué aux extrémités de la polyligne, étant donné qu'il n'est pas possible de leur appliquer une valeur d'angularité. Afin de paramétrer l'amplitude du bruit appliqué aux extrémités, nous proposons une méthode basée sur le degré des nœuds situés aux extrémités des polylignes.

Pondération par le degré des nœuds aux extrémités

Hypothèses Etant donné qu'il n'est pas possible d'appliquer une valeur d'angularité aux extrémités d'une polyligne, nous proposons de pondérer le bruit appliqué à ces extrémités par le degré de leur nœud. En effet, au sein d'un réseau routier par exemple, on peut considérer que plus d'arcs entrent et sortent d'un même nœud, plus la localisation de ce dernier est précise. Au contraire, si une extrémité de polyligne n'est connectée à aucun arc, on peut considérer que sa position est beaucoup plus imprécise (cf. figure 4.12).

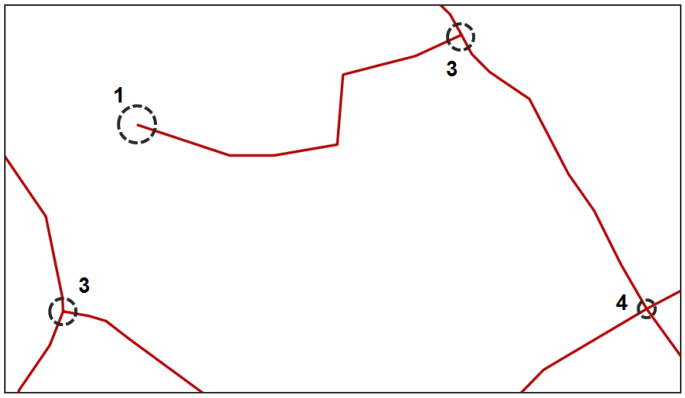


Figure 4.12 – Imprécision de localisation relative au degré des nœuds

Modélisation Cette hypothèse nous amène donc à considérer que le bruit appliqué aux extrémités d'une polyligne doit être pondéré par le degré du nœud. Ainsi, pour un nœud N de degré d localisé en extrémité de polyligne, nous proposons d'appliquer une pondération q_N, conformément à l'équation 4.5.

$$q_N = \frac{1}{d} \tag{4.5}$$

La pondération q_N est ensuite appliquée au bruit du nœud N, suite au tirage d'une valeur aléatoire suivant une loi normale $N(0,\sigma)$. La figure 4.13 reprend le résultat de la simulation réalisée sur les tronçons de route de la figure 4.11, en y ajoutant le bruitage sur les extrémités avec la méthode de pondération par le degré des nœuds.
On observe bien qu'en utilisant cette méthode, les extrémités ayant peu d'arcs entrants ou sortants présentent une imprécision de positionnement beaucoup plus importante que les extrémités connectées à d'autres arcs.

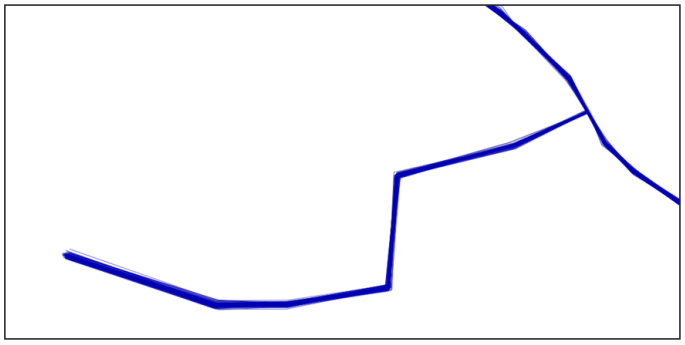


Figure 4.13 – Bruitage d'une polyligne en prenant en compte l'angularité et le degré des nœuds

Nous venons de présenter deux méthodes permettant de pondérer le bruit réalisé sur chaque point de la géométrie, à travers des contraintes d'angularité entre points successifs et de degré des nœuds situés aux extrémités. Ces méthodes permettent de préserver la forme générale des objets après simulation, et ainsi proposer une simulation réaliste. Nous allons voir dans la section suivante comment intégrer une seconde contrainte dans notre modèle de simulation de l'erreur de pointé, afin de prendre en compte la corrélation des erreurs entre points successifs de la géométrie simulée.

4.1.3.2 Corrélation des erreurs entre points successifs

De nombreux auteurs considèrent que la simulation d'erreurs sur une géométrie doit intégrer une corrélation entre points successifs (Heuvelink *et al.*, 2007; De Bruin, 2008). Pour mieux le comprendre, nous allons tout d'abord présenter les raisons justifiant l'intégration d'une corrélation de l'erreur entre points successifs lors du bruitage d'une géométrie, puis proposerons une méthode permettant de prendre en compte cette contrainte dans notre modèle de simulation de l'erreur de pointé.

Hypothèse Pour justifier la prise en compte de la corrélation des erreurs durant la saisie d'une géométrie, nous pouvons nous appuyer sur un exemple simple : imaginons un opérateur numérisant une route. Si une erreur est intégrée aux extrémités de la polyligne, c'est-à-dire aux carrefours, nous partons de l'hypothèse que l'opérateur va chercher à compenser cette erreur petit à petit sur l'ensemble de la géométrie, plutôt que d'une seule fois. En effet, la saisie est un processus dynamique (et rapide) supposant que si l'opérateur vise légèrement à côté, il cherchera à corriger l'erreur sans s'arrêter ou baisser son rythme de saisie. On peut donc considérer à travers cet exemple que l'erreur de positionnement des extrémités du tracé de la route est propagée sur l'ensemble des points de la géométrie, et qu'il existe donc une corrélation de l'erreur entre chacun des points. Cette hypothèse suppose donc que la saisie soit opérée d'un point de départ à un point d'arrivée, ce qui est cohérent avec la saisie des objets linéaires.

Afin de modéliser la corrélation des erreurs entre points successifs, deux approches ont été envisagées : l'utilisation d'un lissage gaussien, et la translation des sommets en fonction du bruit aux extrémités.

Lissage gaussien des objets bruités

La première approche envisagée pour corréler les erreurs entre points successifs repose sur l'application d'un lissage gaussien sur les géométries bruitées.

Le **lissage gaussien** (déjà présenté en section 3.3.2.3) est un algorithme qui a pour but d'éliminer les détails sans importances d'une polyligne. Cet algorithme fonctionne selon le principe d'une fenêtre glissante sur l'ensemble de la polyligne. Chaque point de la polyligne est déplacé vers le barycentre calculé avec les autres points de la polyligne localisés dans la fenêtre, en fonction de la distance curviligne les séparant (et en suivant une loi gaussienne). La force du lissage gaussien est déterminée selon un paramètre σ, qui correspond à l'écart-type de la loi gaussienne utilisée.

Application sur les géométries bruitées L'application d'un lissage gaussien en aval du bruitage d'une géométrie peut être envisagée afin d'intégrer une corrélation des erreurs entre points successifs d'une géométrie. En appliquant ce lissage, l'erreur de position en chaque point est recalculée en fonction de l'erreur appliquée sur les points voisins, ce qui permet de corréler ces erreurs.

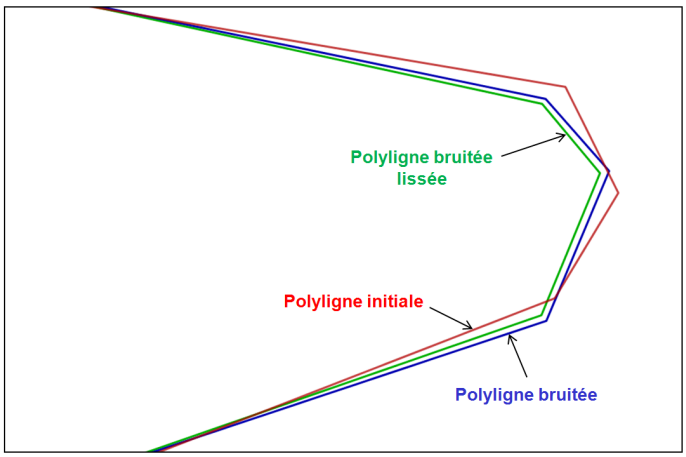


Figure 4.14 – Effet du lissage gaussien sur des tronçons de route bruités de la BDCARTO

Cependant, l'inconvénient de cet algorithme est qu'il a pour principal effet de "moyenner" la position des points de la géométrie lissée, en fonction de celle des points voisins. Ceci se traduit par une diminution systématique des longueurs calculées, comme le montre la figure 4.14, où la polyligne lissée suite au bruitage a tendance à "prendre la corde" dans les virages.

Cette observation rend l'utilisation du lissage gaussien problématique dans le cadre de la mise en œuvre du modèle de simulation de l'erreur de pointé, puisque les mesures géométriques calculées en aval risquent fortement d'être sous-estimées, ce qui n'est pas envisageable. Ainsi, pour prendre en compte la corrélation des erreurs entre points successifs d'une géométrie, nous proposons donc de nous orienter vers une autre approche, basée sur une translation des points de la géométrie selon l'erreur portée aux extrémités.

Translation pondérée par les erreurs aux extrémités

Une autre approche, inspirée des travaux de Bonin (2002) et Fouqué (1999) a donc été préférée pour modéliser la corrélation entre points successifs d'une géométrie. En reprenant l'exemple d'un réseau routier, cette approche se décline en deux étapes :

1. les nœuds du réseau (les carrefours) sont bruités indépendamment selon une loi normale $N(0,\sigma)$. Chacun des deux points A et C situés aux extrémités subissent donc un déplacement suivant un vecteur $\overrightarrow{v}$ et $\overrightarrow{u}$, pondéré par degré du nœud.
2. les points intermédiaires B des polylignes composant le réseau sont translatés selon un vecteur $\overrightarrow{w}$, calculé par une somme pondérée des vecteurs $\overrightarrow{v}$ et $\overrightarrow{u}$. La pondération est le résultat du rapport entre la longueur de la polyligne allant d'un point intermédiaire à l'extrémité de la polyligne, par rapport à la longueur totale de la polyligne, comme expliqué dans l'équation 4.6.

$$\overrightarrow{w} = a * \overrightarrow{v} + b * \overrightarrow{u} \tag{4.6}$$

où a est le rapport entre la longueur de $[AB]$ sur la longueur de $[AC]$
où b est le rapport entre la longueur de $[BC]$ sur la longueur de $[AC]$

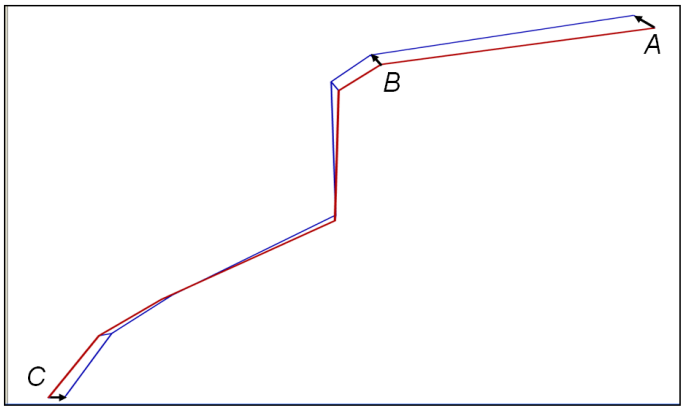


Figure 4.15 – Principe de translation d'un point B en fonction des erreurs aux extrémités A et C d'une polyligne

La figure 4.15 permet d'illustrer ce principe de corrélation des erreurs entre sommets successifs, en fonction de l'erreur affectant les extrémités de la polyligne.
On notera bien que le bruitage aléatoire n'est ici appliqué qu'aux extrémités de la polyligne, c'est-à-dire aux nœuds du réseau. La translation de chacun des points intermédiaires de la géométrie est effectuée en amont du bruitage de la géométrie par le biais du modèle aléatoire présenté précédemment.

Cette méthode basée sur une translation des sommets préalable au bruitage aléatoire permet donc d'intégrer une contrainte de corrélation des erreurs entre points successifs dans notre approche. Nous allons maintenant décrire comment combiner l'ensemble de ces contraintes en fonction de la géométrie des objets, et de leurs caractéristiques.

4.1.3.3 Combinaison des contraintes pour la simulation d'erreurs

Afin de reproduire de manière réaliste l'impact de l'erreur de pointé de l'opérateur à l'aide de simulations, l'approche développée dans ce travail vise donc à intégrer des contraintes en termes de respect de la forme initiale des objets, et de corrélation d'erreurs entre points successifs. Cependant, nous allons voir que ces contraintes ne sont pas toutes prises en compte en fonction de la géométrie et, le cas échéant, de la nature des objets étudiés. Nous étudierons donc successivement le cas des objets linéaires et des objets surfaciques.

Objets linéaires Pour les objets linéaires, comme exposé dans les sections précédentes, nous considérons que le bruit aléatoire doit être intégré en prenant en compte les contraintes de corrélation des erreurs et de respect de la forme initiale des objets.
Ainsi, pour chaque réalisation d'une simulation, les deux contraintes sont intégrées successivement. En prenant l'exemple d'un réseau routier, les différentes étapes du modèle de simulation sont :

1. Bruitage des nœuds (carrefours du réseau) pondéré par leur degré,
2. Translation des sommets intermédiaires en fonction des erreurs aux extrémités,
3. Bruitage des sommets intermédiaires translatés, pondéré par l'angularité du sommet.

Objets surfaciques Pour les géométries surfaciques, il est nécessaire de prendre en compte les caractéristiques des objets traités. Ainsi, nous proposons de différencier les objets représentant une partition de l'espace (comme les limites administratives) des objets n'en représentant pas une (comme des surfaces d'eau par exemple). En effet, ces objets n'ayant pas les mêmes propriétés, nous considérons qu'il faut les contraindre différemment pour obtenir des résultats réalistes.

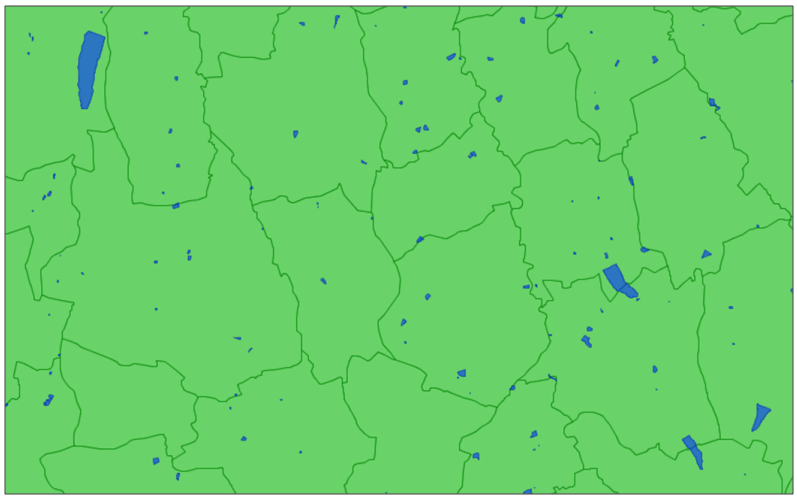


Figure 4.16 – Exemple d'objets surfaciques représentant une partition de l'espace (les limites communales, en vert) et n'en représentant pas une (les surfaces d'eau, en bleu)

Les objets représentant une partition de l'espace, comme les limites administratives par exemple, supposent l'existence de frontières communes. Nous pouvons supposer qu'il existe une corrélation des erreurs entre sommets successifs, puisqu'un opérateur qui saisit une limite communale s'appuiera sur des nœuds existants (une limite entre trois communes)

pour saisir une nouvelle frontière. Nous pouvons imaginer qu'il cherchera à amortir l'erreur en chacun des nœuds sur l'ensemble de la frontière saisie. Dans ce cas de figure, la même méthodologie que celle proposée pour les objets linéaires est appliquée.

Ceci n'est pas le cas pour les objets ne représentant pas une partition de l'espace, comme la délimitation d'un lac ou d'une forêt par exemple. En effet, un polygone étant une géométrie close, la notion de points d'arrivée et de départ n'existe pas. Il ne semble donc pas pertinent d'intégrer une corrélation entre points successifs. Dans ce cas de figure, seule la contrainte impliquée par l'angularité entre sommets successifs, afin de préserver la forme générale des objets simulés, sera introduite dans la simulation.
Ces différentes propositions de combinaison des contraintes, en fonction de la géométrie et des caractéristiques des objets, seront mises en œuvre dans le cadre de l'application du modèle de simulation de l'erreur de pointé (cf. section 4.1.4).

Nous venons de présenter dans cette section les principes permettant de procéder à une simulation réaliste de l'erreur de pointé sur des objets vectoriels, en décrivant tout d'abord comment paramétrer le bruit en fonction de l'échelle de saisie, puis en proposant des méthodes de simulation respectant des contraintes en termes de corrélation d'erreurs et de respect de la forme initiale des objets. Nous allons à présent proposer une expérimentation qui aura pour objectif de valider nos hypothèses en termes de simulation de l'erreur, et le cas échéant, de calibrer notre modèle.

Synthèse Section Afin d'étudier l'impact de l'erreur de pointé sur les mesures géométriques, nous proposons une approche par simulations intégrant le respect de contraintes en termes de corrélation d'erreurs et d'angularité entre points successifs de la géométrie des objets évalués. Cependant toutes ces contraintes ne sont pas systématiquement intégrées en fonction de la géométrie et des propriétés des objets évalués.

4.1.4 Expérimentation pour la validation du modèle de simulation

Afin de valider les hypothèses et les choix de modélisation que nous venons de présenter pour simuler l'erreur de pointé de l'opérateur, une expérimentation a été menée. Cette expérimentation repose sur la saisie manuelle d'un tronçon de route par un échantillon d'utilisateurs, à différentes échelles fixées (cf. Annexe D).

4.1.4.1 Présentation de l'expérimentation

Objectifs Les objectifs de cette expérimentation de saisie d'un tronçon de route sont multiples. Cette expérimentation doit tout d'abord nous permettre de quantifier l'impact de l'échelle de saisie sur l'imprécision de pointé, ainsi que sur l'imprécision des mesures de longueur. Les résultats devant confirmer, ou non, nos hypothèses en termes de paramétrage du bruit gaussien. Egalement, cette expérimentation doit nous permettre de valider la méthode d'estimation de l'échelle moyenne de saisie. Au final, nous chercherons à calibrer le modèle de simulation de l'erreur de pointé, et, le cas échéant, de le valider.

Méthodologie Le test de saisie manuelle a été réalisé sur un tronçon de route (la route D112) extrait de la BDTOPO, dans le département de l'Isère. Les individus participant au test ont pour mission de saisir l'axe de la route à deux échelles : au 1 :10.000 puis au 1 :25.000 sans avoir possibilité de zoomer ou de changer l'échelle de visualisation, qui demeure fixe. Chacun des deux tronçons (pour chacune des échelles expérimentées) doit être saisis en une seule fois, c'est-à-dire en une géométrie unique. Le test de saisie a été réalisé avec le logiciel libre QuantumGIS (version 1.7), comme illustré dans la figure 4.17, sur PC.

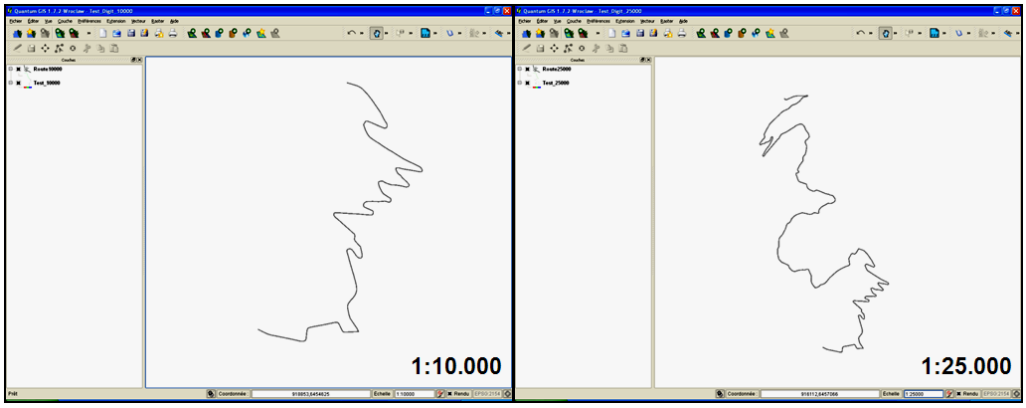


Figure 4.17 – Tronçons de route utilisés pour le test de saisie, représentés au 1 :10.000 et au 1 :25.000

Afin de permettre une représentation appropriée du tronçon de route en fonction de l'échelle de représentation, une taille de symbole de 0,7 millimètres a été appliquée, en référence aux spécifications cartographiques utilisées en France. Ainsi, des zones tampons de 3,5 mètres au 1 :10.000 et de 8,75 mètres au 1 :25.000 ont été générées, puis converties au format image afin d'éviter toute possibilité d'accrochage entre les tronçons saisis et le tronçon de route de référence.

Echantillon L'échantillon exploité dans le cadre de cette expérimentation est constitué de 20 individus, tous membres à cette période du laboratoire COGIT de l'IGN. Cet échantillon est constitué dans l'ensemble d'individus ayant une bonne connaissance des données géographiques et pratiquant, dans leur grande majorité, la saisie d'objets géographiques vectoriels de manière occasionnelle.

En effet, à la question "Quelle est votre fréquence de saisie d'objets géographiques vectoriels ?", 75% ont répondu "Occasionnellement", 10% "Fréquemment", 15% "Jamais", et 0% "Quotidiennement".

Nous convenons qu'il aurait été préférable de réaliser cette expérimentation sur un échantillon composé de restituteurs photogrammètres ou de dessinateurs cartographes, qui réalisent ce type de tâche de manière quotidienne, mais par manque de temps, ceci ne pu être organisé. Nous prendrons cependant en compte le profil de cet échantillon dans notre analyse finale.

4.1.4.2 Résultats de l'expérimentation

Nous présenterons dans cette section les résultats de l'expérimentation relatifs à l'estimation de l'imprécision de pointé, puis ceux relatifs à l'imprécision de la longueur de la route, en fonction de l'échelle de saisie. Enfin, une analyse des distances inter-points sera proposée pour valider la méthode d'estimation de l'échelle de saisie.

Imprécision de pointé

Suite aux saisies réalisées par les 20 individus de l'échantillon, 10 points significatifs ont été extraits sur chacun des tronçons de route saisis au 1 :10.000 et au 1 :25.000, afin d'étudier l'imprécision de pointé (cf. figure 4.18).

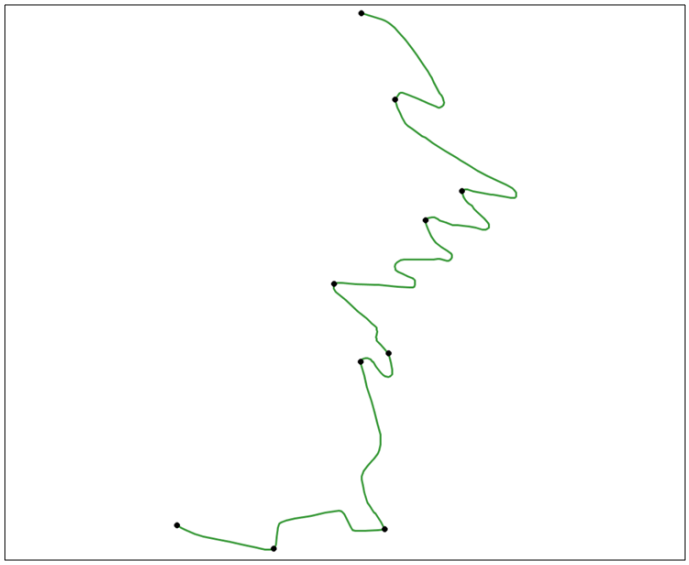


Figure 4.18 – Localisation des 10 points sélectionnés pour étudier l'imprécision de pointé des tronçons saisis

Ces 10 points ont été sélectionnés du fait de leur facilité d'identification sur chacune des routes numérisées. La position de chacun de ces points est comparée, par le biais d'un contrôle ponctuel, avec la position du point homologue du tronçon de route de la BD-TOPO. Les résultats de cette comparaison sont proposés dans la table 4.2.

	1 :10.000			**1 :25.000**		
Indicateur	Δ_d(m)	Δ_x(m)	Δ_y(m)	Δ_d(m)	Δ_x(m)	Δ_y(m)
Moyenne	5,06	0,36	-1,63	10,68	-0,12	-2,58
Ecart-type	2,74	3,67	4,12	5,59	8,18	8,50

Table 4.2 – Imprécision de pointé (en mètres) à partir des 10 points sélectionnés

Les résultats de la comparaison de la position de ces 10 points montrent tout d'abord que l'échelle de saisie impacte directement l'imprécision de pointé. En effet l'imprécision de pointé des dix points est en moyenne deux fois plus importante au 1 :25.000 qu'au 1 :10.000. L'écart-type de l'erreur de distance des 10 points, qui correspond à l'imprécision de pointé σ_p, est de 2.7 m au 1 :10.000 et de 5.6 m au 1 :25.000. La figure 4.19 permet d'illustrer cette imprécision de pointé sous forme de cible.

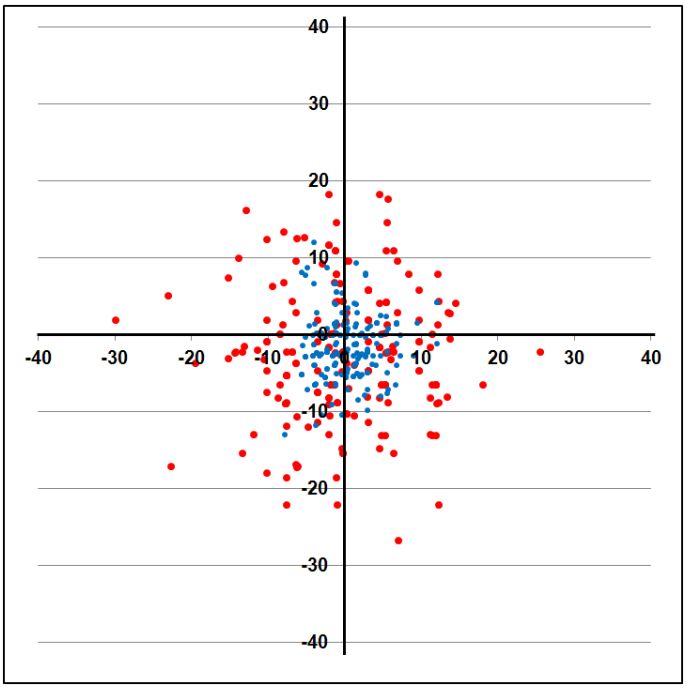


Figure 4.19 – Résultats de l'étude de l'imprécision de pointé (en mètres) à partir des 10 points sélectionnés, au 1 :10.000 (en bleu) et au 1 :25.000 (en rouge)

Nous constatons cependant suite à ces comparaisons, que σ_p^2 est environ égale à 7 mètres au 1 :10.000, ce qui suppose une valeur de $\sqrt{2\sigma_c} = 0{,}7$ mm, alors qu'en théorie, comme le propose Chandelier (2011), cette dernière est supposée être égale à 0,2 mm. Nous pouvons justifier cette différence, relativement importante, par trois raisons principales.

Tout d'abord, les conditions dans lesquelles cette expérimentation a été menée ne sont pas comparables à celles proposées en restitution photogrammétrique, tant au niveau logiciel que matériel (où par exemple une *trackball* remplace généralement la souris). Egalement, le profil de l'échantillon n'est pas celui de restituteurs photogrammètres qualifiés, ce qui peut se répercuter sur l'imprécision de pointé. Enfin, la symbologie appliquée sur le tronçon de référence de la BDTOPO peut avoir pour impact, principalement au 1 :25.000, de rendre délicate la saisie de l'axe de la route, étant donné que l'emprise au sol est plus importante.

Egalement, les résultats montrent l'existence d'un biais de saisie en y (par la moyenne des Δ_y), mais pas en x. Ce biais peut s'expliquer par les habitudes de saisie prises par certains individus de l'échantillon, qui ont davantage tendance à pointer en dessous de la route, plutôt que dans son axe. L'impact de la symbologie évoqué précédemment peut également justifier ce biais. Cependant, puisque nous nous intéressons dans le cadre de cette expérimentation à l'imprécision de pointé, et donc à la dispersion des erreurs, ce biais n'impacte pas véritablement la suite de notre travail.

Imprécision de longueur

Afin d'étudier l'imprécision des longueurs calculées à partir de la géométrie des routes saisies par l'échantillon, nous avons isolé la section sud de la route saisie au 1 :25.000, qui correspond à la route saisie au 1 :10.000, comme le montre la figure 4.20. L'étude de la saisie du même tronçon au 1 :10.000 et au 1 :25.000 nous permet d'étudier l'impact de l'échelle de saisie sur l'imprécision de la longueur calculée.

Figure 4.20 – Localisation du tronçon de route (en rouge) utilisé pour étudier l'imprécision des mesures de longueur

Le tronçon de route initial, extrait de la BDTOPO, présente une longueur de 4,27 km. Cette longueur constitue donc notre longueur de référence. Nous l'avons comparée avec les longueurs des tronçons saisis au 1 :10.000 et au 1 :25.000 par l'échantillon.

	1 :10.000	**1 :25.000**
Minimum(m)	4222,74	4068,03
Maximum(m)	4336,37	4382,99
Moyenne(m)	4266,17	4236,38
Ecart-type(m)	28,51	79,93

Table 4.3 – Imprécision des mesures de longueur (en mètres) au 1 :10.000 et au 1 :25.000

Les résultats de ces comparaisons, présentés dans la table 4.3, montrent que l'échelle de saisie impacte également l'imprécision de la mesure de longueur calculée. En effet, l'écart-type des erreurs de longueur des routes saisies au 1 :10.000 est de 28,5 mètres, tandis qu'au 1 :25.000, il est de 79,9 mètres (soit un rapport de 2,8). Ces résultats montrent bien que plus l'échelle de saisie diminue, plus la longueur calculée est imprécise.

Ces résultats nous permettent également de constater une tendance à la sous-estimation des longueurs avec la réduction d'échelle. En effet, la moyenne des longueurs des routes saisies au 1 :25.000 est inférieure d'environ 30 mètres à la longueur de référence, tandis que la longueur moyenne des routes saisies au 1 :10.000 est pratiquement identique à la longueur de référence.

Nous pouvons expliquer cela par la baisse du niveau de détail des objets saisis au 1 :25.000. En effet, comme nous l'avons évoqué précédemment, plus l'échelle de saisie diminue, plus le niveau de détail baisse, et plus la distance entre points successifs d'une géométrie a tendance à augmenter (comme dans le cas d'un filtrage de points par l'algorithme de Douglas et Peucker (1973)). Cette baisse du niveau de détail a principalement pour conséquence de réduire la longueur d'une polyligne.

L'étude des distances inter-points va nous permettre de quantifier cette baisse du niveau de détail en fonction de l'échelle de saisie, mais également de valider la méthode d'estimation de l'échelle moyenne de saisie.

Etude des distances inter-points

L'étude des distances inter-points a été entreprise dans le cadre de cette expérimentation afin de valider la méthode d'estimation de l'échelle moyenne de saisie proposée en section 4.1.2.2. Elle offre également la possibilité de quantifier l'impact de la réduction d'échelle sur le niveau de détail des objets.
Les figures 4.21 et 4.22 présentent respectivement les distributions des distances inter-points pour la section homologue des routes saisies au 1 :10.000 et au 1 :25.000.

Pour quantifier l'impact de l'échelle de saisie sur le niveau de détail des objets, la granularité (qui correspond aux plus petites distances inter-points observées) peut être utilisée. Pour les objets saisis au 1 :10.000, la plus petite distance observée et de 5,29 mètres, tandis qu'elle est de 6,61 mètres au 1 :25.000. Ces résultats montrent bien que la baisse de l'échelle de saisie impacte les plus petites distances inter-points observées. Cependant, nous intéresserons davantage à la distance inter-points médiane, puisque cet indicateur est utilisé afin d'estimer l'échelle moyenne de saisie.

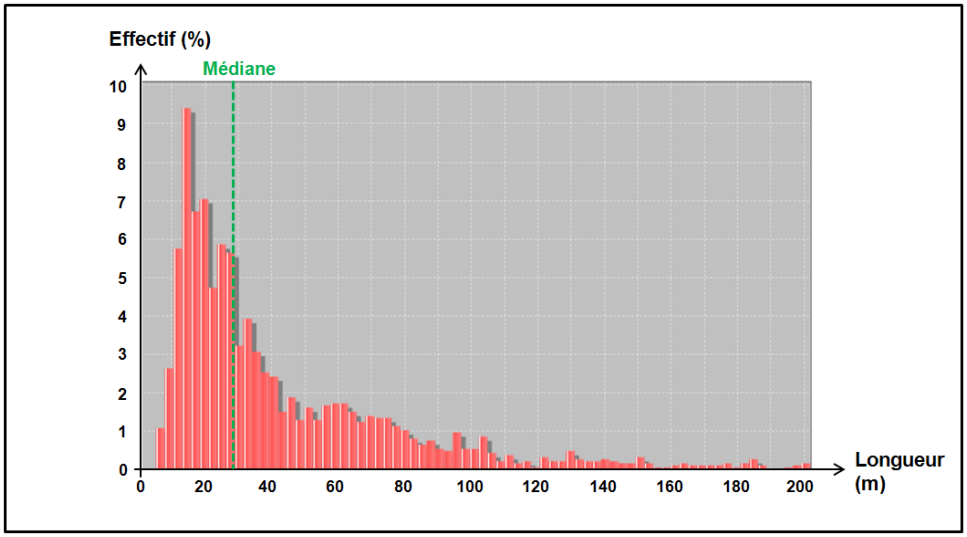


Figure 4.21 – Distribution des distances inter-points des routes saisies au 1 :10.000

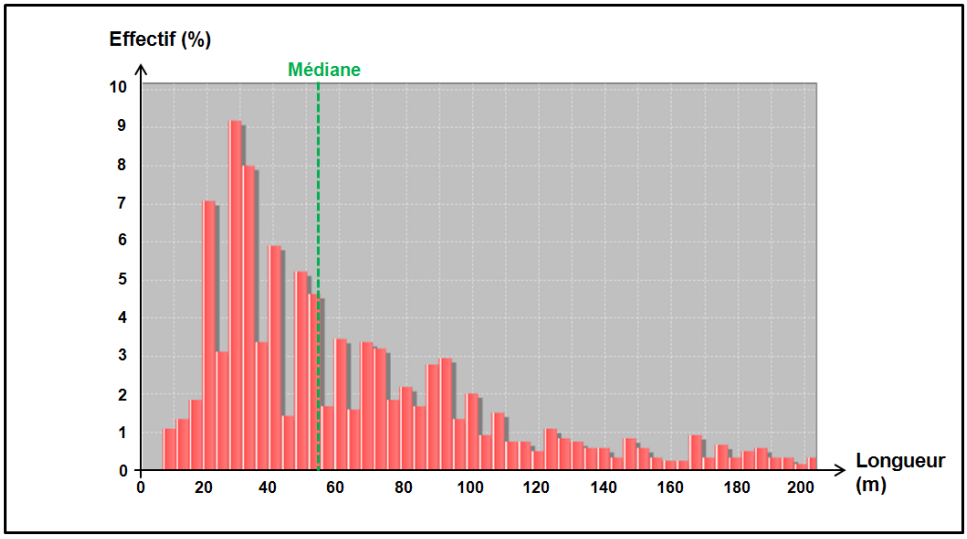


Figure 4.22 – Distribution des distances inter-points des routes saisies au 1 :25.000

La distance inter-points médiane observée est de 29,58 mètres pour les tronçons saisis au 1 :10.000, et de 53,33 mètres pour les tronçons saisis au 1.25.000. Par application de la méthode d'estimation de l'échelle moyenne de saisie proposée en section 4.1.2.2, les échelles de saisie moyennes sont respectivement estimées au 1 :14.790 (au lieu du 1 :10.000) et au 1 :26.665 (au lieu du 1 :25.000).

Ces résultats montrent que la méthode d'estimation basée sur la distance inter-points médiane a tendance à légèrement sur-estimer l'échelle moyenne de saisie (en particulier pour les données saisies au 1 :10.000). Bien qu'incorrecte, cette estimation constitue malgré tout une approximation réaliste. C'est pourquoi nous conserverons la méthode ainsi proposée pour la suite de ce travail.

Bilan de l'expérimentation Les résultats de l'expérimentation qui ont été présentés dans cette section nous ont permis de quantifier les effets de la réduction de l'échelle de saisie sur :
– l'imprécision de pointé de l'opérateur,
– l'imprécision des mesures de longueur calculées,
– le niveau de détail des objets.

Nous proposerons dans la section suivante d'exploiter les résultats de cette expérimentation dans le cadre d'une mise en application du modèle de simulation de l'erreur de pointé, afin de le calibrer et, le cas échéant, de valider ses estimations.

4.1.4.3 Calibration du modèle de simulation

Nous proposons dans cette section de mettre en application le modèle de simulation de l'impact de l'erreur de pointé afin :
– de le calibrer,
– de valider ses estimations sur l'imprécision des mesures de longueur.
Pour cela, nous proposons de l'appliquer sur l'ensemble des tronçons saisis par les individus de l'échantillon, aux deux échelles expérimentées.

Méthodologie Le principe de calibration du modèle de simulation de l'erreur de pointé repose sur l'analyse de la dispersion des longueurs simulées. L'objectif de cette analyse est d'évaluer si la longueur de référence du tronçon de route de la BDTOPO (qui est de 4,27 km) est inscrite dans un intervalle de confiance compris entre la longueur du tronçon simulé le plus court L_{min} et la longueur du tronçon simulé le plus long L_{max}. Cette analyse est effectuée pour chacun des tronçons saisis par l'échantillon, aux deux échelles expérimentées.

Le modèle de simulation sera mis en application avec plusieurs combinaisons des paramètres :
– du bruit gaussien σ_p, en nous basant sur des valeurs théoriques de $\sqrt{2\sigma_c}$ allant de 0,2 à 0,9 mm en fonction de l'échelle de saisie,
– de la fonction d'amplitude du bruit gaussien selon l'angularité entre points successifs $f(q) = a\alpha + b$.
L'objectif de cette étude est d'identifier la combinaison des paramètres qui offre le meilleur pourcentage d'estimation pour calibrer le modèle de simulation. Si les résultats obtenus par une combinaison de paramètres sont satisfaisants, le modèle ainsi calibré sera validé.

Calibration du modèle de simulation

Le modèle de simulation de l'erreur de pointé a donc été appliqué sur chacun des tronçons de route, saisis au 1 :10.000 et au 1 :25.000, avec les paramètres décrits précédemment. Pour chaque tronçon saisi, une simulation de 1.000 réalisations a été générée.

Les tableaux 4.4 et 4.5 présentent les pourcentages d'estimations correctes de l'imprécision de longueur (c'est-à-dire pour lesquels la longueur de référence est inscrite dans l'intervalle de confiance) selon les paramètres utilisés.

Paramétrage du bruit au 1 :10.000								
$\sqrt{2\sigma_c}$(mm)	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
σ_p^2(m)	2	3	4	5	6	7	8	9
σ_p(m)	1.41	1.73	2	2.23	2.44	2.64	2.83	3
Fonction utilisée	**Pourcentage d'estimations correctes**							
$f(q) = \alpha$	25%	30%	30%	30%	35%	45%	45%	45%
$f(q) = 0.75\alpha + 0.25$	30%	30%	40%	50%	50%	50%	60%	60%
$f(q) = 0.5\alpha + 0.5$	40%	45%	50%	55%	65%	70%	70%	75%
$f(q) = 0.25\alpha + 0.75$	45%	55%	70%	70%	70%	75%	80%	90%

Table 4.4 – Pourcentages d'estimations de longueurs correctes pour les routes saisies au 1 :10.000

Au 1 :10.000 Les résultats de l'application du modèle sur les routes saisies au 1 :10.000 confirment que pour estimer l'impact de l'erreur de pointé sur l'imprécision de mesure, une valeur d'imprécision de pointé $\sqrt{2\sigma_c}$ supérieure à 0,2 mm doit être utilisée. Ceci s'explique en grande partie par le profil de l'échantillon, composé en grande majorité de personnes saisissant "occasionnellement" des données géographiques, et par le matériel utilisé. Les résultats montrent qu'afin d'obtenir un pourcentage acceptable (supérieur ou égal à 75%), une imprécision de pointé de 0,7 mm doit être utilisée. En ce qui concerne le choix de la fonction d'amplitude du bruit selon l'angularité inscrite entre points successifs, les résultats montrent que la fonction $f(q) = 0.25\alpha + 0.75$ propose les résultats les plus satisfaisants. Cette fonction a pour effet de réduire l'effet de l'angularité sur le paramétrage du bruit, mais son application permet malgré tout de respecter la forme générale des objets.

Paramétrage du bruit au 1 :25.000								
$\sqrt{2\sigma_c}$(mm)	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
σ_p^2(m)	5	7.5	10	12.5	15	17.5	20	22.5
σ_p(m)	2.23	2.74	3.16	3.53	3.87	4.18	4.47	4.74
Fonction utilisée	**Pourcentage d'estimations correctes**							
$f(q) = \alpha$	25%	30%	30%	35%	35%	35%	40%	45%
$f(q) = 0.75\alpha + 0.25$	30%	35%	35%	35%	40%	45%	45%	45%
$f(q) = 0.5\alpha + 0.5$	35%	35%	40%	45%	50%	50%	60%	60%
$f(q) = 0.25\alpha + 0.75$	35%	40%	45%	50%	55%	55%	60%	60%

Table 4.5 – Pourcentages d'estimations de longueurs correctes pour les routes saisies au 1 :25.000

Au 1 :25.000 Les résultats de l'application du modèle de simulation de l'erreur de pointé sont moins satisfaisants sur les objets saisis au 1 :25.000. Le pourcentage d'estimations correctes n'atteint que 60% avec une imprécision de pointé de 0,9 mm combinée à la fonction $f(q) = 0.25\alpha + 0.75$. Ces résultats peuvent s'expliquer par différentes raisons.
Au delà du profil de l'échantillon ou des conditions matérielles de l'expérimentation, nous avons souligné précédemment que l'imprécision de longueur augmente lorsque l'échelle diminue. Ceci s'explique par la baisse du niveau de détail des objets saisis, ou encore par des pointés trop imprécis. Ainsi, sur certains objets présentant une longueur beaucoup plus courte (de près de 200 m. pour la route la plus courte), l'intervalle de confiance estimé par le modèle ne s'avère pas suffisant pour encadrer la longueur de référence.

Egalement, afin de faciliter la saisie au 1 :25.000, une symbologie a du être appliquée sur la route à saisir. A cette échelle, cette symbologie implique des effets relatifs à la généralisation cartographique (que nous étudierons par ailleurs). En effet, l'emprise du symbole en unités terrain devenant importante, elle a pour effet de générer des imprécisions de pointé, et par conséquent des erreurs de longueur importantes. On ne peut donc pas négliger l'impact de la symbologie sur ces résultats. La mise en application du modèle sur les routes correctement saisies, en éliminant celles présentant des erreurs de longueur trop importantes (que nous pourrions qualifier d'*outliers*), aurait fourni de bien meilleurs résultats.

Calibration du modèle A partir des résultats obtenus sur les routes saisies au 1 :10.000 et au 1 :25.000, nous observons que la combinaison offrant les meilleurs résultats est la suivante :
– une imprécision de pointé $\sqrt{2\sigma_c} = 0{,}7$ mm,
– la fonction d'amplitude bruit selon l'angularité $f(q) = 0.25\alpha + 0.75$.
Afin d'éviter de générer un bruit trop important, un paramétrage de l'imprécision de pointé à 0,7 mm offre des résultats relativement satisfaisants au 1 :10.000. Cette estimation rejoint les résultats de l'étude de l'imprécision de pointé présentée auparavant.

Ainsi, la première conclusion qui peut être tirée de ce paramétrage, c'est qu'au delà de la valeur $\sqrt{2\sigma_c}$, qui est supérieure à la valeur théorique de 0,2 mm, nous observons qu'un contrôle ponctuel à partir d'un échantillon de points, offre la possibilité de paramétrer le modèle de simulation. Cette observation rejoint celle de Hunter et Goodchild (1996), qui proposent d'utiliser l'Erreur Moyenne Quadratique afin de paramétrer le bruit de leur modèle basé sur une approche par simulations.

Pour pondérer l'amplitude du bruit gaussien en fonction de l'angularité, nous préconisons d'utiliser une fonction $f(q) = 0.25\alpha + 0.75$, qui offre la possibilité de préserver, sans trop contraindre, la forme générale des objets évalués.
Enfin, en termes de mise en application, nous avons vu que le modèle de simulation est relativement performant pour les objets saisis au 1 :10.000, mais moins au 1 :25.000, pour les raisons évoquées précédemment. Des études complémentaires devront ainsi être réalisées afin de valider sa mise en application au 1 :25.000.

Synthèse Section Une expérimentation, reposant sur la saisie d'un tronçon de route au 1 :10.000 et au 1 :25.000 par un échantillon de 20 personnes, a été proposée dans cette section. Cette expérimentation a permis de confirmer que l'échelle de saisie impact directement l'imprécision de pointé, mais également l'imprécision des mesures de longueur. Le modèle de simulation a ensuite été appliqué sur les tronçons de route saisis, afin de voir s'il permettait d'estimer la longueur de référence, en utilisant différents paramètres. L'application du modèle est satisfaisante au 1 :10.000, mais peu au 1 :25.000. Les résultats montrent que l'exploitation de l'erreur moyenne quadratique pour paramétrer l'amplitude du bruit est une solution efficace. Egalement, une fonction de pondération de l'amplitude du bruit gaussien en fonction de l'angularité peu contraignante offre les meilleurs résultats.

4.1.5 Application du modèle d'estimation

Cette section a pour objectif de mettre en application le modèle d'estimation de l'impact de l'erreur de pointé à présent calibré, suite à l'expérimentation présentée dans la section précédente. Comme nous l'avons vu, différents scénarios de simulation sont proposés en fonction de la géométrie des objets et de leurs caractéristiques. Nous proposerons donc tout d'abord de mettre en application le modèle sur des géométries linéaires, puis ensuite sur des géométries surfaciques.

4.1.5.1 Application sur des géométries linéaires

Données utilisées Afin de mettre en application le modèle de simulation de l'impact de l'erreur de pointé sur les mesures de longueur, une extraction du réseau routier de la BDCARTO a été utilisée (cf. figure 4.23). Cette extraction du réseau routier correspond à l'ensemble des tronçons de route localisés dans les cantons d'Hendaye et de Saint-Jean-de-Luz, dans le département des Pyrénées-Atlantiques. La longueur totale du réseau extrait est de 420.0 kilomètres.

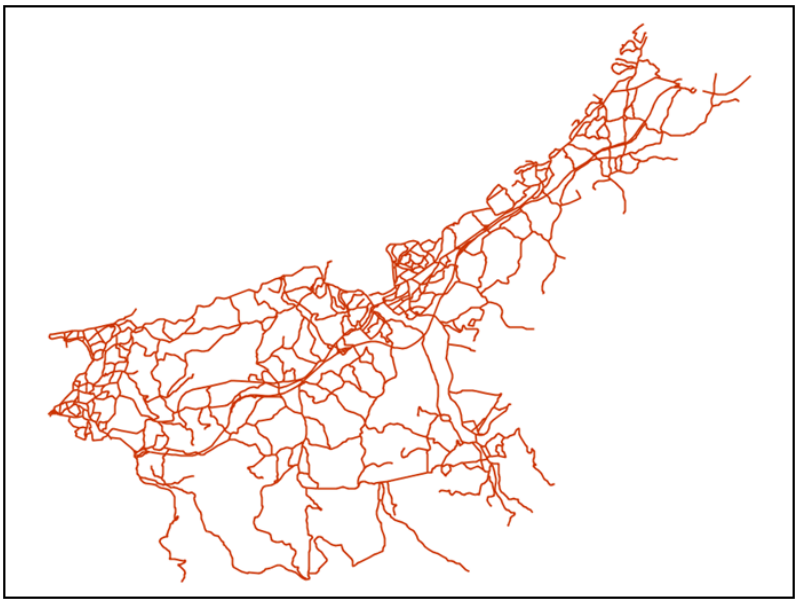


Figure 4.23 – Extraction du réseau routier de la BDCARTO sur les cantons d'Hendaye et de Saint-Jean-de-Luz

Nous avons bien conscience qu'en plus de l'impact de l'erreur de pointé, la BDCARTO est affectée par l'impact de la généralisation cartographique. Cependant, nous nous contenterons dans cette section d'estimer uniquement l'impact de l'erreur de pointé sur les mesures de longueur. L'impact de la généralisation cartographique sera traité dans la section 4.2.

Paramétrage du modèle Pour paramétrer le modèle de simulation, l'échelle de saisie moyenne de la BDCARTO a du être déterminée. Bien que nous sachions que la BDCARTO est initialement saisie à partir de cartes représentées au 1 :50.000, la méthode d'estimation de l'échelle de saisie caractéristique, basée sur l'exploration des distances inter-points, a été utilisée.

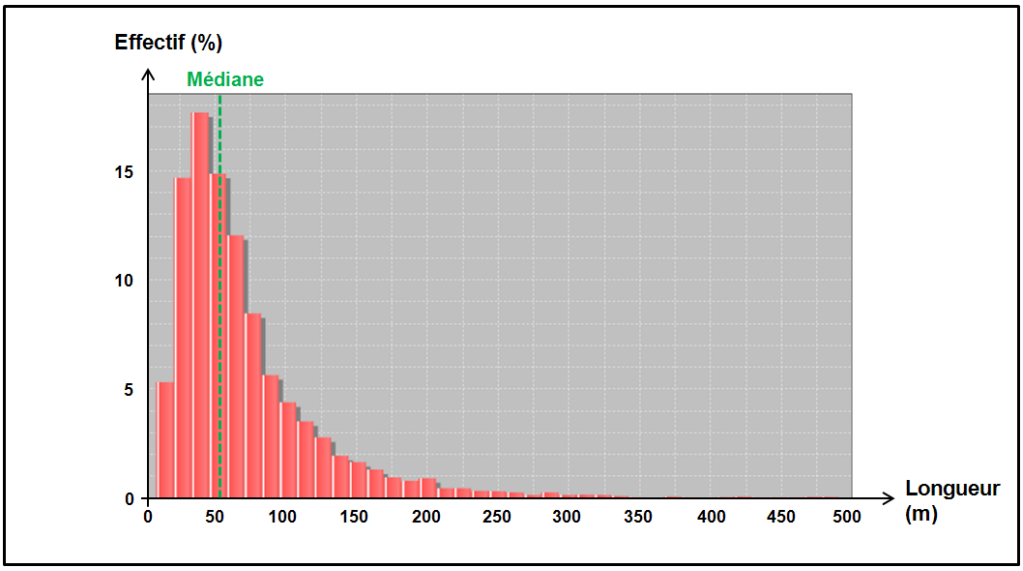


Figure 4.24 – Distribution des distances inter-points sur le réseau routier de la BDCARTO

La distribution des distances inter-points est présentée en figure 4.24. Une distance inter-points médiane de 55,54 mètres y est observée. En nous basant sur la méthode d'estimation présentée en section 4.1.2.2, l'échelle de saisie moyenne a été estimée au 1 :27.771.

A partir de cette estimation de l'échelle de saisie, et en utilisant une valeur $\sqrt{2\sigma_c} = 0{,}7$ mm conformément aux résultats de l'expérimentation, une valeur $\sigma_p = 4{,}41$ m est donc utilisée pour paramétrer l'amplitude du bruit lors de la simulation. La fonction $f(q) = 0.25\alpha + 0.75$ est retenue pour pondérer le bruit gaussien en fonction de l'angularité entre points consécutifs. Pour mener à bien la simulation sur les tronçons de route de la BDCARTO, 1.000 réalisations sont générées. Sur chaque réalisation générée, l'erreur de longueur est déterminée par la différence entre la longueur initiale et la longueur du tronçon simulé.

Résultats de la simulation L'application du modèle de simulation de l'erreur de pointé sur les tronçons de route montre tout d'abord que ce modèle a pour tendance de sur-estimer les longueurs géométriques. En effet, on observe que la moyenne des erreurs de longueur est négative, d'environ -652,92 mètres. Ceci signifie que les tronçons simulés sont en moyenne plus longs que les tronçons initiaux, d'environ 0,15%. Ces résultats montrent que le bruitage des géométries linéaires a pour tendance de surestimer systématiquement la longueur initiale, malgré l'intégration de contraintes de respect de la forme initiale des objets, nécessaire en vue de préserver le réalisme des objets simulés.

Cependant, le biais observé dans la distribution des erreurs de longueur était relativement prévisible. Cette sur-estimation peut en effet s'expliquer par le fait que de nombreux tronçons de route sont localisés en milieu urbain, où les routes présentent généralement des formes rectilignes qui minimisent leurs longueurs. Par conséquent, l'application d'un bruit, même avec une contrainte de respect de forme (peu importante dans notre application), ne peut empêcher la surestimation des longueurs simulées. Ce premier résultat nous incite donc à réfléchir à l'intégration de nouvelles contraintes dans le modèle de simulation, afin d'éviter ce biais dans les longueurs simulées.

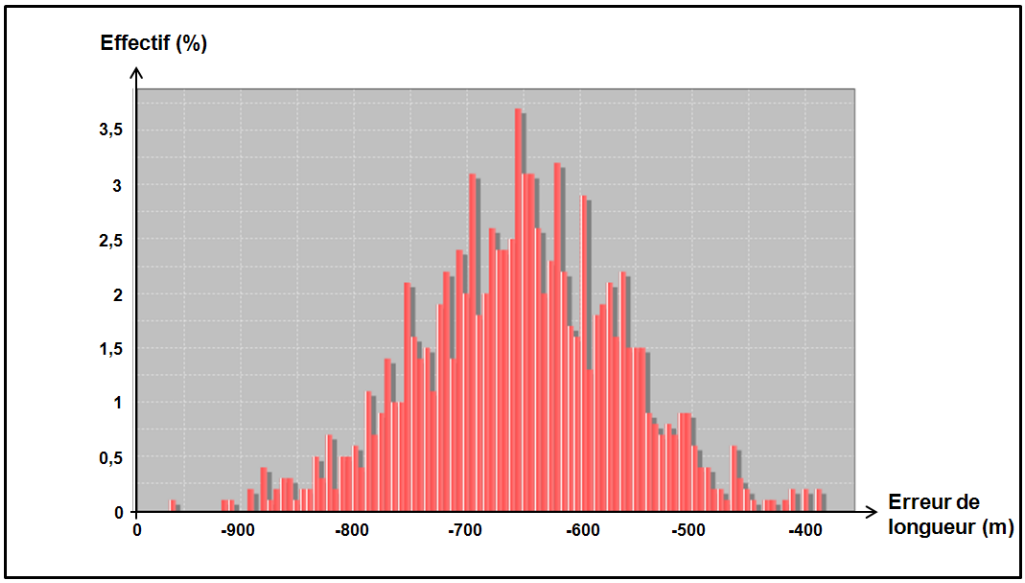


Figure 4.25 – Distribution des erreurs de longueur entre routes initiales et simulées

L'imprécision des mesures de longueur se caractérise par un écart-type σ_L de 89,33 mètres, comme le montre la distribution des erreurs de longueur de la figure 4.25. Ainsi, en prenant comme hypothèse que la distribution des erreurs suit une loi normale, nous pouvons estimer que l'imprécision de la mesure de longueur est de +/- $3\sigma_L$, soit +/- 267,9 mètres. Cette estimation est cohérente avec les erreurs maximale (-963,4 mètres) et minimale (-385,8 mètres) observées suite à la simulation.

Les résultats de l'application du modèle de simulation de l'erreur de pointé sur les géométries linéaires soulèvent donc des interrogations complémentaires quant au réalisme du bruitage proposé. En effet, bien que prévisible, la biais constaté dans les erreurs de longueur confirme que les contraintes de respect de forme, bien que peu importantes, doivent être enrichies de contraintes supplémentaires, afin de proposer une estimation non-biaisée de l'imprécision des mesures de longueur. Différentes pistes, prenant notamment en compte la courbure des objets, pourront être explorées. Ces nouvelles contributions ne seront cependant pas intégrées dans le périmètre de ce travail de thèse. La section suivante nous permettra de présenter la mise en application du modèle de simulation sur les mesures de surface.

4.1.5.2 Application sur des géométries surfaciques

L'application du modèle de simulation de l'erreur de pointé sur les géométries surfaciques n'est pas entreprise de la même manière selon les propriétés des objets évalués. Nous distinguons notamment les objets représentant une partition de l'espace de ceux n'en représentant pas. La mise en application du modèle de simulation de l'erreur de pointé proposée dans cette section ne sera réalisée que sur les objets représentant une partition de l'espace, en exploitant la classe d'objets des limites communales.

Données utilisées L'application du modèle a été réalisée sur les limites communales des cantons d'Hendaye et de Saint-Jean-de-Luz, également extraites de la BDCARTO (cf. figure 4.26). Les limites communales ainsi extraites couvrent 8 communes, pour une superficie totale de 133,43 km^2. En guise de comparaison, les mêmes limites communales extraites de la BDTOPO présentent une superficie totale de 132,73 km^2.

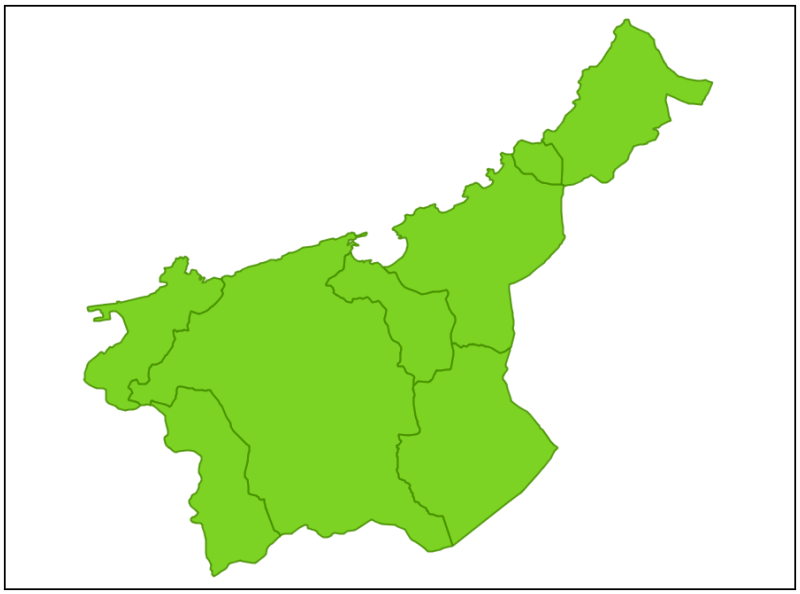


Figure 4.26 – Limites communales de la BDCARTO sur les cantons d'Hendaye et de Saint-Jean-de-Luz

Paramétrage du modèle Tout comme pour les tronçons de route, l'échelle caractéristique de saisie a été estimée par exploration des distances inter-points (cf. figure 4.27).

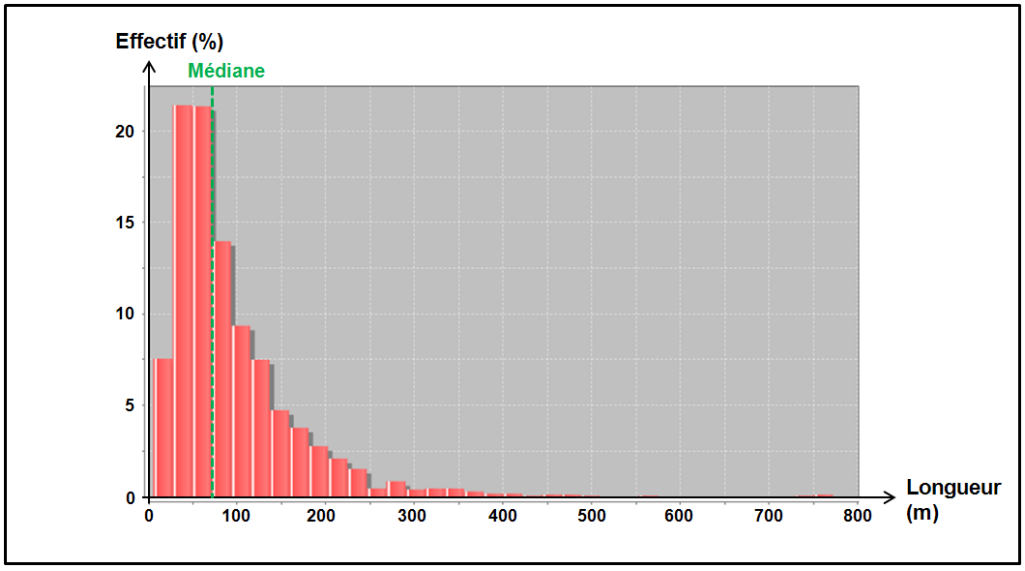


Figure 4.27 – Distribution des distances inter-points sur les limites communales de la BDCARTO

Pour les limites communales de la BDCARTO, la distance inter-points médiane observée est de 70,58 mètres. En nous basant sur la méthode d'estimation présentée en section 4.1.2.2, l'échelle de saisie moyenne a été estimée au 1 :35.288. En utilisant cette estimation de l'échelle de saisie, ainsi qu'une valeur $\sqrt{2\sigma_c} = 0{,}7$ mm conformément aux résultats de l'expérimentation, une valeur $\sigma_p = 4{,}97$ m sera donc utilisée pour paramétrer l'amplitude du bruit lors de la simulation. Egalement, la fonction $f(q) = 0.25\alpha + 0.75$ sera retenue pour pondérer l'amplitude du bruit gaussien en fonction de l'angularité entre points consécutifs. De le même manière, 1.000 réalisations ont été générées pour mener à bien la simulation.

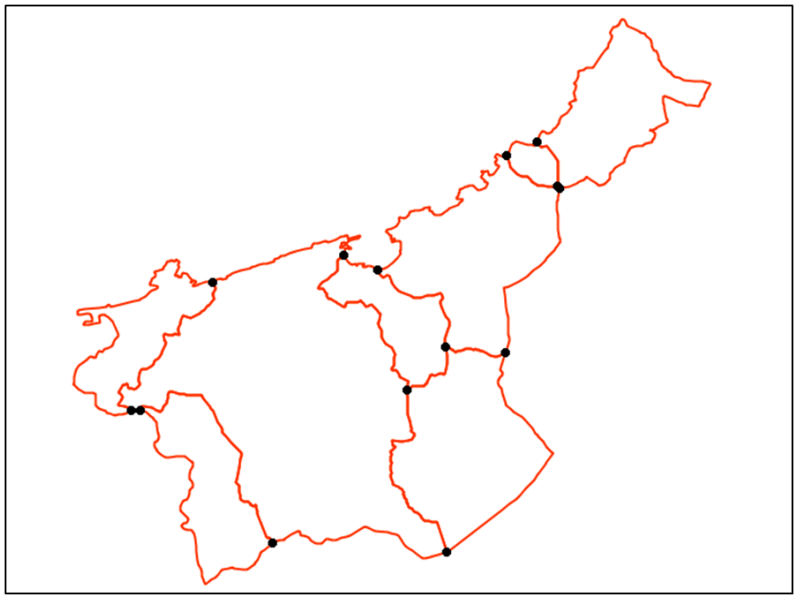


Figure 4.28 – Extraction des nœuds des limites communales

Méthodologie de simulation Les polygones simulés ont été générés en suivant une procédure identique à celle des objets linéaires :

1. L'ensemble des limites communales sont converties en carte topologique, afin d'extraire les nœuds du réseau (cf. figure 4.28),
2. Les polylignes correspondant aux limites communales localisées entre deux nœuds sont bruitées, comme pour les géométries linéaires (en intégrant les contraintes de respect de forme et de corrélation d'erreur),
3. Les polygones bruités sont reconstruits en extrayant chacune des faces du réseau bruité.

Résultats de la simulation Les résultats de la simulation de l'erreur de pointé sur les limites communales des cantons d'Hendaye et de Saint-Jean-de-Luz montrent tout d'abord que contrairement aux objets linaires, l'application d'un bruit sur la géométrie des objets n'a pas pour conséquence de sur-estimer les surfaces simulées. En effet, le graphique de la figure 4.29 montre bien que la distribution des erreurs de surface (correspondant à l'écart entre surface initiale et surface simulée) est centrée. L'écart-type σ_S des erreurs de surface est quant à lui estimé à 0,008 km^2.

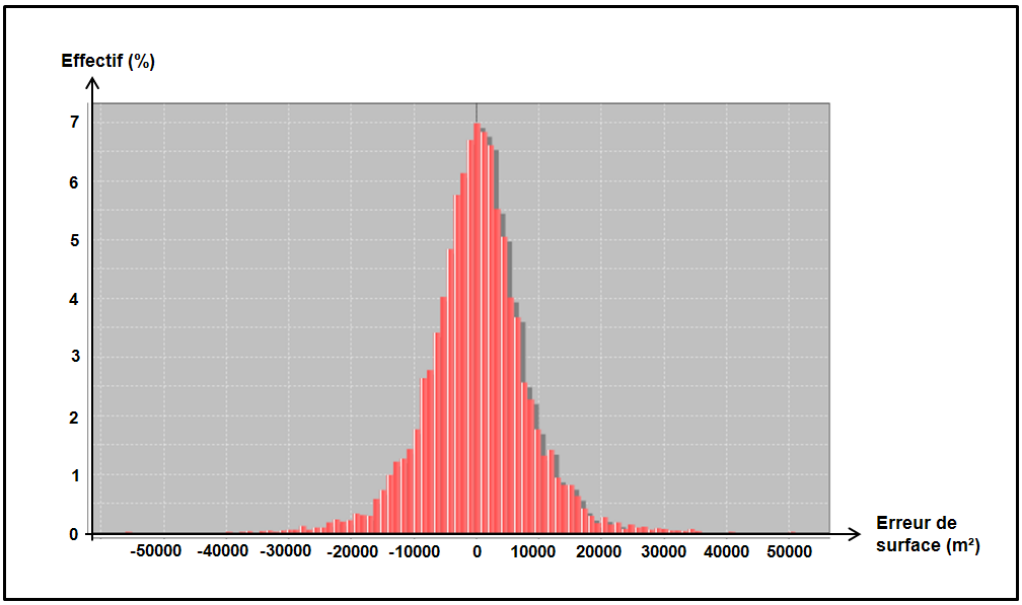


Figure 4.29 – Distribution des erreurs de surface entre limites communales initiales et simulées

En partant de l'hypothèse que la distribution des erreurs de surface suit une loi normale centrée, nous pouvons estimer l'imprécision de la mesure de surface à +/- $3\sigma_S$, soit +/- 0,024 km^2 (ou 2,4 hectares). Cette imprécision de la mesure de surface correspond ainsi à 0.018% de la surface initiale des communes des cantons d'Hendaye et de Saint-Jean-de-Luz.

L'application du modèle de simulation de l'erreur de pointé sur les limites communales des cantons d'Hendaye et de Saint-Jean-de-Luz offre la possibilité d'estimer globalement l'imprécision de la mesure géométrique de surface. Cependant, nous pouvons constater que l'imprécision de mesure ainsi produite ne permet pas d'approximer la mesure géométrique de surface de référence, calculée à partir des objets de la BDTOPO, qui est inférieure de 0,6 km^2, soit 60 hectares. Ces erreurs de surface importantes entre les deux bases de données sont impliquées par l'impact d'un autre processus de production, la généralisation cartographique, que nous étudierons dans la section suivante.

Conclusions de l'application du modèle de simulation L'application du modèle de simulation de l'impact de l'erreur de pointé que nous venons de mener dans cette section nous permet de tirer quelques conclusions quant à la sensibilité des mesures géométriques. Nous avons tout d'abord pu observer que les mesures de surface sont moins sensibles à l'imprécision de pointé que les mesures de longueur. En effet, suite aux simulations, nous avons observé une sur-estimation systématique des longueurs simulées (par un biais dans la distribution des erreurs), mais pas sur les surfaces simulées.
Cette observation demeure pourtant prévisible. Comme le montre la figure 4.30, l'intégration d'un bruit peut faire fortement varier une mesure de longueur, comme un périmètre par exemple, sans que la mesure de surface ne varie véritablement. Ceci explique pourquoi une sur-estimation systématique est observée sur les longueurs simulées.

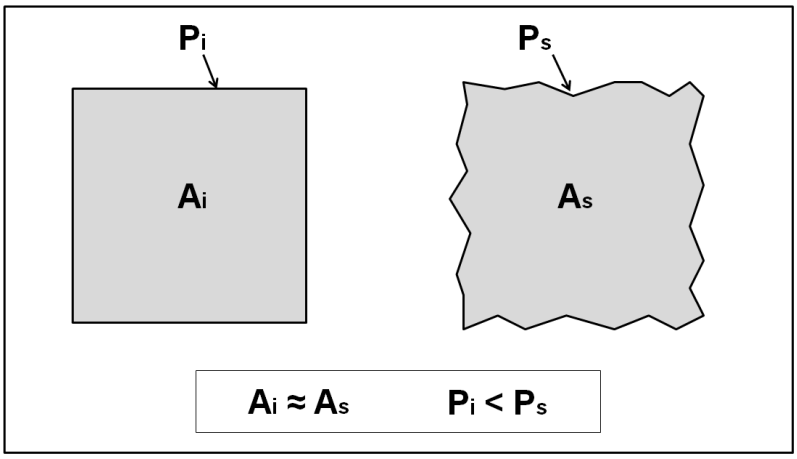


Figure 4.30 – Sur-estimation systématique du périmètre d'un objet surfacique simulé par rapport à sa surface.

Bien évidemment, cette sur-estimation demeure faible sur des jeux de données de faibles longueurs, mais lorsque la simulation de l'erreur de pointé est appliquée à l'ensemble d'un réseau routier (comme nous avons pu le voir dans cette section), la sur-estimation ainsi générée peut s'avérer importante.

L'intégration de contraintes de forme permet pourtant de réduire les exagérations liées au bruitage, qui ont pour conséquence d'augmenter la longueur des objets simulés. Cependant ces contraintes ne s'avèrent pas suffisantes pour éviter ce type d'effets et ainsi générer un bruitage véritablement réaliste, car non biaisé. Malgré ces imperfections, le modèle de simulation de l'erreur de pointé sera utilisé tel quel pour la suite de ce travail, en ayant conscience de ses limites, et des améliorations à y apporter.

Conclusion de la section

Cette section nous a permis de proposer un modèle d'estimation de l'impact de l'erreur de pointé de l'opérateur sur les mesures géométriques de longueur et de surface. Conformément à plusieurs contributions, une approche par simulations a été utilisée afin de modéliser l'erreur de pointé sur la géométrie des objets vectoriels. Les principes généraux de cette approche, basée sur un bruitage aléatoire de la géométrie des objets, ont tout d'abord été exposés, ainsi que les problèmes classiques auxquels elle est généralement confrontée.

Différentes méthodes permettant de réaliser une simulation réaliste de l'erreur de pointé sur les objets simulés ont ensuite été proposées. Ces propositions concernent tout d'abord le paramétrage de l'amplitude du bruit, qui est fonction de l'échelle moyenne de saisie des données. Une méthode d'estimation de l'échelle de saisie caractéristique (basée sur l'exploration des distances inter-points) a donc été proposée, afin d'accompagner l'utilisateur du modèle lorsque celui-ci ne la connait pas.
Pour préserver le réalisme des objets simulés, différents types de contraintes sont intégrés dans le modèle de simulation. Le premier type de contrainte concerne le respect de la forme initiale des objets. Nous avons pour cela proposé une méthode de pondération du bruit en fonction :
– de l'angularité inscrite entre points successifs de la géométrie évaluée,
– du degré des nœuds localisés aux extrémités d'une polyligne.
Le second type de contrainte est relatif à la corrélation des erreurs entre points successifs. Pour cela, une translation pondérée des points par rapport à leur distance curviligne aux extrémités a été proposée. L'application de ces contraintes dépend de la géométrie des objets évalués et de leurs caractéristiques.

Une expérimentation, basée sur la digitalisation de tronçons de route à différentes échelles, a ensuite été menée afin de calibrer le modèle de simulation. Les résultats de l'expérimentation montrent que pour paramétrer le bruit affecté aux objets, l'erreur moyenne quadratique, calculée par le biais d'un contrôle ponctuel, s'avère être une solution efficace. Egalement, l'expérimentation nous a permis de calibrer le modèle de simulation, en lui appliquant des contraintes relativement souples en termes de pondération du bruit par rapport à l'angularité entre points successifs.

Enfin, le modèle de simulation a été appliqué sur une extraction du réseau routier et des limites communales de la BDCARTO. Les résultats montrent que malgré les contraintes de respect de forme appliquées au modèle, les longueurs simulées sont sur-estimées, ce qui n'est pas le cas des surfaces. Ces résultats mettent donc en lumière les limites de ce modèle. En effet, si un biais n'est pas systématiquement observé lors de la simulation d'un objet linéaire, nous avons vu que ce n'est pas le cas lorsque la simulation est appliquée à l'ensemble d'un réseau routier.

Ces limites soulignent l'intérêt de prolonger les recherches sur la simulation de l'erreur de pointé, en intégrant d'autres types de contraintes, afin de modéliser son impact de la manière la plus réaliste possible. Malgré ces imperfections, le modèle d'estimation de l'impact de l'erreur de pointé sera intégré à ce stade d'avancement au sein du modèle général d'estimation de l'imprécision des mesures géométriques développé dans le cadre ce travail de thèse.

4.2 Généralisation cartographique

Le processus de généralisation cartographique peut être considéré comme un processus de production de base de données à part entière, bien qu'il soit généralement destiné à produire des cartes. En effet, le processus de généralisation cartographique est traditionnellement mis en œuvre lorsque l'on souhaite dériver une carte à partir d'une autre à plus grande échelle. Cependant, pour différentes raisons, il peut arriver que des objets géographiques soient saisis à partir de cartes. Dans ce cas, les objets géographiques saisis sont impactés par les effets de la généralisation cartographique. Ces effets se manifestent sur la géométrie des objets géographiques par modifications de leur forme et/ou de leur position, et par conséquent, les mesures de longueur et de surface calculées à partir de la géométrie des objets généralisés sont également affectées.
Cependant, peu d'études ont cherché à quantifier les effets de la généralisation cartographique sur les mesures géométriques. Cela peut se comprendre, puisqu'en théorie, il ne faudrait jamais créer une base de données à partir d'une carte. Pourtant cela peut être le cas, lorsqu'aucune autre source d'information n'est disponible, ou lorsque c'est la source d'information la plus facile à acquérir (pour des raisons de temps ou de moyens), voire parfois la seule solution envisageable (dans le cas des cartes anciennes).
Ainsi, dans le cadre de l'élaboration du modèle d'estimation, nous proposons dans cette section des méthodes permettant de quantifier l'impact de la généralisation cartographique sur les mesures géométriques de longueur ou de surface. Cependant, estimer l'impact de la généralisation cartographique peut s'avérer complexe, du fait de la diversité des opérations géométriques potentiellement mises en œuvre en fonction de la configuration spatiale des objets. De ce fait, il ne s'avère pas pertinent de modéliser les effets de la généralisation cartographique à l'aide d'approches par simulations. Nous privilégierons ici une approche par tests, basée sur des comparaisons, afin d'élaborer un modèle statistique d'estimation de l'imprécision des mesures géométriques.
Une présentation de l'impact de la généralisation cartographique et des objectifs de cette étude sera donc tout d'abord proposée, avant de présenter les méthodes mises en œuvre afin d'estimer l'impact de généralisation cartographique sur les mesures géométriques de longueur et de surface.

4.2.1 Les impacts de la généralisation cartographique

Préalablement à l'élaboration de méthodes d'estimation de l'impact de la généralisation cartographique sur les mesures géométriques, quelques généralités relatives aux impacts de la généralisation cartographique sont proposées. Cela nous permettra de présenter ensuite les objectifs de cette étude, ainsi que la méthodologie retenue.

4.2.1.1 Opérations de généralisation

La généralisation cartographique est définie par Cuenin (1972) comme "*l'opération qui, par sélection, schématisation et harmonisation, reconstitue sur une carte la réalité de la surface représentée dans ses traits essentiels en fonction du but de la carte, de son thème, de son échelle et des particularités de la région cartographiée*".
Comme l'explique Gaffuri (2008), les opérations de généralisation cartographique s'avèrent nécessaires lorsque l'on souhaite réaliser une carte à partir d'une autre dont l'échelle est plus grande. En effet, une simple réduction de la carte ne peut être réalisée du fait des contraintes de visibilité et d'interprétation à respecter.

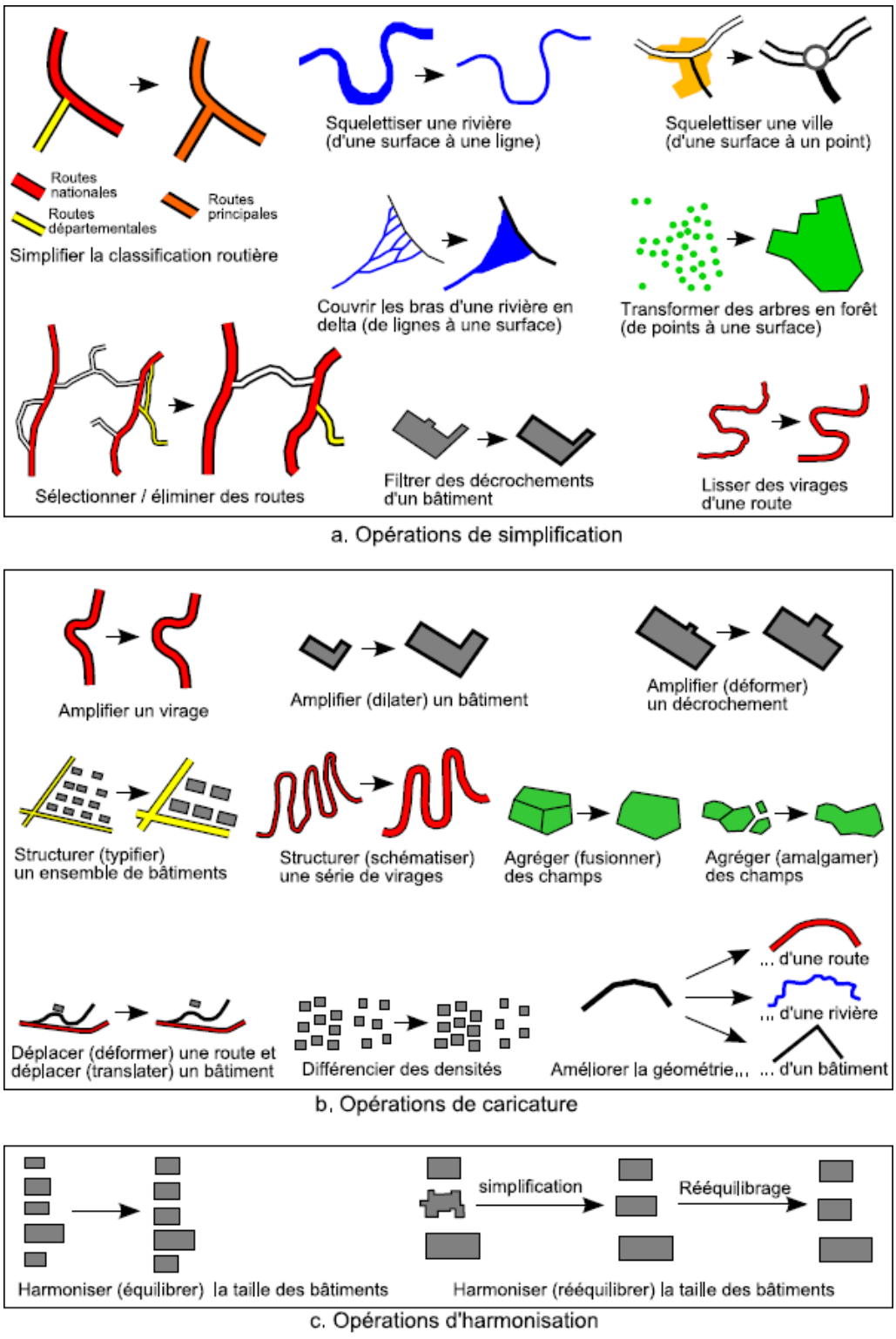


Figure 4.31 – Les opérations de généralisation, d'après Mustière (2001) et Gaffuri (2008)

L'opération de généralisation cartographique suppose donc une transformation de la carte initiale, et donc des objets qui la composent, afin d'être adaptée à l'échelle de la nouvelle carte. Depuis les années 80, les opérations de généralisation cartographique se sont peu à peu automatisées (voir par exemple Ruas (1999); Mustière (2001); Gaffuri (2008); Touya (2011)).

Opérations de généralisation De nombreuses classifications des opérations de généralisation ont été proposées (Gaffuri, 2008). Nous retiendrons la classification proposée par Mustière (2001), qui différencie trois grandes familles d'opérations (cf. figure 4.31) :
- les opérations de **simplification**, permettant de réduire la quantité d'information,
- les opérations de **caricature**, permettant de mettre en valeur l'information jugée importante,
- les opérations d'**harmonisation**, permettant d'éliminer les différences entre objets.

La figure 4.32 illustre l'impact des opérations de généralisation sur la représentation d'un même lieu :
- au 1 :25.000, à partir du SCAN25 (IGN, 2012d),
- au 1 :100.000, à partir de la Carte TOP100 (IGN, 2012c),
- au 1 :250.000, à partir de la Carte Régionale (IGN, 2012b).

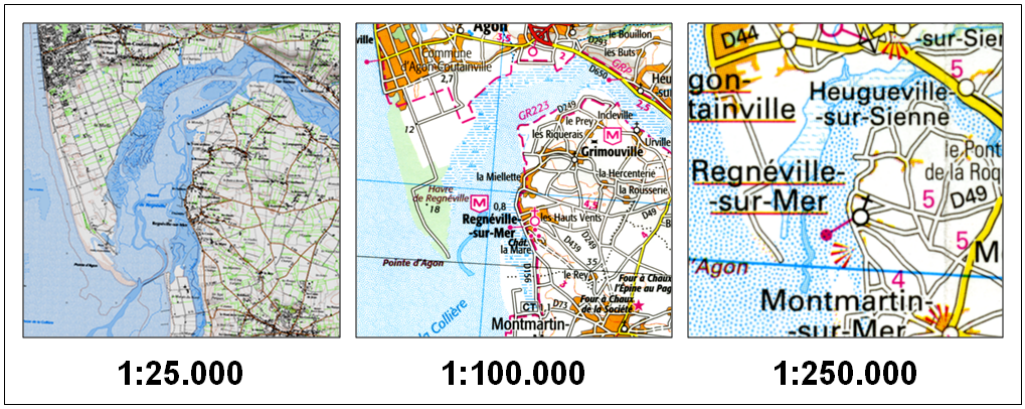


Figure 4.32 – Impact de la généralisation cartographique sur la représentation du Havre de Regnéville à trois échelles de représentation différentes

On s'aperçoit bien à travers cette illustration que l'impact de la généralisation cartographique se manifeste par de nombreuses modifications sur la géométrie des objets conservés. Ceci est particulièrement visible en ce qui concerne le réseau routier. Si jamais une carte est utilisée comme source pour saisir la géométrie des objets, ces derniers seront affectés par les effets de la généralisation cartographique, impactant également les mesures géométriques de longueur et de surface.

4.2.1.2 Impact sur les mesures géométriques

Les opérations de généralisation affectent la géométrie des objets vectoriels par modification de leur forme, leur taille, et leur position. En conséquence, les mesures de longueur et de surface calculées à partir de la géométrie des objets vectoriels (si ceux-ci ont été saisis à partir d'une carte ou d'une base de données généralisée) sont également affectées par les effets de la généralisation.

Pour illustrer cet impact de la généralisation cartographique sur les mesures géométriques, la figure 4.33 présente la rocade de Rennes, numérisée à partir de trois sources cartographiques différentes, à savoir la carte TOP100 au 1 :100.000 (IGN, 2012c), la Carte Régionale au 1 :250.000 (IGN, 2012b), et la Carte routière de la France au 1 :1.000.000 (IGN, 2012a).

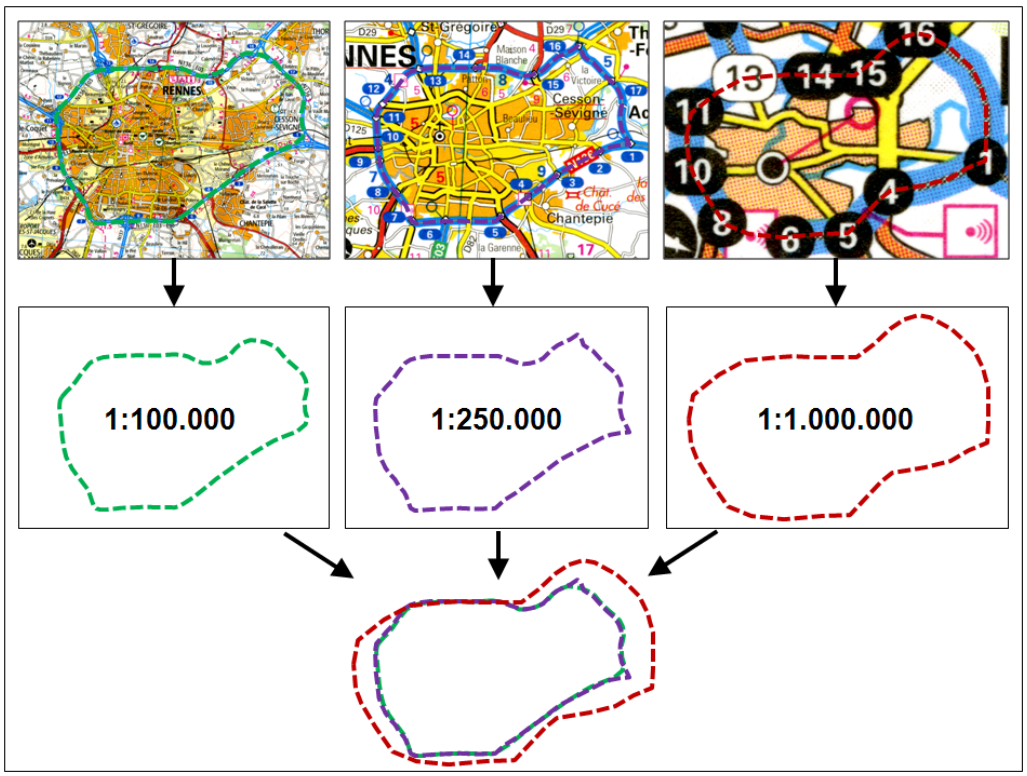


Figure 4.33 – La rocade de Rennes numérisée à partir de trois sources cartographiques

Nous voyons bien à travers cet exemple que la longueur mesurée à partir de ces trois objets, représentant pourtant la même entité du monde réel, n'est pas identique en fonction des opérations de généralisation mises en œuvre lors de la production de la source cartographique.

Différents auteurs ont ainsi étudié les effets de la généralisation cartographique sur les mesures géométriques, comme Blakemore (1984), João (1998), ou encore Bard (2004). L'une des principales contributions à l'étude de l'impact de la généralisation cartographique sur les mesures géométriques est celle de João (1998).

João (1998) a cherché à quantifier les effets de la généralisation cartographique sur les mesures de longueur, calculées à partir de différentes classes d'objets géographiques (routes, voies ferrées, rivières et limites administratives). L'auteur s'est pour cela appuyé sur des cartes britanniques et portugaises saisies à différentes échelles.

La figure 4.34 illustre l'impact de la généralisation manuelle sur les mesures de longueur calculées sur quatre classes d'objets extraites des cartes portugaises et britanniques. La tendance générale montre que la généralisation cartographique a pour effet de réduire les longueurs calculées à partir de la géométrie des objets.

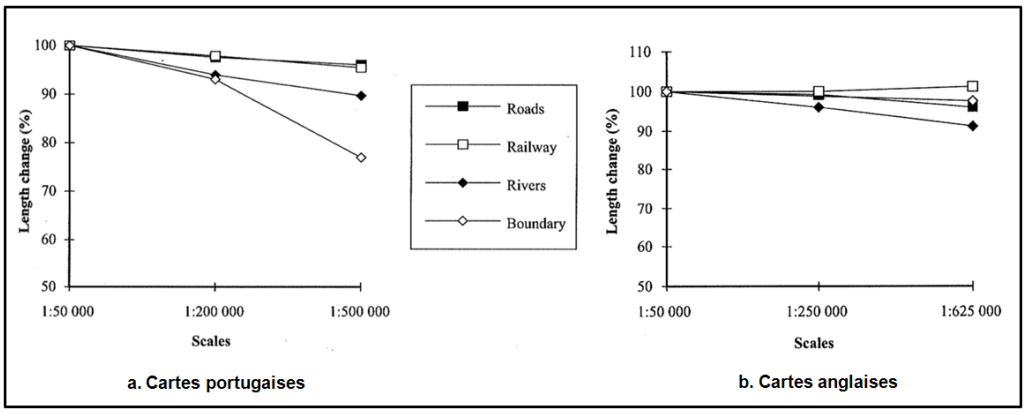


Figure 4.34 – Impact du changement d'échelle sur la longueur d'objets généralisés, d'après João (1998)

Cependant, João (1998) a dans un second temps focalisé cette étude sur des échantillons de routes (également extraits des cartes britanniques et portugaises) afin d'évaluer l'hétérogénéité des modifications de longueur au sein d'une même classe d'objets. Alors que la tendance générale tend davantage vers une réduction des longueurs calculées, l'auteur montre bien que dans certains cas, les effets de la généralisation manuelle sur les routes peuvent occasionnellement se caractériser par des augmentations des longueurs calculées (cf. figure 4.35). L'auteur cite à ce titre l'exemple de la rocade périphérique de Canterbury, pour laquelle la réduction d'échelle a eu pour effet une augmentation de longueur de 70 %. Selon la classification de Mustière (2001), cette route est affectée par une opération de caricature, générant une importante augmentation de longueur.

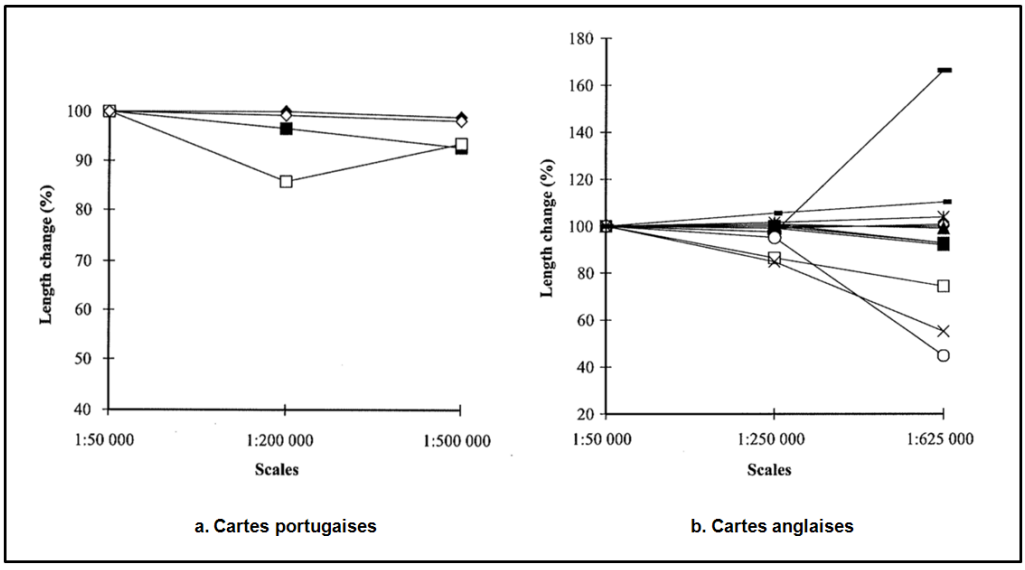


Figure 4.35 – Impact du changement d'échelle sur la longueur d'objets routiers généralisés, d'après João (1998)

Ces cas particuliers illustrent bien qu'en fonction des opérations de généralisation réalisées, les mesures géométriques peuvent être affectées positivement ou négativement. Les opérations de caricature (notamment celles d'amplification et de schématisation) peuvent tout particulièrement impliquer des modifications importantes en termes de mesures géométriques.

Enfin, João (1998) a cherché à comparer les effets de la généralisation manuelle et automatique sur les mesures de longueur des quatre classes d'objets géographiques (routes, voies ferrées, rivières et limites administrative). La généralisation automatique est ici basée sur l'application d'un filtrage par l'algorithme de Douglas et Peucker (1973). Les résultats montrent que la généralisation automatique préserve davantage les longueurs initiales (en passant d'une échelle à une plus petite) que la généralisation manuelle. Les déformations impliquées par la généralisation manuelle, plus importantes, assurent par contre une meilleure préservation des formes et du sens de la carte, par comparaison à l'algorithme de Douglas et Peucker (1973), qui a pour tendance de détruire les formes. On peut d'ailleurs regretter que cette étude n'ait été réalisée que sur un seul objet dans chaque classe, ce qui parait insuffisant pour en déterminer une loi générale, et que seul l'algorithme de Douglas et Peucker (1973), qui conserve la position des points non filtrés, ait été utilisé.

Ainsi, dans le cadre de ce travail de thèse, nous allons chercher à prolonger le travail initié par João (1998) sur l'impact de la généralisation cartographique sur les mesures de longueur, mais également de surface. A la différence de João (1998), nous ne cherchons pas à comparer les impacts entre généralisation manuelle et automatique, puisqu'aujourd'hui les processus de généralisation automatique sont majoritairement utilisés en production. L'étude proposée aura pour but de quantifier l'impact des processus de généralisation sur les mesures géométriques. Pour cela, des méthodes spécifiques de comparaison automatique seront mises en place.

4.2.1.3 Proposition de méthode d'estimation

Objectifs L'objectif de cette étude réside dans l'élaboration d'un modèle statistique permettant d'estimer l'impact de la généralisation cartographique sur les mesures géométriques de longueur et de surface. Pour cela, nous utilisons une approche basée sur des tests, puisque comme nous avons pu le voir précédemment, l'impact de la généralisation cartographique peut se manifester de manières très hétérogènes, ne rendant pas pertinente une approche par simulations. L'approche par tests repose sur des comparaisons préalables entre des bases de données ayant subi les effets de la généralisation (à différentes échelles) et une base de données de référence (censée représenter le terrain nominal).

En vue d'estimer l'impact de la généralisation cartographique sur les mesures géométriques nous chercherons à estimer l'erreur moyenne de mesure, ainsi que la dispersion des erreurs de mesure. En effet, comme l'a montré João (1998), l'impact de la généralisation cartographique se manifeste de manière générale par une sous-estimation des longueurs mesurées, mais dans certains cas, cet impact peut également se manifester par une sur-estimation des longueurs, du fait de certains processus mis en œuvre. Ainsi, l'erreur moyenne ne parait pas suffisante pour quantifier les effets de généralisation cartographique sur les mesures géométriques. Une étude de la dispersion de ces erreurs (par une variance ou un écart-type) s'avère donc nécessaire.

Hypothèses de travail João (1998) a déjà montré que l'impact de la généralisation cartographique sur les mesures de longueur varie selon l'échelle de représentation et la classe d'objets représentés. Cependant, cette étude ne différencie pas les effets de la généralisation cartographique en fonction du contexte géographique des objets. Or, comme le montre Touya (2011), différents processus de généralisation cartographique sont développés en fonction de l'espace géographique au sein duquel sont localisés les objets (nous utiliserons ici le terme de contexte géographique).

En effet, ces dernières années, des modèles de généralisation automatique ont été spécifiquement développés pour les zones urbaines (Ruas, 1999), rurales (Duchêne, 2002), ou de montagne (Mustière (2001). On peut de ce fait considérer que l'impact de la généralisation cartographique est hétérogène, du fait des processus mis en œuvre en fonction du contexte géographique des objets, et que cela se traduit également sur les mesures géométriques calculées en aval.

Ainsi, nous pouvons supposer que les erreurs de mesures géométriques calculées à partir d'objets généralisés varient en fonction de trois paramètres :
– l'échelle de représentation,
– la classe d'objets représentée,
– le contexte géographique des objets.
Nous chercherons donc à élaborer une fonction permettant de quantifier l'erreur de mesure selon ces trois paramètres, comme le montre l'équation 4.7.

$$Erreur = f(Echelle, Classe, Contexte) \tag{4.7}$$

Nous ne nous limiterons dans le cadre de ce travail, qu'à trois contextes géographiques, à savoir les zones urbaines, rurales et de montagne. En effet, les processus de généralisation sur les objets localisés dans les zones urbaines et de montagne impliquent des modifications géométriques importantes, dont nous chercherons en priorité à évaluer l'impact sur les mesures. D'autres délimitations impliquant des processus de généralisation spécifiques (comme les zones côtières ou les zones péri-urbaines) ont été proposées par Touya (2011), mais ne seront pas prises en compte dans le cadre de ce travail.

Les comparaisons entre les bases de données généralisées et la base de données de référence seront donc réalisées successivement dans chaque contexte géographique. L'objectif est d'évaluer si les erreurs de longueur affectant les objets généralisés varient bien en fonction du contexte géographique. Cependant, cette différenciation par contextes géographiques ne sera appliquée que pour l'étude des erreurs de longueur. Pour l'étude portant sur les erreurs de surface, nous considèrerons que la répartition des erreurs est homogène sur l'ensemble de la zone étudiée, et donc aucune différenciation de contexte géographique ne sera appliquée dans le cadre des comparaisons.

Mise en œuvre Cette étude sur l'estimation de l'impact de la généralisation cartographique sur les mesures géométriques de longueur et de surface se décline en deux temps, comme illustré dans la figure 4.36 :

1. Elaboration du modèle statistique par comparaisons,
2. Application du modèle statistique sur des objets à évaluer.

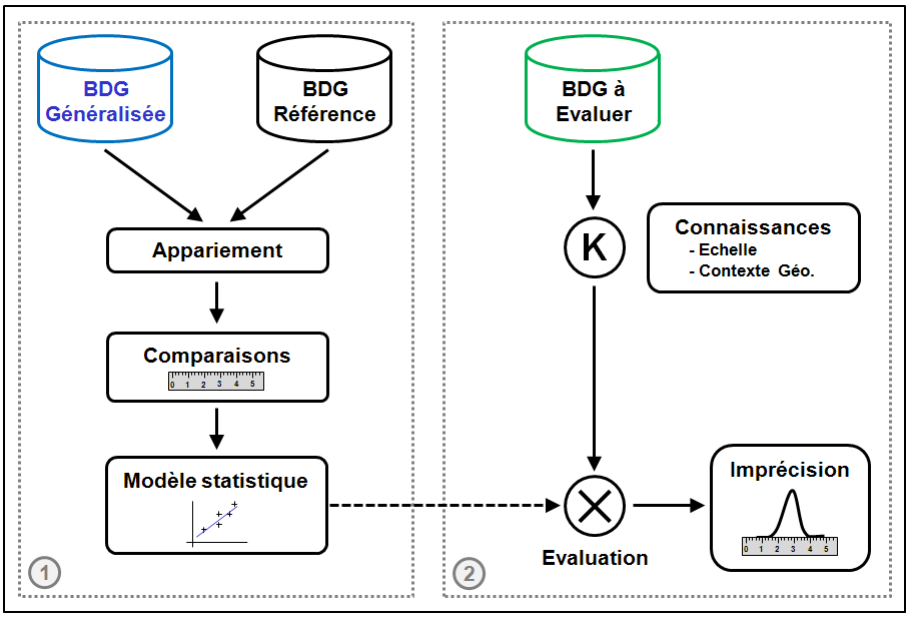


Figure 4.36 – Principe d'élaboration d'un modèle statistique (1) et application de ce modèle sur une base de données à évaluer (2)

L'élaboration du modèle statistique se base, comme nous l'avons expliqué, sur des comparaisons préalables entre données généralisées (à différentes échelles) et une base de données de référence, selon un ou plusieurs contextes géographiques. La méthodologie de comparaison est élaborée afin d'être la plus automatisée possible, dans le but de quantifier l'erreur de mesure sur de grandes quantités d'objets, et ainsi produire un modèle statistique représentatif. Pour cela, des méthodes d'appariement automatique seront mobilisées.
Une fois le modèle statistique élaboré, nous chercherons à l'appliquer sur une base de données dont on essaie d'estimer l'impact du processus de généralisation cartographique sur les mesures géométriques. L'application du modèle statistique sera tout d'abord réalisée pour sa validation, puis dans le cadre de la mise en œuvre du modèle général d'estimation.

Avant d'étudier l'impact de la généralisation cartographique sur les mesures de longueur ou de surface, il est préalablement nécessaire de développer des méthodes d'acquisition des connaissances afin d'estimer l'échelle de représentation de la carte source, ou encore de délimiter le contexte géographique des objets. Ces méthodes, indispensables au processus de comparaison, puis au paramétrage du modèle statistique, seront présentées dans la section suivante.

Synthèse Section Nous avons vu dans cette section que pour certaines raisons, la géométrie des objets d'une base de données peut être saisie à partir de cartes. Par conséquent, celle-ci est affectée par les effets de la généralisation cartographique, qui se répercutent sur les mesures géométriques de longueur ou de surface. Afin d'estimer son impact sur les mesures géométriques, une approche par tests est proposée. L'objectif de cette approche est d'élaborer un modèle statistique d'estimation de l'impact de la généralisation cartographique, à partir de comparaisons réalisées entre données généralisées et données de référence.

4.2.2 Acquisition de connaissances pour le paramétrage du modèle

En considérant que l'impact de la généralisation cartographique varie en fonction de l'échelle de représentation des données sources (lorsqu'il s'agit d'une carte) et du contexte géographique des objets évalués, il est nécessaire de disposer de ces informations afin de paramétrer le modèle d'estimation. Plusieurs cas de figure peuvent être envisagés : soit l'utilisateur dispose de ces informations, et est donc capable de paramétrer le modèle, soit il n'en dispose pas, et dans ce cas des méthodes spécifiques doivent être développées afin de l'accompagner. Nous allons donc proposer dans cette section des méthodes permettant d'estimer l'échelle de représentation des données sources, et de délimiter le contexte géographique des objets évalués.

4.2.2.1 Estimation de l'échelle de représentation des données sources

Nous avons évoqué dans la section 4.1.2.2 les concepts d'échelles de saisie et de représentation des bases de données géographiques. Nous avons vu qu'au contraire des cartes, les bases de données n'ont pas d'échelle de représentation fixe, mais sont utilisables dans une gamme d'échelle.

Echelle de représentation caractéristique Dans le cas des données impactées par la généralisation cartographique, le concept d'échelle de représentation caractéristique prend toute son importance. En effet, les bases de données cartographiques sont, par définition, des "*bases de données qui regroupent des données géographiques dont la position de chaque objet est proche de la précision géométrique de saisie, mais dont la taille et la position de chaque objet est conforme à l'échelle mathématique de représentation*" (Ruas, 2002). A travers cette définition, nous voyons que la géométrie des objets est impactée par les contraintes de représentation cartographique à une échelle donnée (l'échelle de représentation caractéristique). L'application d'un symbole sur la géométrie des objets constitue une contrainte de représentation, puisque les objets symbolisés doivent demeurer lisibles à l'échelle de représentation souhaitée.

Exploitation de la symbologie associée aux objets Pour estimer l'échelle de représentation caractéristique d'une base de données cartographiques, ou bien l'échelle de la carte ayant servi de source à la saisie d'objets, nous proposons de prendre en compte la taille de symbole associée aux objets de la base de données.
En effet, pour que les objets d'une base de données cartographiques soient rendus lisibles à l'échelle de représentation désirée, il est nécessaire de leur associer une symbologie adaptée. La taille du symbole varie en fonction de la classe d'objets et de leur classification. Par exemple, dans la Carte Régionale au 1 :250.000 de l'IGN, la largeur de symbole des autoroutes est de 1,5 millimètres, tandis qu'elle n'est que de 1 millimètre pour les routes régionales. L'emprise des objets symbolisés sur la carte devient beaucoup plus importante que sur le terrain. Dans l'exemple de la Carte Régionale, la largeur d'une autoroute équivaut à 375 mètres terrain, et celle d'une route régionale à 250 mètres.

L'application d'un symbole sur les objets de la base de données génère donc des conflits de lisibilité, qui sont résolus par l'application d'opérations de généralisation, qui modifient la géométrie des objets (par simplification, agrandissement, déplacement) ou les éliminent. Ainsi, pour un problème spécifique de représentation, comme la superposition de deux

symboles de routes, des opérations spécifiques de généralisation sont réalisées : déplacement des objets routiers, suppression de virages, élargissement de virages...
Dans ce contexte, la connaissance des opérations de généralisation cartographique peut nous aider à estimer l'échelle de représentation des objets d'une base de données cartographiques, lorsque la taille de symbole est connue. Nous allons donc proposer dans cette section deux méthodes originales permettant d'estimer l'échelle de représentation d'une base de données cartographiques par connaissance des opérations de généralisation associées, à travers l'exemple de la représentation des tronçons de route.

a. Superposition de symboles de routes voisines

La première méthode que nous proposons fait référence à un problème classique de généralisation cartographique : la représentation de routes voisines, et en particulier de routes parallèles. Afin d'assurer la lisibilité de la carte, en généralisation manuelle ou automatique, il est préconisé de conserver une distance minimale entre les deux routes symbolisées (Cuenin, 1972). La figure 4.37 illustre cet exemple par deux routes parallèles représentées sur la carte TOP100. On observe qu'un espace vide est conservé entre les symboles des routes afin d'assurer leur visibilité, bien que l'emprise des symboles en unités terrain soit beaucoup plus importante que dans la réalité.

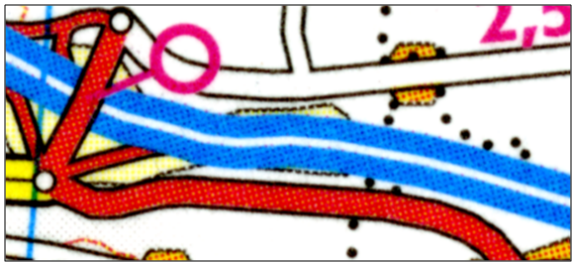


Figure 4.37 – Deux routes voisines représentées sur la carte TOP100 (n°106)

Hypothèses La distance minimale mesurée entre les axes de deux routes voisines est fonction de :
– la taille du symbole et de l'espace vide séparant les deux routes,
– l'échelle de représentation.
L'équation 4.8 formalise l'expression de cette distance minimale entre les axes de deux routes voisines, où D_{Min} est la distance minimale (en mètres) entre les axes de deux routes voisines, L_{DS} est la largeur du demi-symbole de chaque route, L_G est la largeur de l'espace vide entre les deux routes, et $1/E$ est l'échelle de représentation de la carte.

$$D_{Min} = \frac{2 * L_{DS} + L_G}{E} \tag{4.8}$$

A l'aide de cette équation, il est possible d'estimer l'échelle de représentation des objets de la base lorsque :
– la distance minimale entre les axes de deux routes voisines est connue,
– la taille de symbole et de l'espace vide est spécifiée.

Ainsi, pour estimer l'échelle de représentation d'une base de données à évaluer, nous proposons d'affecter une symbologie type aux routes étudiées, et d'analyser la largeur de l'espace vide entre les routes parallèles symbolisées. En réduisant l'échelle de représentation, la taille des symboles augmente, et l'espace vide entre les deux routes symbolisées se réduit. Nous pouvons considérer que lorsque les deux symboles simulés se superposent, l'échelle de représentation est approximée.

Symbologie Pour estimer l'échelle de représentation d'un réseau routier d'une base de données cartographiques, les spécifications de symbolisation de la carte TOP100 (IGN-SDC, 2007) au 1 :100.000 et de la Carte Régionale (IGN-SDC, 2004) au 1 :250.000 ont été utilisées, comme exposé dans la table 4.6.

	Echelle	
Type de route	**1 :100.000**	**1 :250.000**
Autoroute	1,5 mm	1,5 mm
Route large (principale ou régionale)	0,92 mm	1 mm
Route étroite (principale ou régionale)	0,69 mm	0,75 mm
Chemin	0,25 mm	0,25 mm

Table 4.6 – Spécifications simplifiées des largeurs de symboles routiers pour la carte TOP100 et la Carte Régionale

Algorithme L'algorithme développé se base sur les tailles de symboles spécifiées et sur une échelle de représentation incrémentale. La figure 4.38 illustre les différentes étapes de l'estimation de l'échelle de représentation.

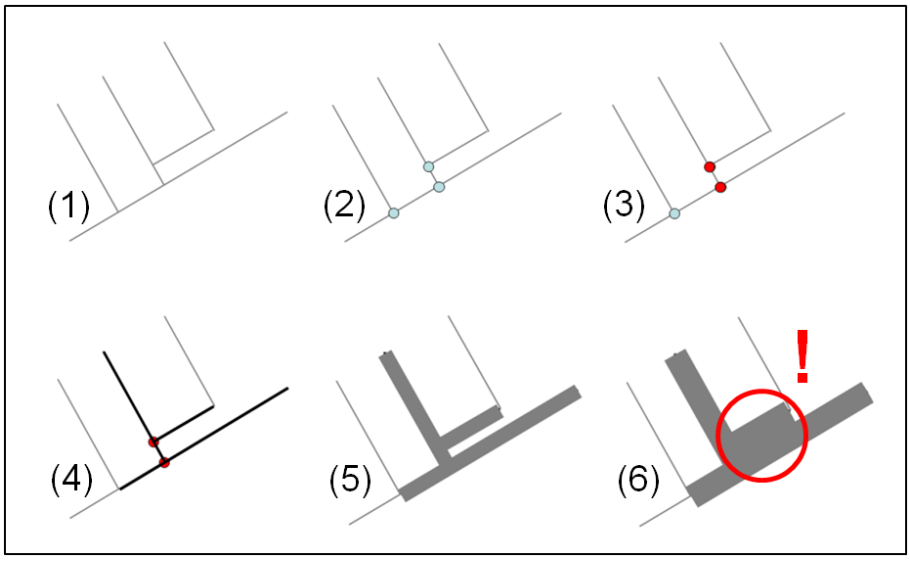


Figure 4.38 – Algorithme d'estimation de l'échelle de représentation basé sur la superposition de symboles de routes voisines

L'algorithme cherche dans un premier temps à détecter les routes voisines, pouvant potentiellement conserver une distance vide entre deux symboles. Pour cela :
- le réseau est transformé en graphe topologique (1),
- chaque intersection (c'est-à-dire chaque nœud connecté à trois arcs ou plus) est sélectionnée (2),
- les intersections proches (en dessous d'une valeur seuil, 400 mètres par exemple) sont ensuite sélectionnées (3),
- les arcs connectés aux intersections proches sont sélectionnés (4).

Dans un second temps, une symbologie est affectée aux routes sélectionnées, et l'algorithme recherche celles où les symboles s'intersectent. Pour cela :
- une zone tampon (de largeur relative à la taille de symbole et à l'échelle courante) est calculée sur les tronçons sélectionnés (5),
- si un nombre toléré de superpositions est atteint l'algorithme s'arrête (6), sinon il reprend à 5 avec une échelle incrémentée.

Le pseudo code suivant formalise l'algorithme d'estimation de l'échelle de représentation par la superposition des symboles de routes voisines, où *scale* est l'échelle de représentation courante, *buffer* est la distance de la zone tampon appliquée sur les routes sélectionnées, *symbolSize* est la taille du symbole en millimètres, *overlapNumber* est le nombre de superpositions de symboles, et t est le nombre toléré de superpositions.

```
1  scale = 10000
2  while overlapNumber < t
3      buffer = 0,5 * symbolSize * scale
4      for each buffer :dbbuffer
5          if buffer.isIntersected() = true then
6              overlapNumber = overlapNumber + 1
7          end if
8      end for
9      if overlapNumber >= t then
10         return scale
11     else
12         scale = scale + 10.000
13     end if
14 end while
```

Expérimentation Pour valider la méthode proposée, une expérimentation est réalisée sur deux extractions du réseau routier des données TOP100 et de la Carte Régionale, dont les échelles de représentation sont respectivement le 1 :100.000 et le 1 :250.000. L'expérimentation a été réalisée avec une échelle initiale au 1 :10.000, incrémentée de 10.000 à chaque itération. A chaque échelle de représentation testée, la surface de symboles intersectés est calculée.

Le graphique de la figure 4.39 présente les résultats de l'expérimentation sur une extraction du réseau routier des données TOP100. Ce graphique montre que même si quelques intersections sont détectées à des échelles supérieures au 1 :100.000, au delà de cette échelle, la surface intersectée de symboles de routes voisines augmente significativement.

Une telle augmentation de la surface intersectée confirme que la symbologie utilisée n'est plus adaptée à des échelles inférieures au 1 :100.000, qui est l'échelle de représentation des données TOP100. Les résultats de cette expérimentation s'avèrent donc très encourageants et montrent que cette méthode offre la possibilité d'approximer l'échelle de représentation, lorsque la surface intersectée augmente fortement.

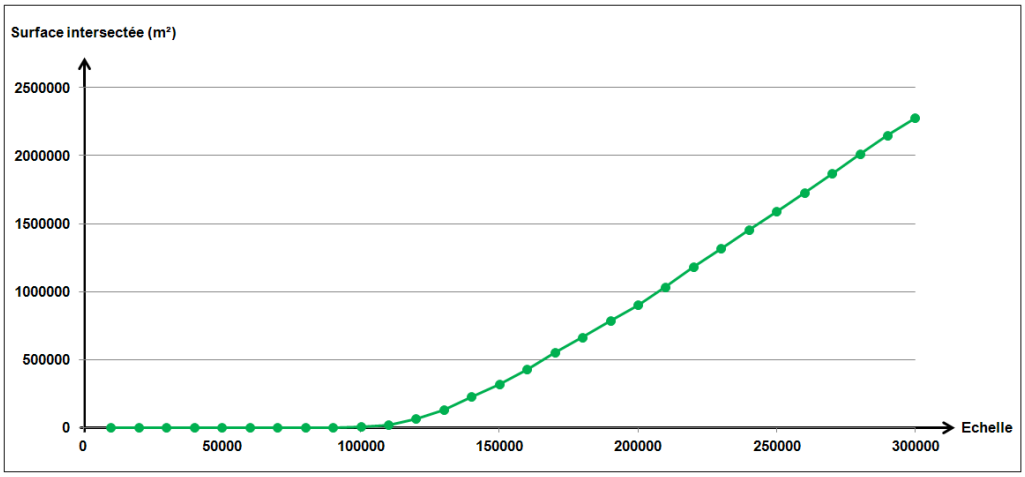


Figure 4.39 – Surface intersectée entre symboles de routes voisines en fonction de l'échelle de représentation du réseau routier de données TOP100

Le graphique de la figure 4.40 présente quant à lui les résultats de l'expérimentation sur une extraction identique du réseau routier des données de la Carte Régionale. Contrairement aux résultats de l'expérimentation sur les données TOP100, le graphique montre que des intersections de symboles entre routes voisines sont détectées à de plus grandes échelles (dés le 1 :140.000) que l'échelle de représentation des données (au 1 :250.000).

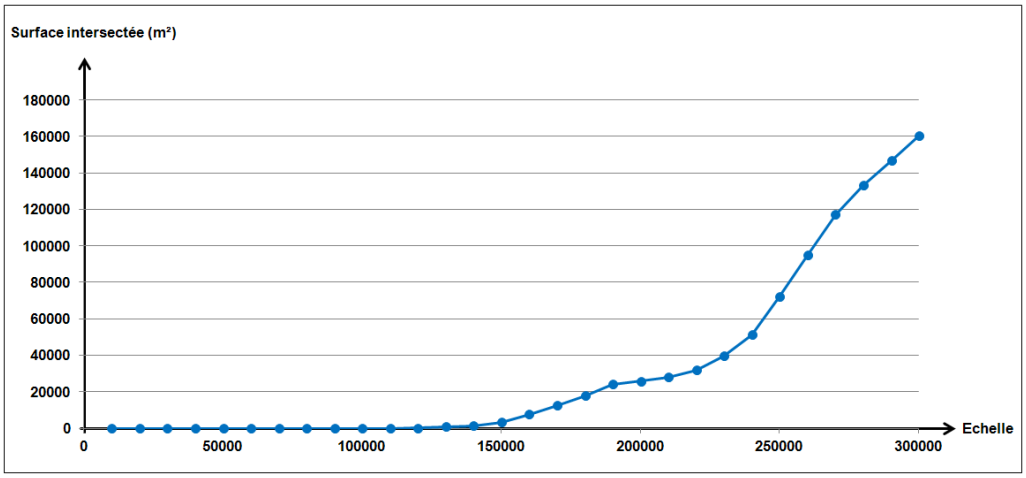


Figure 4.40 – Surface intersectée entre symboles de routes voisines en fonction de l'échelle de représentation du réseau routier de la carte Régionale

Par exemple, la figure 4.41 illustre un cas de superposition de symboles routiers détecté au 1 :150.000 sur les données de la Carte Régionale, et présent sur la carte elle-même. Le graphique de la figure 4.40 montre cependant qu'à partir du 1 :220.000, c'est-à-dire en approchant l'échelle de représentation des données, la surface intersectée par les symboles routiers se superposant augmente fortement.

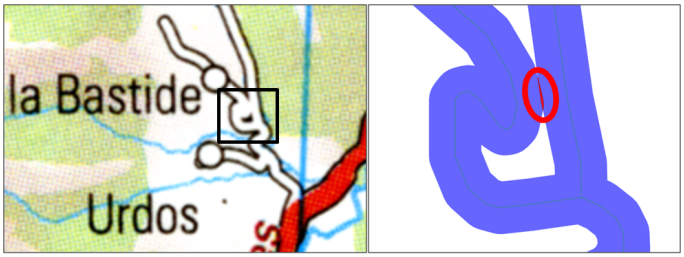


Figure 4.41 – Superposition de symboles routiers détectée sur les données de la Carte Régionale au 1 :150.000

Nous voyons donc bien que cette méthode basée sur l'intersection de symboles entre routes voisines permet assez facilement d'approximer l'échelle de représentation des données lorsque la surface intersectée augmente fortement. L'expérimentation réalisée sur l'extraction du réseau routier des données TOP100 le montre très clairement. Ceci est par contre nettement moins significatif sur les données de la Carte Régionale.

Cependant cette méthode n'est véritablement effective que pour le cas d'un réseau routier dense, comme en milieu urbain par exemple. En zones de montagne ou en milieu rural, si les axes de routes ne sont pas suffisamment proches les uns des autres, cette méthode devient rapidement inadaptée. Ainsi, dans le cas des routes de montagne, une seconde méthode est proposée, basée sur les problèmes d'empâtement de symboles de virages.

b. Empâtement des symboles de virages

Le problème d'empâtement des symboles de virages est également un problème classique de généralisation cartographique. Comme l'a montré Mustière (2001), lorsque des virages sont trop serrés, un empâtement du symbole de route se produit, empêchant une bonne visualisation des virages sur la carte (cf. figure 4.42).

Figure 4.42 – Empâtement d'une ligne symbolisée, d'après Mustière (2001)

Différents processus de généralisation sont proposés afin de permettre une bonne représentation des virages, notamment la suppression ou l'élargissement. Pour détecter à quel moment il est nécessaire d'activer ces processus de généralisation, Mustière (2001) propose une méthode basée sur l'empâtement maximal toléré d'un symbole routier.

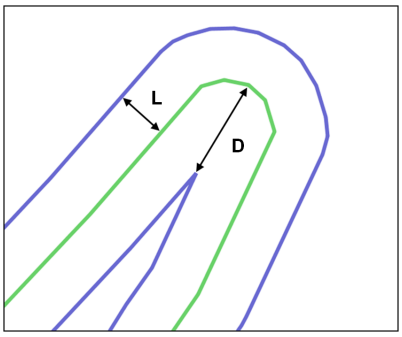


Figure 4.43 – Rapport entre la distance bord-axe de la route D et la largeur du demi-symbole L

Comme le montre la figure 4.43, il est possible de qualifier l'empâtement d'un symbole de route en exploitant le rapport entre la distance D bord-axe de la route avec la largeur du demi-symbole de route L. Selon Mustière (2001), lorsque ce rapport est supérieur à 1.7, le symbole est considéré comme empâté, et un processus de généralisation doit être activé.

L'équation 4.9 permet de formaliser cette méthode de détection de l'empâtement d'un symbole de route, où D est la distance (en mètres) entre l'axe de la route et l'extrémité du symbole, L la largeur (en millimètres) du demi-symbole de route, et $1/E$ l'échelle de représentation.

$$D > \frac{L * 1.7}{E} \tag{4.9}$$

Hypothèses Nous pouvons supposer que la distance maximale tolérée entre l'axe de la route et l'extrémité du symbole correspond à celle de l'équation 4.9. Cette hypothèse nous permet d'estimer l'échelle de représentation de la base de données cartographiques, à partir du moment où sont connues :
– la distance entre l'axe de la route et l'extrémité du symbole,
– la largeur du demi-symbole de route.
Nous allons utiliser cette hypothèse afin de développer une méthode permettant d'estimer l'échelle de représentation des routes d'une base de données cartographiques.

Algorithme Un algorithme, basé sur la taille de symbole et l'échelle de représentation, a donc été développé. Comme dans la méthode de détection de superpositions de symboles proposée précédemment, une zone tampon est calculée pour chaque tronçon de route, conformément aux spécifications des données TOP100 et de la Carte Régionale, présentées en table 4.6. A chaque échelle expérimentée, l'algorithme calcule le nombre d'empâtements détectés, jusqu'à ce qu'un nombre d'empâtements toléré soit atteint.

Le pseudo code ci-dessous présente cet algorithme, où *scale* est l'échelle de représentation courante, *buffer* est la distance de la zone tampon appliquée sur les routes, *symbolSize* est la taille du symbole en millimètres, *distance* est la distance bord-axe en mètres, *coalescenceNumber* est le nombre d'empâtements détecté, et t est le nombre toléré d'empâtements.

```
1  scale = 10.000
2  while coalescenceNumber < t
3      buffer = 0,5 * symbolSize * scale
4      for each lineString :dbroads
5          for each pt :lineString
6              if distance > 1,7 * buffer then
7                  coalescenceNumber = coalescenceNumber + 1
8              end if
9          end for
10     end for
11     if coalescenceNumber >= t then
12         return scale
13     else
14         scale = scale + 10.000
15     end if
16 end while
```

Expérimentation Une expérimentation a été réalisée sur le même échantillon de tronçons de route que dans la méthode de détection de superpositions de symboles routiers, à partir des données TOP100 et de la Carte Régionale. De la même manière, l'échelle initiale de représentation testée est le 1 :10.000, incrémentée de 10.000 à chaque itération.

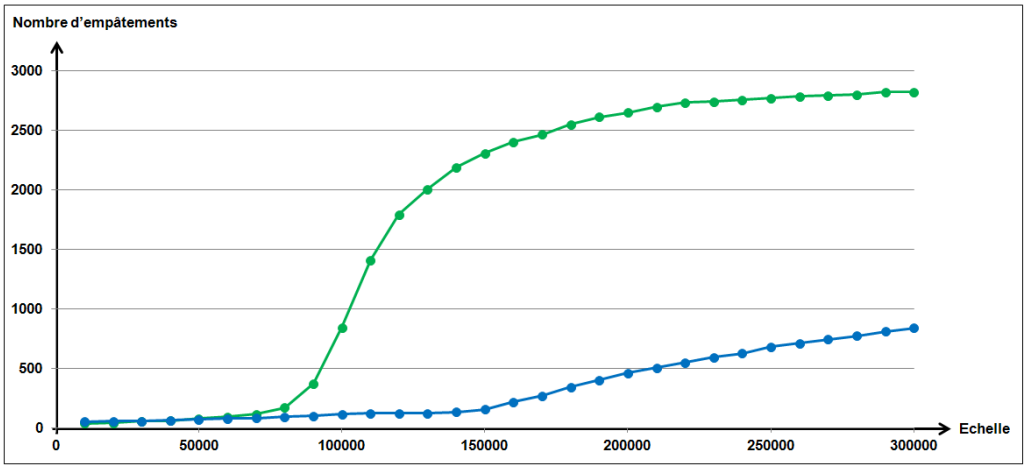


Figure 4.44 – Nombre d'empâtements détectés en fonction de l'échelle de représentation sur les tronçons de route extraits des données TOP100 (en vert) et de la Carte Régionale (en bleu)

Comme illustré dans le graphique de la figure 4.44, les résultats montrent une augmentation significative du nombre d'empâtements détectés à partir de l'échelle 1 :100.000 pour les données TOP100, ce qui confirme bien l'intérêt de la méthode.

La figure 4.45 illustre la symbolisation d'un virage extrait des données TOP100 à différentes échelles, en comparaison avec sa représentation sur une carte. On observe qu'à partir du 1 :100.000, la distance bord-axe (représentée avec un trait bleu épais) devient trop importante, ce qui laisse supposer que l'échelle de représentation est atteinte.

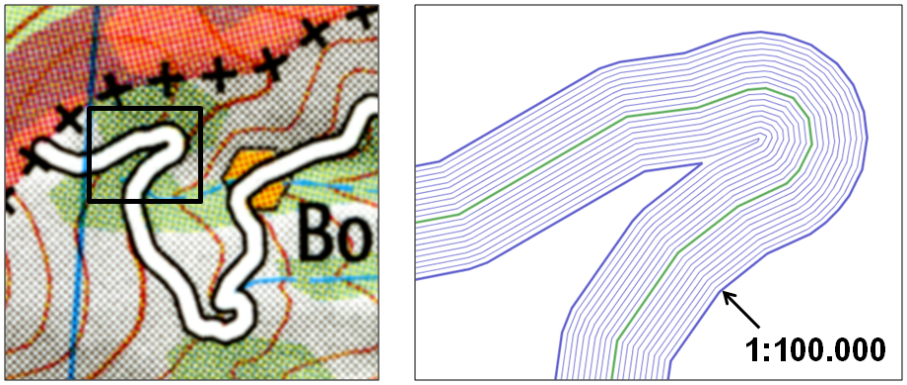


Figure 4.45 – Limite d'empâtement d'un virage extrait des données TOP100, en comparaison avec sa représentation cartographique

Cependant, pour les données de la Carte Régionale, les résultats sont moins significatifs, puisque le nombre d'empâtements de symboles routiers augmente dés l'échelle 1 :150.000. On peut penser à ce titre que l'impact de la généralisation de modèle, qui implique la suppression de nombreuses petites routes, est responsable du faible taux de détection d'empâtements. En effet, la majorité des routes concernées par ces cas d'empâtements de symboles de virages sont des petites routes, au caractère sinueux, dont beaucoup sont éliminées lors de la généralisation des données.

c. Combinaison des méthodes pour le paramétrage du modèle

Les deux méthodes proposées offrent des résultats encourageants pour estimer l'échelle de représentation de bases de données cartographiques. Ces méthodes ne concernent pour l'instant que les tronçons de route. Nous allons à présent voir comment les combiner afin de paramétrer le modèle d'estimation de l'impact de la généralisation cartographique.

Hypothèse Nous avons pu constater que la surface de symboles intersectés (pour la première méthode) et le nombre d'empâtements détectés (pour la seconde méthode) augmentent fortement lorsque l'échelle de représentation caractéristique est approchée. Nous avons donc décidé d'exploiter le taux de variation de ces valeurs à chaque changement d'échelle par l'algorithme, pour estimer l'échelle de représentation caractéristique.

En partant de l'hypothèse que les taux de variation (de la surface intersectée entre routes voisines et du nombre d'empâtements détectés) atteignent leurs maximums lorsque l'échelle de représentation caractéristique est approchée, nous proposons d'estimer une gamme d'échelles de représentation comprise entre ces taux de variation maximaux.

Expérimentation Les graphiques des figures 4.46 et 4.48 présentent (pour les données TOP100 et de la Carte Régionale), à chaque saut d'échelle, les taux de variation :
– de la surface intersectée par des symboles de routes voisines,
– du nombre d'empâtements de virages détectés.

Pour les données TOP100, l'analyse des taux de variation maximaux permet d'établir une gamme d'échelle de représentation comprise entre le 1 :100.000 et le 1 :110.000. Cette estimation est cohérente avec l'échelle de représentation de ces données, établie au 1 :100.000.

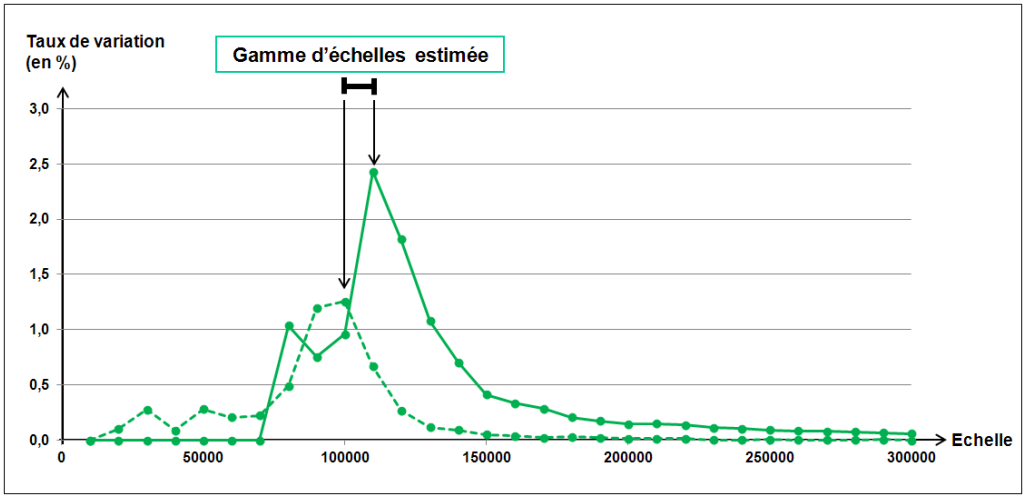


Figure 4.46 – Variation de la surface intersectée (en trait plein) et du nombre d'empâtements détectés (en pointillés) entre deux échelles successives pour les données TOP100

La figure 4.47 illustre les résultats d'une détection conjointe de superposition de symboles de routes voisines, et d'empâtements de virages, à partir des données TOP100. Les symboles des routes y sont représentés au 1 :100.000.

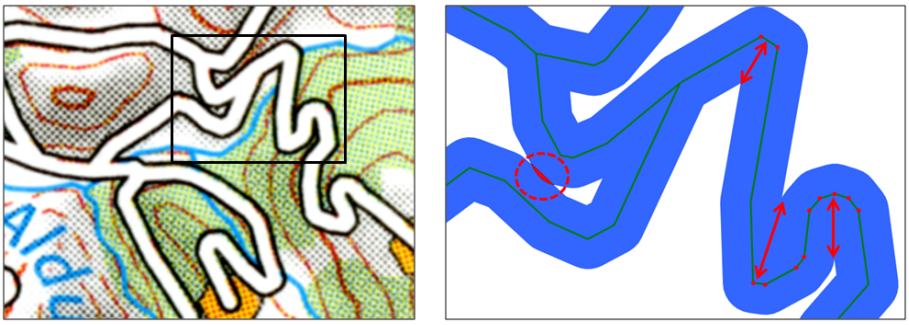


Figure 4.47 – Détection conjointe de superpositions de symboles de routes et de zones d'empâtement à partir des données TOP100

Pour les données de la Carte Régionale, l'analyse des taux de variation maximaux montre que les méthodes d'estimation proposées ont tendance à sous-estimer la gamme d'échelle de représentation. En effet, cette gamme d'échelle est ici comprise entre le 1 :150.000 et le 1 :160.000, ce qui est inférieur à l'échelle de représentation réelle de ces données.

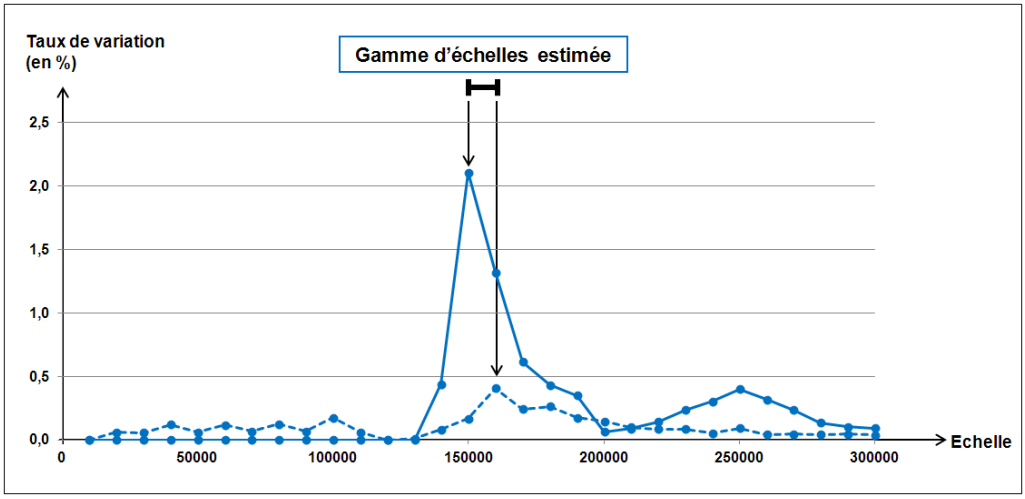


Figure 4.48 – Variation de la surface intersectée (en trait plein) et du nombre d'empâtements détectés (en pointillés) entre deux échelles successives pour les données de la Carte Régionale

La figure 4.49 permet d'illustrer la sous-estimation de l'échelle de représentation des routes de la Carte Régionale. Cette figure montre une extraction de routes sur lesquelles des superpositions de routes voisines sont détectées au 1 :150.000 et des empâtements de virages au 1 :160.000. Cette sous-estimation a tendance à montrer que les données de la Carte Régionale sont sous-généralisées.

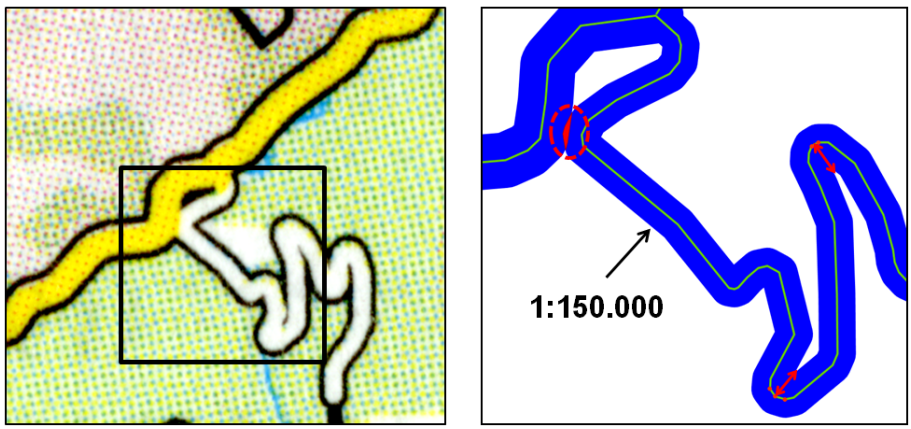


Figure 4.49 – Détection conjointe de superpositions de symboles de routes et de zones d'empâtement sur les données de la Carte Régionale

Cependant, il n'est pas étonnant que les données de la Carte Régionale soient légèrement sous-généralisées. En effet, produire une carte au 1 :250.000 est une entreprise très délicate, du fait de la proportion du territoire cartographié occupée par les symboles. Il est donc envisageable que les contraintes de représentation appliquées à cette échelle soient plus souples (que pour une carte au 1 :100.000 par exemple), tolérant des superpositions de symboles de routes ou des empâtements de virages qui ne nuisent pas à la compréhension générale du message cartographique. Ceci tend à expliquer pourquoi les méthodes proposées ont tendance à sous-estimer l'échelle de représentation de ces données.

Paramétrage du modèle Pour paramétrer le modèle, nous proposons donc d'estimer l'échelle de représentation par la borne inférieure de la gamme d'échelles établie à l'aide des deux méthodes d'estimation proposées précédemment.
Le pseudo-code ci dessous présente l'algorithme d'estimation de l'échelle caractéristique de représentation $repScale$ à partir des bornes inférieures $minScale$ et supérieures $maxScale$ de la gamme d'échelles déterminée par les taux de variation maximaux du nombre d'empâtements détectés $coalescenceVarMax$ et de la surface intersectée entre routes voisines $overlapVarMax$.

```
1  repScale = 10.000
2  minScale = 10.000
3  maxScale = 10.000
4  coalescenceVarMax = 0
5  coalescenceScale = 10.000
6  overlapVarMax = 0
7  overlapScale = 10.000
8  for each scale
9      if coalescenceVar > coalescenceVarMax then
10         coalescenceVarMax = coalescenceVar
11         coalescenceScale = scale
12     else if overlapVar > overlapVarMax then
13         overlapVarMax = overlapVar
14         overlapScale = scale
15     end if
16 end for
17 if overlapScale > coalescenceScale then
18     minScale = coalescenceScale
19     maxScale = overlapScale
20 else
21     minScale = overlapScale
22     maxScale = coalescenceScale
23 end for
24 repScale = minScale
```

Conclusions Les résultats observés confirment que les méthodes proposées offrent une solution intéressante pour estimer l'échelle de représentation d'une base de données cartographiques, bien qu'une tendance à la sous-estimation soit observée pour les données de la Carte Régionale. Néanmoins, ces méthodes nécessiteraient d'être validées sur d'autres bases de données. De plus, nous nous sommes ici basé sur des symbologies types, extraites des spécifications de cartes IGN françaises. Il semblerait nécessaire de compléter ces méthodes d'estimation de l'échelle de représentation avec les symboles extraits de cartes réalisées par d'autres producteurs, ou dans d'autres pays.

Après avoir proposé des méthodes d'estimation de l'échelle de représentation, l'élaboration du modèle d'estimation de l'impact de la généralisation cartographique suppose également de délimiter le contexte géographique des objets évalués. Les méthodes développées à cet effet sont présentées dans la section suivante.

4.2.2.2 Délimitation du contexte géographique des objets

Le modèle d'estimation de l'impact de la généralisation cartographique sur les mesures géométriques nécessite de connaître le contexte géographique des objets étudiés. Nous proposons dans cette section des méthodes permettant de le délimiter.

Notion de contexte géographique Dans le cadre de ce travail de thèse, nous entendons par contexte géographique une zone possédant une identité physique homogène relative au processus de généralisation cartographique. L'utilité de la délimitation du contexte géographique dans le cadre de l'estimation de l'impact de la généralisation cartographique fait notamment référence aux travaux de Touya (2011), qui propose de faire collaborer des processus de généralisation cartographique différents en fonction de "l'espace géographique" au sein duquel sont localisés les objets de la base de données. Nous emploierons ici le terme de contexte géographique des objets.

Trois contextes géographiques sont concernés dans le cadre de cette étude : les zones de montagne, les zone urbaines, et enfin les zones rurales. Afin de les délimiter, trois possibilités peuvent se présenter à l'utilisateur du modèle :
- soit l'utilisateur les délimite manuellement,
- soit des données de délimitation sont déjà disponibles,
- soit des méthodes de délimitation sont proposées à l'utilisateur.

Les deux premiers cas ne seront pas présentés dans cette section. En effet, la délimitation de ces zones de haut niveau peut être réalisée directement par un utilisateur, à l'aide d'outils de saisie classiques. De la même manière, de nombreuses bases de données disponibles gratuitement proposent des zonages pouvant être utiles à la délimitation du contexte géographique. C'est le cas de la base de données *CorineLandCover* (EEA, 2007), qui propose par exemple une délimitation des zones urbaines et péri-urbaines dans sa nomenclature. C'est d'ailleurs cette base de données qui est utilisée pour délimiter les zones urbaines dans l'étude de l'impact de la généralisation cartographique sur les mesures de longueur des tronçons routiers, en section 4.2.3.

Nous proposons dans cette section des méthodes permettant de faciliter la délimitation du contexte géographique, dans le cas où la connaissance du terrain par l'utilisateur ne lui permet pas de le faire, ou si des bases de données de zonage ne sont pas disponibles. Les méthodes présentées dans cette section sont destinées à permettre la délimitation de deux contextes géographiques : les zones de montagne et les zones urbaines. Nous considérons que les zones rurales correspondent aux zones qui ne sont ni urbaines, ni de montagne.

a. Délimitation des zones de montagne

La délimitation des zones de montagne a fait l'objet de nombreuses contributions. Les travaux les plus significatifs dans ce domaine sont ceux de Chaudhry et Mackaness (2008). Les auteurs proposent une méthode qui permet de déterminer la zone de proéminence d'un sommet en s'appuyant sur les courbes de niveau. Cette méthode permet de distinguer précisément les collines et leur chaine.

Objectifs Dans le cadre de ce travail, la délimitation des zones de montagne que nous cherchons à mettre en œuvre est beaucoup moins précise que celle proposée par Chaudhry et Mackaness (2008). En effet, nous ne cherchons pas à délimiter les zones de proéminence de sommets, mais plutôt une délimitation grossière des zones de montagne sur une surface étendue. Pour cela, l'approche proposée par Touya (2011) a été exploitée.
Dans cette approche, Touya (2011) définit une zone de montagne comme une zone où la pente est forte (et donc la densité en courbes de niveau est importante) et où le dénivelé est important. Cette définition, basée uniquement sur des critères physiques (et donc à la seule utilisation de courbes de niveau) satisfait nos exigences en termes de définition d'une zone homogène pour estimer l'impact de la généralisation cartographique. Nous allons utiliser cette approche, ainsi que la méthode proposée, en y apportant quelques ajustements.

Méthode utilisée Afin de délimiter les zones de montagne de manière automatique, Touya (2011) s'est inspiré des travaux de Walter (2008), qui reposent sur l'utilisation d'une grille carrée. Walter (2008) a développé une méthode générique permettant de classifier les cellules d'une grille selon un certain nombre de critères. Pour chaque critère, l'algorithme attribue une note comprise entre 1 et 3, suivant la valeur des données dans la cellule. Une moyenne pondérée de la note de chaque critère peut ensuite être réalisée, lorsque il est nécessaire d'appliquer plus de poids à certains critères qu'à d'autres.
Contrairement à Walter (2008), qui avait utilisé des critères appliqués à la détection des zones urbaines, Touya (2011) a utilisé deux critères appliqués à la détection des zones de montagne en se basant sur des courbes de niveau : le dénivelé et la densité de courbes de niveau.

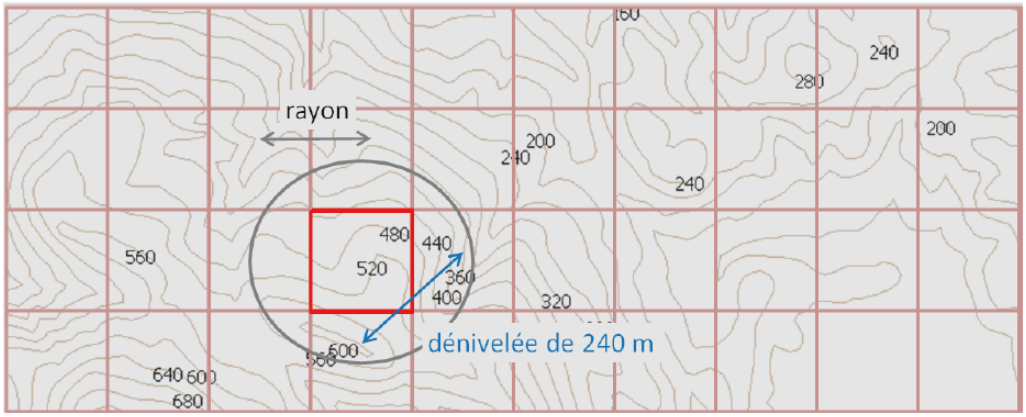


Figure 4.50 – Principe de détéction des zones de montagne avec un critère de dénivelé, d'après Touya (2011)

Pour le critère de dénivelé, comme l'illustre la figure 4.50, l'algorithme cherche dans un cercle de rayon donné (à partir du centre de cellule) les courbes de niveau intersectées, et compare la plus haute à la plus basse.
Pour le critère de densité, il calcule la longueur totale des courbes de niveau dans le cercle, puis affecte une note à la cellule. Une note est ensuite affectée à chaque cellule en fonction de la valeur de dénivelé ou de densité calculée. L'attribution des notes pour chaque critère est paramétrée préalablement par des seuils définis par l'utilisateur.

Comme le souligne Touya (2011), la méthode développée offre des résultats très satisfaisants pour un faible temps de calcul. Cette méthode implique le paramétrage préalable de :
– la taille de cellule,
– le rayon de recherche,
– les seuils pour chaque critère et pour la note finale de classification,
– le poids de chaque critère.
Il est donc nécessaire de procéder à de nombreux tests avant d'en déduire un paramétrage adapté.

Etant donné les bons résultats expérimentaux de cette méthode de délimitation des zones de montagne, ainsi que sa facilité de mise en œuvre, celle-ci a été utilisée dans le but d'accompagner l'utilisateur dans le paramétrage du modèle d'estimation. Cependant, nous avons cherché à nous limiter à un seul critère, relatif au dénivelé. En effet, le critère de densité est dépendant de l'équidistance choisie lors de la création des courbes de niveau. Lorsque l'équidistance augmente (ce qui est souvent le cas en zone de montagne), la densité de courbes de niveau baisse. Une méthode alternative est donc proposée afin de nous abstraire de cette contrainte.

Proposition Pour délimiter les zones de montagne, les résultats obtenus par deux formulations du critère de dénivelé ont été comparés.
La première formulation du critère de dénivelé (cf. équation 4.10) classifie une cellule en zone de montagne si la différence entre les altitudes maximale Z_{Max} et minimale Z_{Min} détectées dans un cercle de rayon donné, est supérieur à un seuil T.

$$Z_{Max} - Z_{Min} > T \tag{4.10}$$

La seconde formulation est inspirée des travaux de Jaara et Lecordix (2011). Elle est présentée dans l'équation 4.11, où Z_{Max} est l'altitude maximale, Z_{Min} l'altitude minimale, Z_{Moy} l'altitude moyenne, et T le seuil de classification.

$$2 * Z_{Max} - Z_{Min} - Z_{Moy} > T \tag{4.11}$$

Application Pour délimiter les zones de montagne, les courbes de niveau, produites avec une équidistance de 40 mètres ont été extraites du Modèle Numérique d'Elevation SRTM, sur l'ensemble du département des Pyrénées-Atlantiques (cf. figure 4.51).

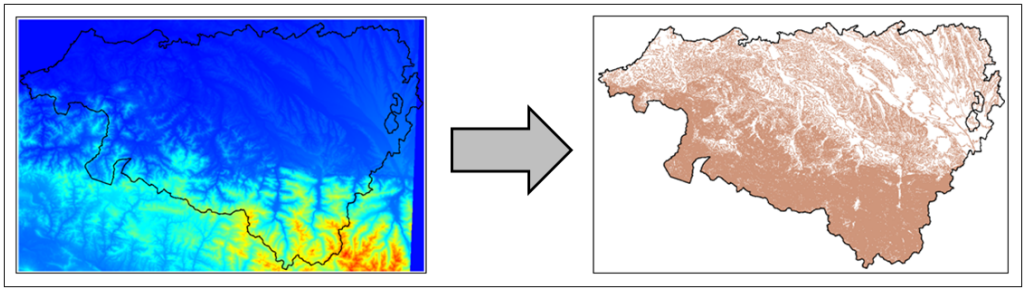


Figure 4.51 – Conversion du MNE SRTM en courbes de niveau

Une grille carrée de résolution de 500 mètres a été créée, le rayon de recherche étant également paramétré à 500 mètres. Les deux critères ont été testés avec des valeurs de seuil de 100, 200 et 300 mètres. Les résultats sont présentés dans la figure 4.52.

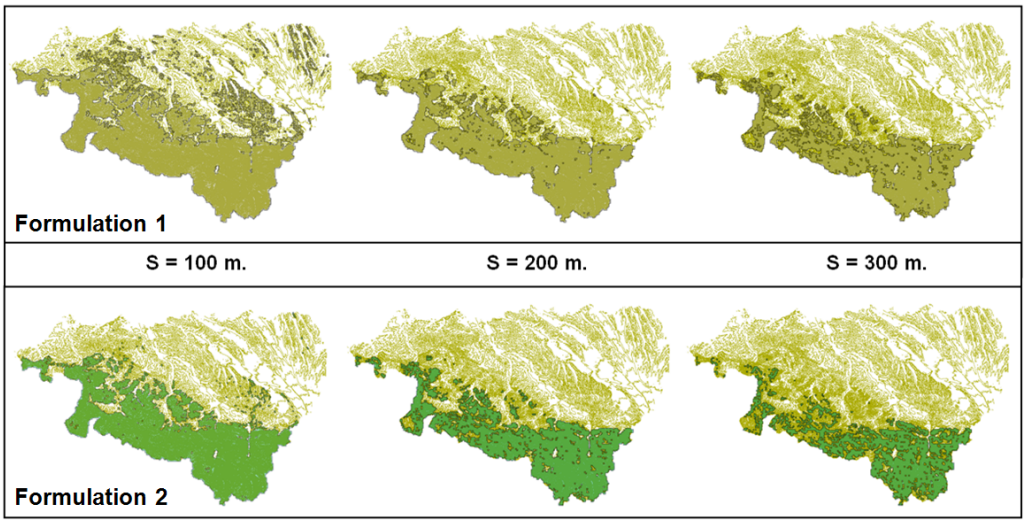


Figure 4.52 – Zones de montagne détectées selon les deux formulations du critère de dénivelé

Résultats Les délimitations obtenues avec les deux formulations du critère de dénivelé et les différentes valeurs du seuil S ont été comparées. Nous avons observé que la première formulation du critère, s'il est utilisé seul, a pour tendance de sur-détecter les zones de montagne, en comparaison avec la seconde formulation. Par conséquent, la seconde formulation, avec un valeur de seuil de 200 mètres, s'est avérée être la plus réaliste pour détecter les zones de montagne.

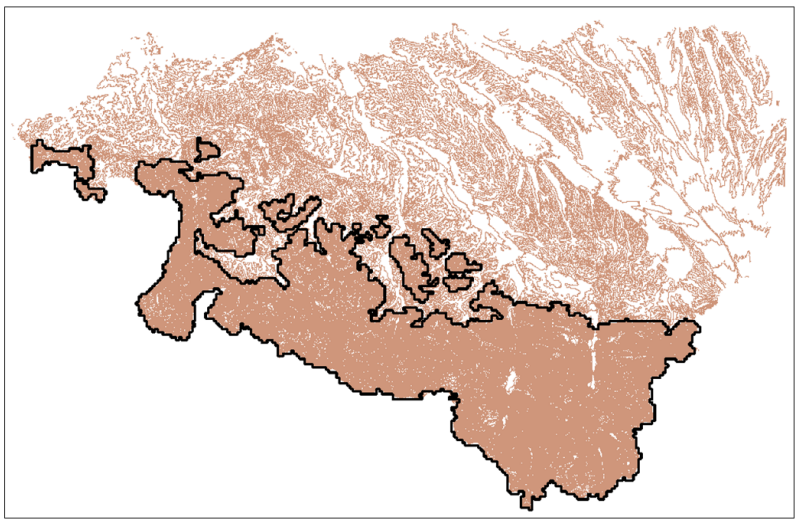


Figure 4.53 – Résultat de la délimitation de la zone de montagne suite aux opérations de nettoyage et de lissage

Afin de disposer d'une zone de montagne homogène, plusieurs post-traitements ont été réalisés :
– suppression des petits polygones détectés (dont la superficie est inférieure à 4 km^2),
– suppression des trous dans la zone de montagne.
Enfin, une opération morphologique de dilatation puis d'érosion (Serra, 1983) a été réalisée pour lisser la délimitation des zones de montagne obtenue. Celle-ci est présentée dans la figure 4.53.

Cette délimitation des zones de montagne a été utilisée dans le cadre des expérimentations permettant d'estimer l'impact de la généralisation cartographique sur les mesures de longueur d'un réseau routier (voir également Girres (2011a)). La méthode de détection ainsi proposée offre des résultats satisfaisants, limités à un seul critère de dénivelé. De plus, cette méthode de détection ne nécessite que des courbes de niveau en entrée, qui peuvent être extraites de Modèles Numériques d'Elévation aisément disponibles. Cette méthode de délimitation des zones de montagne est donc intégrée au modèle d'estimation, afin d'accompagner l'utilisateur dans son paramétrage.

b. Délimitation des zones urbaines

Le paramétrage du modèle d'estimation de l'impact de la généralisation cartographique implique également de délimiter les zones urbaines. Dans le cadre des expérimentations qui seront réalisées en section 4.2.3, des données de délimitation des zones urbaines déjà produites (la base de données *CorineLandCover*) sont utilisées. Cependant, ces données ne sont disponibles que sur le territoire européen.

Afin de mettre en œuvre le modèle d'estimation de l'impact de la généralisation cartographique, peu importe la zone étudiée ou la disponibilité de données de délimitation des zones urbaines, nous proposons donc dans cette section des méthodes permettant de délimiter automatiquement les zones urbaines. De la même manière que pour les zones de montagne, les zones urbaines que nous souhaitons ici délimiter sont des zones de haut niveau, à la géométrie relativement grossière.

La détection des zones urbaines a fait l'objet de nombreux travaux, en particulier en généralisation cartographique. Différentes méthodes ont été proposées en fonction des données en entrée. Par exemple, Boffet (2000) ou Chaudry et Mackaness (2008) ont proposé des méthodes de délimitation des zones urbaines basées sur la proximité du bâti. Walter (2008) a quant à lui proposé une méthode par grille basée sur la densité du réseau routier.
Pour faciliter le paramétrage du modèle d'évaluation de l'impact de la généralisation cartographique, nous avons retenu deux méthodes permettant de délimiter les zones urbaines, l'une basée sur les bâtiments et l'autre sur le réseau routier.

Méthode basée sur les bâtiments Afin de délimiter les zones urbaines et péri-urbaines à partir d'une base de données de bâtiments, l'algorithme issu de Boffet (2000) est utilisé. Cet algorithme fonctionne en créant des zones tampons (de 25 mètres en zone urbaine, et de 75 mètres en zone péri-urbaine) autour de chaque bâtiment. Les zones tampons obtenues sont ensuite assemblées.

Un succession d'opérations est appliquée sur la surfaces résultante :
– élimination des petites surfaces,
– opération morphologique de dilatation puis d'érosion (Serra, 1983),
– filtrage de la zone à l'aide de l'algorithme de Douglas et Peucker (1973),
– suppression des trous dans la zone.

Cette méthode, plutôt simple à mettre en œuvre, offre des résultats satisfaisants, comme le montre la figure 4.54, extraite de Touya (2011), qui différencie les zones urbaines des zones péri-urbaines.

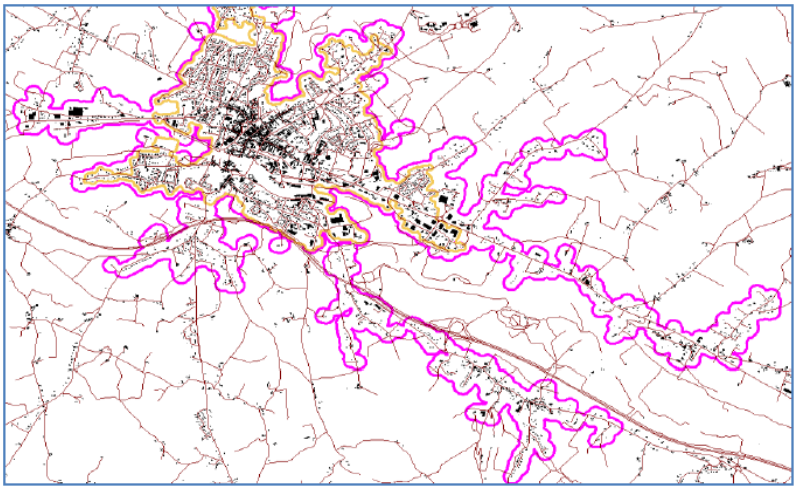


Figure 4.54 – Résultat de la délimitation des zones urbaines et péri-urbaines par la méthode de Boffet (2000), d'après Touya (2011)

La méthode proposée par Boffet (2000) offre donc une solution relativement aisée à mettre en œuvre, lorsque l'utilisateur dispose d'une base de données sur les bâtiments. Si ce n'est pas le cas, nous proposons également une méthode de délimitation des zones urbaines basée sur le réseau routier.

Méthode basée sur le réseau routier Pour détecter les zones urbaines à partir d'un réseau routier, nous aurions pu utiliser la méthode de Walter (2008), qui s'appuie sur un grille carrée afin de délimiter les zones urbaines, en fonction de la densité de routes inscrites dans les cellules de la grille. Cependant, lorsque le réseau routier présente des problèmes de complétude, il est nécessaire de disposer de cellules de taille importante afin de bien différencier celles contenant une forte densité de routes. Les résultats obtenus sont par conséquent peu satisfaisants, puisque les zones détectées présentent des formes rectangulaires. Ainsi, une méthode basée sur la surface des faces du graphe routier a été spécialement développée.

A partir d'un réseau routier, converti en graphe planaire, nous pouvons considérer que les zones urbaines sont celles correspondant aux petites faces du graphe. En effet, le réseau routier permet de délimiter des pâtés de maisons, appelés également blocs. En milieu urbain, ces blocs présentent généralement de petites surfaces, comme le montre la figure 4.55.

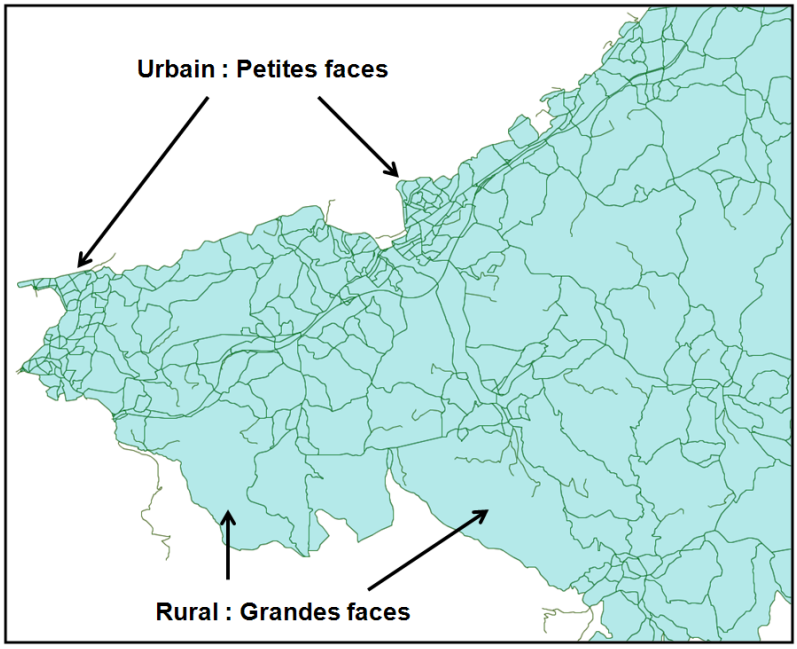


Figure 4.55 – Petites faces en milieu urbain et grandes faces en milieu rural

La méthode que nous proposons pour délimiter les zones urbaines repose donc sur une sélection des faces du graphe qui présentent une superficie $surfaceFace$ inférieure à une valeur seuil $T1$. Cependant, l'application brute de cette méthode n'est pas satisfaisante, puisque de nombreux polygones résiduels subsistent. C'est par exemple le cas des polygones étroits, formés entre deux routes parallèles, comme par exemple les tronçons d'autoroutes (cf. figure 4.56). Pour éliminer ces polygones, nous proposons d'utiliser un indicateur de compacité des faces $compacityFace$ qui doit être supérieur à un seuil $T2$.

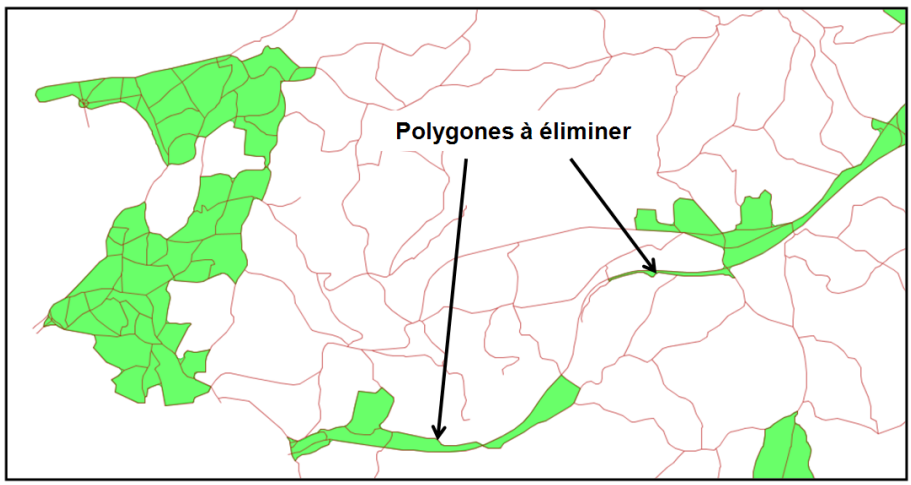


Figure 4.56 – Polygones à éliminer pour la détection des zones urbaines

Les faces résultantes sont ensuite assemblées afin de former une zone urbaine, qui doit elle aussi présenter une superficie $surfaceZone$ supérieure à un seuil $T3$.

L'algorithme de sélection des zones urbaines est présenté dans le pseudo code ci-dessous.

```
1  for each Face :dbFaces < t
2     if surfaceFace < T1 then
3          if compacityFace > T2 then
4               dbUrbanZone.union(Face)
5          end if
6     end if
7  end for
8  for each Zone :dbUrbanZone
9     if surfaceZone < T3 then
10         dbUrbanZone.remove(Zone)
11    end if
12 end for
```

Suite à la sélection des zones urbaines, plusieurs post-traitements sont réalisés :
- suppression des trous dans la zone,
- opération morphologique de dilatation puis d'érosion (Serra, 1983).

Afin de tester la méthode de délimitation des zones urbaines à partir du réseau routier, une expérimentation a été réalisée sur le réseau routier de la BDCARTO dans le département des Pyrénées-Atlantiques. Les seuils utilisés pour réaliser cette délimitation sont les suivants :
- taille maximale des blocs = 0,2 km^2,
- taille minimale des zones urbaines = 0,4 km^2,
- seuil de compacité = 0,2.

La figure 4.57 présente le résultat des zones urbaines obtenues.

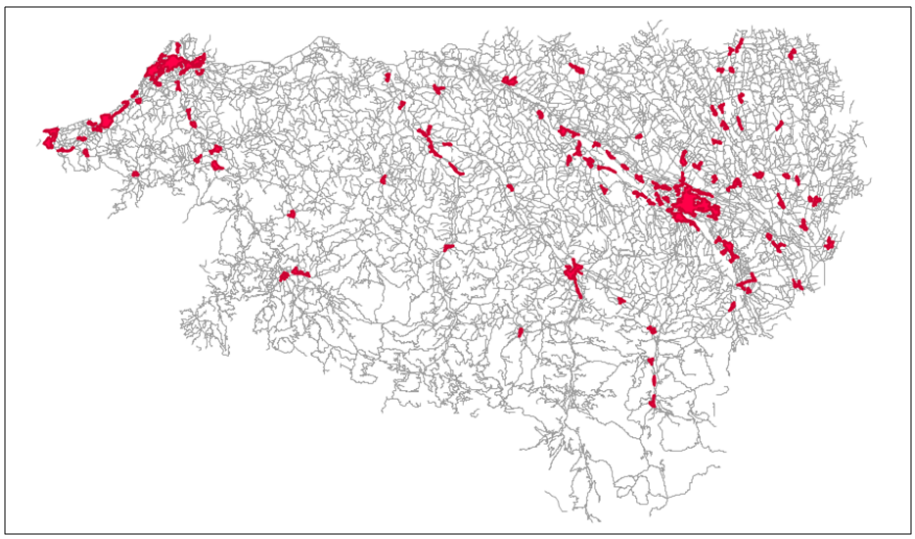


Figure 4.57 – Résultats de la délimitation des zones urbaines sur le département des Pyrénées-Atlantique, à partir du réseau routier de la BDCARTO

Les résultats de l'expérimentation, bien que sous-estimant légèrement les zones urbaines du département des Pyrénées-Atlantiques, sont globalement satisfaisants. En effet, il est important de prendre en compte l'impact de la complétude du réseau routier utilisé sur les résultats obtenus.
Cependant, bien que cette méthode impose d'expérimenter de manière empirique chacun des paramètres en fonction du réseau routier utilisé, elle offre malgré tout des résultats intéressants afin de délimiter les zones urbaines. Cette méthode a donc été intégrée au sein du modèle d'estimation, pour permettre à un utilisateur de délimiter les zones urbaines à partir d'un réseau routier.

Après avoir présenté les méthodes permettant d'estimer l'échelle de représentation d'une base de données cartographiques et de délimiter le contexte géographique des objets évalués, il est à présent possible de mettre en œuvre le processus de comparaison permettant d'élaborer le modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures de longueur.

Synthèse Section Nous avons proposé dans cette section des méthodes permettant de paramétrer le modèle d'estimation de l'impact de la généralisation cartographique. Tout d'abord, des méthodes d'estimation de l'échelle de représentation d'une base de données cartographiques (ou bien de l'échelle de la carte ayant permis de saisir la géométrie des objets) ont été proposées. Ces méthodes, appliquées à l'étude du réseau routier, reposent sur la connaissance de la taille du symbole appliqué aux objets de la base. Egalement, des méthodes de délimitation du contexte géographique, à savoir les zones de montagne et les zones urbaines, ont été proposées, en se basant sur des données externes.

4.2.3 Impact de la généralisation sur les mesures de longueur

Cette section a pour objectif de proposer un modèle d'estimation de l'impact de la généralisation cartographique sur les mesures géométriques de longueur. Le modèle développé doit permettre d'estimer la longueur de référence à partir d'une mesure de longueur calculée sur une base de données généralisée. Ceci équivaut à répondre à la question : En mesurant une longueur de x mètres sur une carte généralisée, quelle peut-être la mesure de longueur réelle ?
Pour répondre à cette question, nous proposons d'élaborer un modèle statistique basé sur des comparaisons de longueurs entre des bases de données généralisées (ayant des échelles de représentation différentes), et une base de données de référence (représentant le terrain nominal). Ces comparaisons sont réalisées dans trois contextes géographiques, à savoir en milieu urbain, rural, et de montagne, étant donné que nous supposons que les impacts ne sont pas identiques. Pour produire une estimation la moins biaisée possible de l'impact de la généralisation cartographique sur les mesures de longueur, il est nécessaire de réaliser les comparaisons de longueurs sur une grande quantité d'objets. Pour cela, un processus de comparaison automatisé est élaboré.

La mise en œuvre d'un processus automatisé de comparaison de longueurs soulève cependant une question opérationnelle, à savoir si l'erreur de longueur impliquée par la généralisation cartographique est sensible à la longueur étudiée. En d'autres termes, nous pouvons nous demander si l'erreur relative de longueur estimée est la même sur un tronçon de 1 kilomètre que sur un tronçon de 100 kilomètres, dans des conditions identiques (même base de données et même contexte géographique). Nous tenterons donc de répondre à cette question dans la première section, puis présenterons la méthodologie d'élaboration du modèle d'estimation et sa validation dans les sections suivantes.

4.2.3.1 Sensibilité de l'erreur relative à la longueur

Pour savoir si l'erreur de longueur impliquée par la généralisation cartographique est sensible à la longueur étudiée, la méthodologie et les résultats proposés dans Girres (2011a) seront exposés dans cette section. Au cours cette étude, nous avons cherché à estimer l'erreur moyenne de longueur sur des tronçons de route généralisés, en nous basant sur une méthodologie de comparaison normalisée. Nous présenterons tout d'abord la méthodologie de comparaison développée au cours de cette étude, les bases de données utilisées, puis discuterons des résultats en vue de l'élaboration du modèle statistique.

Méthodologie de comparaison

La mise en œuvre d'un processus automatisé de comparaison pour étudier la sensibilité de l'erreur relative à la longueur étudiée requiert des traitements préalables permettant de normaliser la géométrie des objets homologues comparés. Ainsi, la méthodologie générale de comparaison se décline en trois étapes :

1. Appariement des objets homologues entre les deux bases de données,
2. Normalisation de la géométrie des objets appariés,
3. Comparaison des longueurs entre objets normalisés.

Appariement des données La première étape consiste à mettre en relation les objets homologues des deux bases de données comparées. De nombreux travaux ont été réalisés ces dernières années afin de mettre en correspondance les objets géographiques de plusieurs bases de données, comme les travaux de Devogèle (1997) pour les objets linéaires, ceux de Bel Hadj Ali (2001) pour les objets surfaciques, ou encore ceux de Olteanu (2008).

Dans le cadre de ce travail, l'algorithme développé par Mustière et Devogèle (2008) est utilisé afin d'apparier les objets linéaires homologues entre les deux bases de données. Cet algorithme offre la possibilité d'établir des liens d'appariement avec plusieurs cardinalités : 1-1, 1-n ou m-1.

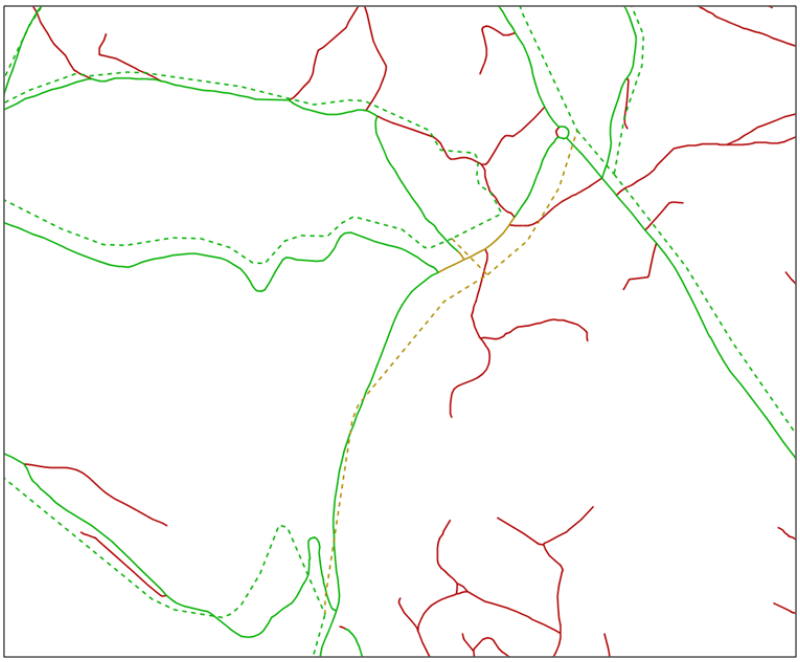


Figure 4.58 – Résultats d'un appariement entre les tronçons de route de la BDTOPO (en trait plein) et de la BDCARTO (en tiretés) en trois classes : "Apparié" (en vert), "Incertain" (en orange), "Non apparié" (en rouge)

Les résultats de l'appariement utilisant l'algorithme de Mustière et Devogèle (2008) sont classés en trois catégories de lien d'appariement : "Apparié", "Incertain" et "Non-apparié" (cf. figure 4.58). Bien que cet algorithme d'appariement soit relativement pessimiste, nous n'utilisons dans le cadre de cette étude que les polylignes correctement appariées. Ce choix, plus sûr, nous fait par contre perdre des informations potentiellement intéressantes. En effet, certains objets, pourtant homologues, verront parfois leur appariement classé "Incertain" pour différentes raisons (une différence d'orientation trop importante par exemple). Ces objets peuvent pourtant présenter des erreurs de longueur importantes, qu'il serait intéressant d'étudier. Cependant, pour procéder à l'automatisation du processus de comparaison, seuls les liens d'appariement certains seront conservés.

Normalisation géométrique La seconde étape du processus de comparaison réside dans une phase de pré-traitement nécessaire des objets à comparer afin de procéder à une comparaison normalisée. Le principe de comparaison normalisée implique de découper les objets appariés de la base de données de référence en tronçons de longueurs identiques, et de comparer leur longueur avec les tronçons homologues découpés à partir des objets appariés de la base de données généralisée. Pour permettre une comparaison normalisée entre tronçons homologues, plusieurs étapes successives de pré-traitement des géométries doivent être entreprises, comme illustré en figure 4.59.

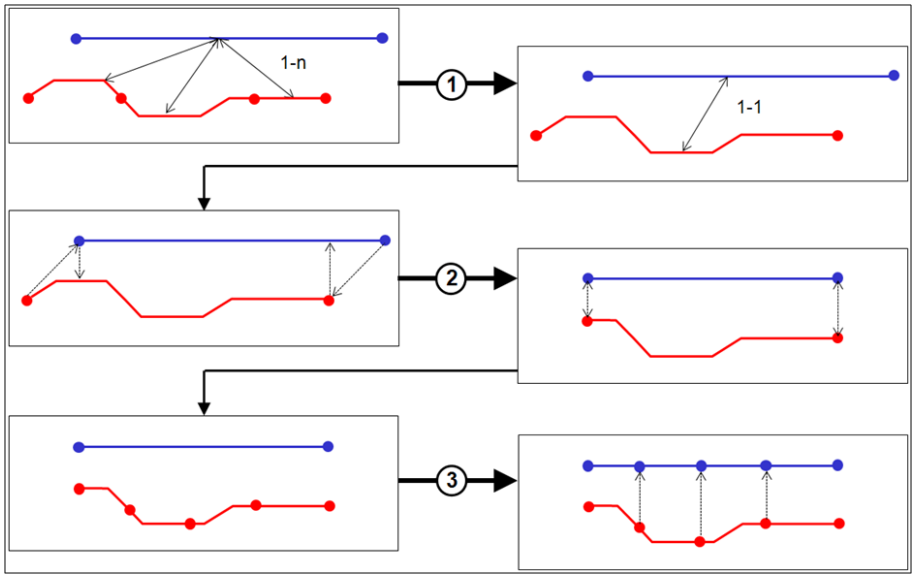


Figure 4.59 – Pré-traitements des polylignes pour une comparaison normalisée

Le premier pré-traitement consiste en une **fusion des objets continus** appariés avec le même lien. En effet, suite à l'appariement automatique, à un objet généralisé peuvent correspondre plusieurs objets de référence, ou l'inverse (c'est à dire un lien 1-n). Les objets appariés peuvent de plus présenter des discontinuités, ce qui ne facilite pas les comparaisons. L'objectif de cette étape vise donc à ne disposer que de liens 1-1.

La second pré-traitement réside dans le **découpage des extrémités des polylignes** fusionnées, qui peuvent générer des erreurs de longueur importantes lors d'une comparaison "brute". Le découpage des extrémités des polylignes est réalisé par projection des extrémités de chaque polyligne vers son homologue.

Une fois les extrémités des polylignes découpées, la polyligne de référence est ensuite sectionnée en tronçons de longueurs identiques. Différentes longueurs de tronçons sont pour cela utilisées, afin d'étudier la sensibilité de l'erreur de mesure en fonction de la longueur du tronçon de référence sectionné.

Enfin, la dernière étape réside dans le **découpage de la polyligne généralisée en tronçons**, par projection des extrémités de chacun des tronçons de la polyligne de référence sur la polyligne généralisée.

Une fois ces différents pré-traitement réalisés, il est possible de procéder à une comparaison des longueurs entre chaque couple de tronçons homologues (cf. figure 4.60).

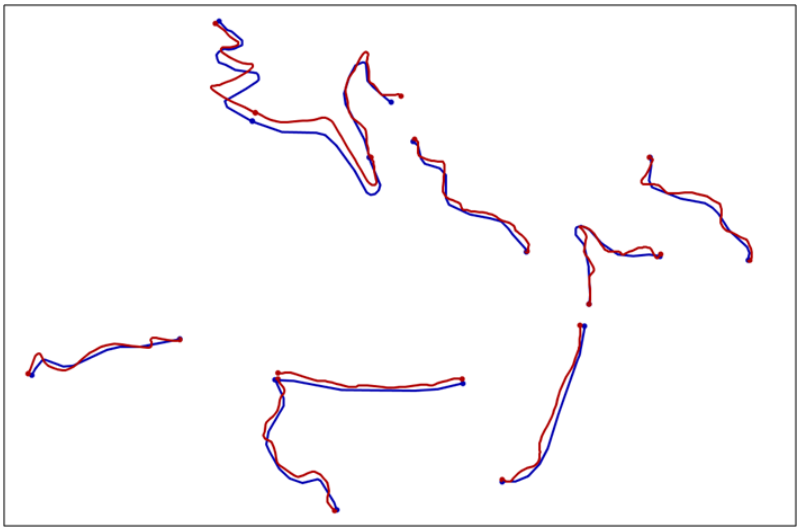


Figure 4.60 – Comparaison de tronçons de routes avec une longueur de référence de 800 m, entre les données TOP100 (en bleu) et la BDTOPO (en rouge)

Comparaison des longueurs La comparaison des longueurs entre objets généralisés et objets de référence s'effectue donc par tronçons, disposant tous d'une même longueur de référence. Même si les objets de référence disposent d'une géométrie tridimensionnelle, permettant des mesures de longueur en 2D5, les comparaisons de longueurs ne sont effectuées qu'en deux dimensions, afin de ne pas biaiser l'analyse (puisque la base de données généralisée ne contient pas d'information altimétrique). Des longueurs croissantes (allant de 200 à 2000 mètres) des tronçons de référence sont expérimentées afin d'étudier la sensibilité de l'erreur de longueur relative. Enfin, les comparaisons sont effectuées isolément dans chacun des contextes géographiques, afin d'étudier son impact sur l'erreur de longueur.

Bases de données utilisées

Pour estimer l'impact de la généralisation cartographique sur les mesures de longueur dans chacun des trois contextes géographiques (urbain, rural et montagne), il a été décidé de travailler sur le département des Pyrénées-Atlantiques. Ce département offre l'avantage de présenter sur son territoire les trois contextes géographiques sélectionnés. Les tests de comparaison proposés ici ne portent que sur l'étude du **réseau routier**. Ceci se justifie puisque, de manière générale, le réseau routier est la classe d'objets la plus souvent utilisée pour réaliser des mesures de longueur (en particulier pour les calculs d'itinéraires), en plus de son rôle structurant de l'espace.

Bases de données comparées Trois bases de données généralisées et une base de données de référence, toutes produites par l'Institut Géographique National, ont été exploitées dans cette étude. Les bases de données généralisées utilisées sont :

- la BDCARTO, saisie initialement à partir des cartes au 1 :50.000,
- les données TOP100, qui servent à la production des cartes à l'échelle 1 :100.000,
- les données utilisées pour la production de la Carte Routière Régionale au 1 :250.000.

La base de données de référence utilisée est la BDTOPO, base de données topographique de précision métrique, saisie principalement par restitution photogrammétrique et relevés GPS de haute précision. Comme illustré dans la figure 4.61, un même réseau routier représenté par ces quatre sources de données présente des différences importantes en termes de position et de longueur géométrique.

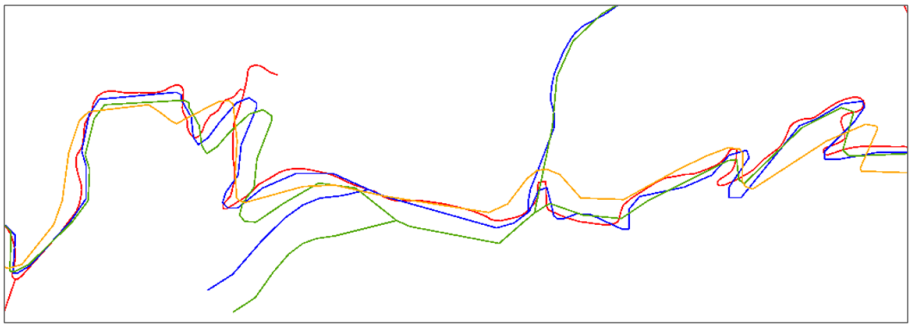


Figure 4.61 – Un même réseau routier représenté par la BDTOPO (en rouge), la BD-CARTO (en bleu), les données TOP100 (en vert) et celles de la Carte Régionale (en orange)

Délimitation du contexte géographique Afin de valider nos hypothèses, les comparaisons de longueurs entre les objets homologues ont été réalisées dans les trois contextes géographiques précédemment évoqués : urbain, rural et de montagne (cf. figure 4.62).

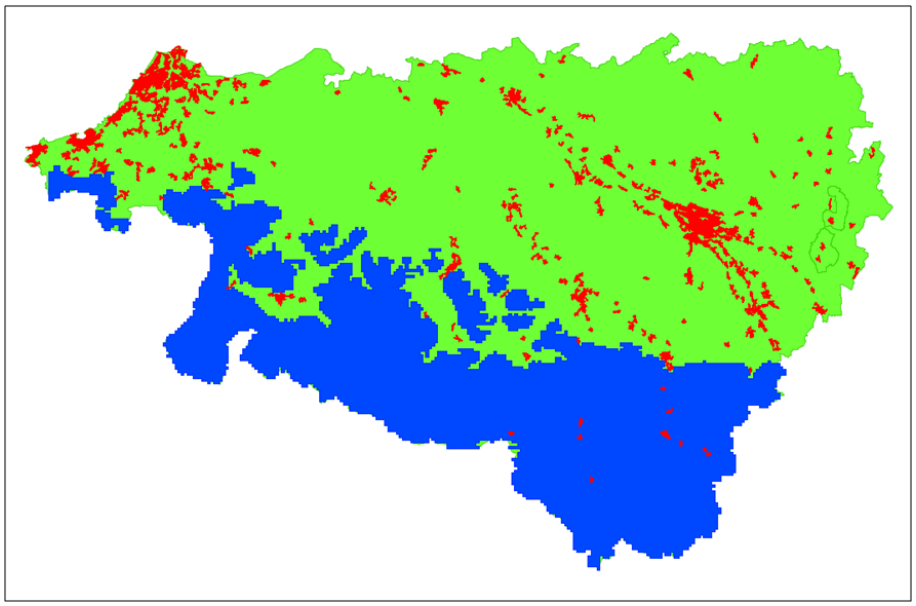


Figure 4.62 – Délimitation du contexte géographique en zones de montagne (en bleu), urbaines (en rouge) et rurales (en vert)

La délimitation de ces trois contextes géographiques a été réalisée en utilisant des bases de données externes, disponibles gratuitement. Pour délimiter les zones de montagne, la méthode proposée en section 4.2.2.2 a été exploitée, à partir des courbes de niveau extraites du modèle numérique d'élévation SRTM (Farr *et al.*, 2007).

En revanche, pour les zones urbaines, le zonage fourni dans la base de données *CorineLand-Cover* (EEA, 2007) a été utilisé. La délimitation des zones urbaines retenues correspond aux zones présentant les codes 111 (*Continuous Urban Areas*) et 112 (*Discontinuous Urban Areas*). Nous considérons que les zones rurales correspondent au complément des zones urbaines et des zones de montagne.

Etude de la sensibilité de l'erreur moyenne

L'expérimentation réalisée dans cette étude porte sur une extraction du réseau routier départemental des Pyrénées-Atlantiques. Des comparaisons basées sur des tronçons de référence de longueur croissante (de 200 à 2000 mètres) ont été réalisées sur les objets localisés en zone rurale et de montagne. En zone urbaine, la longueur maximale des tronçons étudiés n'a pu être portée qu'à 600 mètres. Au delà de cette longueur, du fait des nombreuses intersections routières présentes en milieu urbain, il devenait difficile de disposer d'un échantillon d'objets de population suffisamment importante.

	Montagne						**Rural**					
	CR		**T100**		**BDC**		**CR**		**T100**		**BDC**	
Ref.	**m.**	**%**	**m.**	**%**	**m.**	**%**	**m.**	**%**	**m.**	**%**	**m.**	**%**
200	8,9	*4,4*	9,1	*4,5*	1,0	*0,5*	8,8	*4,4*	6,6	*3,3*	0,4	*0,2*
400	24,1	*6,0*	23,1	*5,8*	3,8	*0,9*	17,6	*4,4*	14,0	*3,5*	1,5	*0,4*
600	36,4	*6,0*	34,3	*5,7*	5,7	*0,9*	26,9	*4,5*	21,8	*3,6*	2,7	*0,5*
800	48,3	*6,0*	45,6	*5,7*	8,2	*1,0*	35,6	*4,5*	29,5	*3,7*	3,7	*0,5*
1000	61,2	*6,0*	58,9	*5,9*	9,4	*0,9*	44,0	*4,4*	37,2	*3,7*	4,5	*0,4*
1200	71,3	*5,9*	68,5	*5,7*	13,3	*1,1*	53,2	*4,4*	46,3	*3,9*	5,4	*0,4*
1400	79,3	*5,7*	80,2	*5,7*	14,3	*1,0*	63,1	*4,5*	54,9	*3,9*	7,3	*0,5*
1600	93,3	*5,8*	89,1	*5,6*	14,3	*0,9*	72,1	*4,5*	65,6	*4,1*	8,7	*0,5*
1800	107,3	*6,0*	105,0	*5,8*	17,3	*1,0*	82,4	*4,6*	74,0	*4,1*	9,8	*0,5*
2000	114,7	*5,7*	116,4	*5,8*	20,1	*1,0*	89,7	*4,5*	84,3	*4,2*	9,4	*0,5*

	Urbain					
	CR		**T100**		**BDC**	
Ref.	**m.**	**%**	**m.**	**%**	**m.**	**%**
200	10,4	*5,2*	5,7	*2,8*	0,3	*0,1*
400	20,5	*5,1*	12,4	*3,1*	0,8	*0,2*
600	28,5	*4,7*	18,8	*3,1*	0,9	*0,1*

Table 4.7 – Erreur de longueur moyenne absolue et relative en fonction de la longueur de tronçon pour les données de la Carte Régionale (CR), TOP100 (T100) et de la BDCARTO (BDC), d'après Girres (2011a)

Erreur moyenne Les résultats des comparaisons sur les objets routiers, exposés en table 4.7, montrent tout d'abord, comme l'a souligné João (1998), que de manière générale, l'impact de la généralisation cartographique se manifeste par une diminution des longueurs moyennes calculées à partir de la géométrie des objets généralisés, peu importe le contexte géographique.

Comme illustré dans le graphique de la figure 4.63, les erreurs moyennes de longueur (relatives à la longueur du tronçon de référence) sont positives, ce qui signifie que les longueurs des tronçons de référence sont plus longues, en moyenne, que celles calculées sur leurs homologues dans les bases de données généralisées.

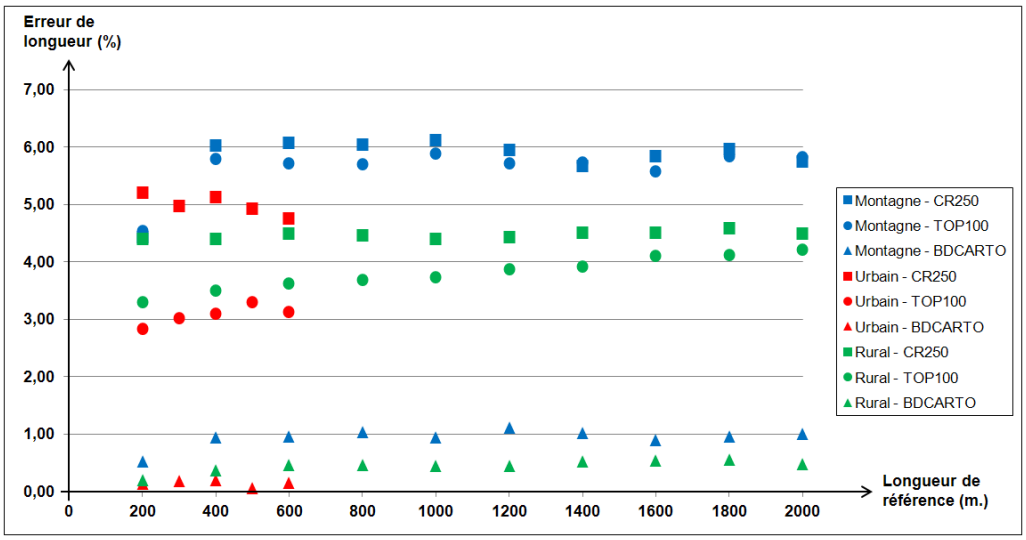


Figure 4.63 – Erreur moyenne de longueur relative à la longueur du tronçon de référence, en zone de montagne (bleu), rurale (vert) et urbaine (rouge), pour les données BDCARTO (rond), TOP100 (triangle), et de la Carte Régionale (carré), d'après Girres (2011a)

Impact du contexte géographique Ces résultats montrent également que le contexte géographique influe sur l'erreur de longueur des objets d'un réseau routier. En effet, pour toutes les bases de données généralisées testées (BDCARTO, TOP100 ou Carte Régionale) les erreurs de longueur calculées sur les objets du réseau routier sont plus importantes en zone de montagne que dans les autres contextes géographiques (cf. figure 4.63). Par exemple, la longueur des tronçons de route extraits des données TOP100, pour une longueur de référence de 400 mètres, est sous-estimée en moyenne de 5,8% en zone de montagne, alors qu'elle ne l'est que de 3,5 et 3,1 % en zones rurale et urbaine. Cette observation est valable quelque soit la longueur de tronçon étudiée ou la base de données utilisée.

Impact de l'échelle de représentation Outre l'importance du contexte géographique sur les erreurs de longueur, ces résultats prouvent avant toute chose que le facteur influant le plus sur l'erreur de longueur des objets généralisés est l'échelle de représentation. En effet, nous observons que l'erreur moyenne de longueur est beaucoup plus importante sur les objets issus des données TOP100 (1 :100.000) que sur ceux de la BDCARTO (1 :50.000). Cette différence d'erreur moyenne de longueur est moins marquée entre les objets de la Carte Régionale (1 :250.000) et ceux des données TOP100 (1 :100.000). On voit donc bien qu'en plus du contexte géographique, l'échelle de représentation impacte fortement les mesures de longueur.

Impact de la longueur de tronçon On observe enfin à la lecture de ces résultats que la longueur du tronçon de référence influe peu sur l'erreur moyenne de longueur. En effet, on peut observer que les erreurs moyennes de longueur relatives sont quasiment constantes pour chaque base de données et dans chaque contexte géographique étudié, peu importe la longueur du tronçon du référence. Par exemple, l'erreur moyenne de longueur relative des tronçons extraits de la Carte Régionale en zone rurale ne varie que de 4,4% à 4,6% par rapport à la longueur de référence.

Cette dernière conclusion permet donc de répondre à notre interrogation initiale, et donc de confirmer que l'erreur de longueur relative impliquée par la généralisation cartographique n'est pas sensible à la longueur étudiée. Dans le cadre de l'élaboration du modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures de longueur, cette conclusion nous permet donc de nous abstenir de procéder par la suite à une comparaison normalisée par tronçons.

D'une manière générale, même si la méthodologie utilisée et les données de référence sont différentes, ces résultats rejoignent ceux de João (1998), qui avait observé une diminution des erreurs moyennes de longueur des routes d'environ 3,8% en passant du 1 :50.000 au 1 :200.000 sur les cartes portugaises, suite à un processus manuel de généralisation. A la différence de João (1998), cette étude a permis de montrer que le contexte géographique influe sur l'erreur moyenne de longueur, en plus de l'échelle de représentation des données. A partir de ces conclusions, nous allons montrer dans la section suivante comment élaborer un modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures de longueur.

4.2.3.2 Elaboration du modèle statistique

Nous allons présenter dans cette section les résultats des comparaisons permettant d'élaborer le modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures géométriques de longueur (voire également en Annexe E).

Méthodologie de comparaison

En nous basant sur les conclusions de la section précédente, les comparaisons de longueurs entre objets homologues sont réalisées sur des tronçons de référence de longueurs hétérogènes. Ainsi, suite à une étape préalable d'appariement automatique utilisant l'algorithme de Mustière et Devogèle (2008), et en ne conservant que les liens d'appariement classés "Apparié", la méthodologie de comparaison utilisée ne reprend que les deux premières étapes de pré-traitement proposées précédemment (cf. figure 4.59), à savoir :
- fusion des objets continus appariés avec le même lien,
- découpage des extrémités de polylignes fusionnées, par projection des extrémités de chaque polyligne.

Les comparaisons de longueurs sont effectuées sur des objets de référence de longueurs hétérogènes. Ainsi, pour l'élaboration du modèle statistique, nous ne nous intéresserons qu'à l'erreur de longueur relative à la longueur de la polyligne généralisé. La figure 4.64 illustre la comparaison de tronçons de route de longueurs hétérogènes entre les données TOP100 et la BDTOPO.

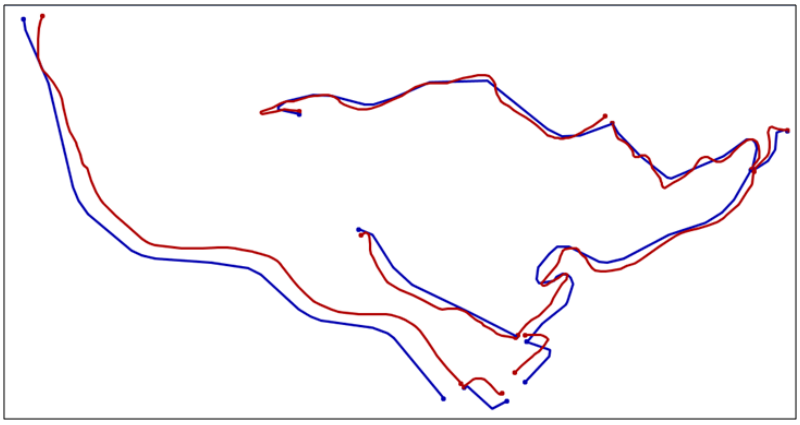


Figure 4.64 – Comparaison de tronçons de route de longueurs hétérogènes, des données TOP100 (en bleu) et de la BDTOPO (en rouge)

Bases de données utilisées

Les comparaisons ont été réalisées uniquement sur la classe d'objet des **tronçons de route**. Cependant, l'ensemble du réseau routier du département des Pyrénées-Atlantiques est ici mobilisé pour l'élaboration du modèle statistique, contrairement à l'étude précédente, où seulement une extraction de ce même réseau avait été utilisée.

Pour élaborer le modèle statistique d'estimation de l'impact de la généralisation cartographique, les mêmes bases de données généralisées ont été utilisées, à savoir :
– la BDCARTO,
– les données de la carte TOP100,
– les données de la Carte Régionale.
De la même manière, la BDTOPO a été utilisée comme base de données de référence.

Enfin, la même délimitation du contexte géographique (urbain, rural et de montagne) sur le département des Pyrénées-Atlantiques que celle proposée dans la section 4.2.3.1 a été utilisée pour réaliser les comparaisons.

Résultats des comparaisons

La table 4.8 présente les résultats des comparaisons de longueurs calculées sur les trois bases de données exploitées, dans les trois contextes géographiques étudiés.

	Montagne		**Rural**		**Urbain**	
	Moy.(%)	ET.(%)	Moy.(%)	ET.(%)	Moy.(%)	ET.(%)
BDCARTO	0,85	8,11	0,53	5,95	0,59	7,09
TOP100	5,62	12,06	3,07	6,92	2,64	9,52
CR250	6,67	11,76	4,03	7,23	5,66	13,34

Table 4.8 – Résultat des comparaisons de longueurs sur l'ensemble du réseau routier des Pyrénées-Atlantiques

Erreur moyenne Les résultats des comparaisons de longueurs sur l'ensemble du réseau routier des Pyrénées-Atlantiques confirment tout d'abord les conclusions observées dans l'étude de Girres (2011a). On observe en effet que l'erreur de longueur relative moyenne dépend :
– de l'échelle de représentation des données généralisées,
– du contexte géographique des objets.
Si les erreurs de longueur relatives moyennes présentées ici diffèrent légèrement de celles calculées lors des comparaisons par sections (cf. section 4.2.3.1), les tendances observées demeurent dans l'ensemble identiques. Ces différences s'expliquent par l'échantillon de tronçons de route utilisé, qui n'est pas le même dans les deux études.

Dispersion des erreurs La quantification de la dispersion des erreurs de longueur est fondamentale dans la cadre de l'élaboration du modèle statistique, car en plus d'estimer une erreur moyenne de longueur, nous souhaitons évaluer son imprécision, à l'aide d'un intervalle de confiance.
Les résultats proposés en table 4.8 montrent tout d'abord que les valeurs d'écart-type augmentent lorsque l'échelle de représentation baisse, dans chacun des contextes géographiques. Ceci signifie que plus l'échelle de représentation diminue, plus les mesures de longueur calculées sont imprécises. Egalement, nous observons que, de manière générale, les écart-types de longueur sont plus importants en zone de montagne et en zone urbaine qu'en zone rurale. Ceci tend à confirmer que dans ces zones, où les effets de la généralisation cartographique peuvent être importants lorsque l'échelle diminue, les mesures de longueur sont également plus imprécises.

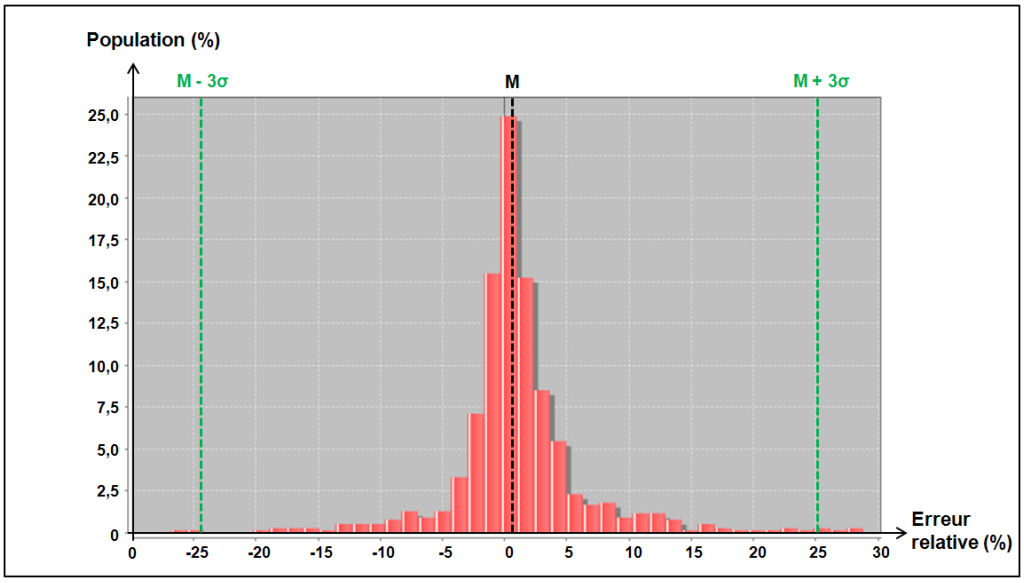


Figure 4.65 – Distribution des erreurs de longueur sur le réseau routier de la BDCARTO localisé en zone de montagne

L'histogramme de la figure 4.65 présente la distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO localisés en zones de montagne. Cette figure montre bien que la distribution des erreurs est centrée autour de sa moyenne M (égale à 0,85 % de la longueur généralisée). L'écart-type des erreurs relatives σ est de 8,11%.

Afin d'estimer l'imprécision des erreurs de longueur relatives sur les données généralisées, une étude du pourcentage d'erreurs relatives inscrites dans un intervalle de confiance de σ, 2σ, et 3σ autour de l'erreur moyenne est proposé (cf. table 4.9).

	Montagne			**Rural**			**Urbain**		
	σ	2σ	3σ	σ	2σ	3σ	σ	2σ	3σ
BDCARTO	88,2	91,2	97,9	94,4	97,9	98,4	92,4	96,9	98,2
TOP100	84,1	95,0	98,1	90,6	96,9	98,4	87,8	95,6	97,8
CR250	84,2	94,0	97,9	90,6	97,2	98,7	89,0	95,3	97,3

Table 4.9 – Pourcentage d'erreurs de longueur relatives comprises dans un intervalle de σ, 2σ, et 3σ autour de l'erreur moyenne relative

Ces résultats montrent que, toutes bases de données et contextes géographiques confondus, au minimum 84,1% des erreurs sont comprises dans un intervalle de confiance de σ autour de la moyenne. Ceci est observable dans l'histogramme de la figure 4.65, qui montre une distribution des erreurs de longueur relatives très concentrée autour de sa moyenne. Egalement, ces résultats montrent que même avec un intervalle de confiance de 3σ autour de la moyenne, il n'est pas possible d'inscrire la totalité des erreurs de longueur potentielles, puisqu'au maximum 97,8% d'entre elles y sont comprises.

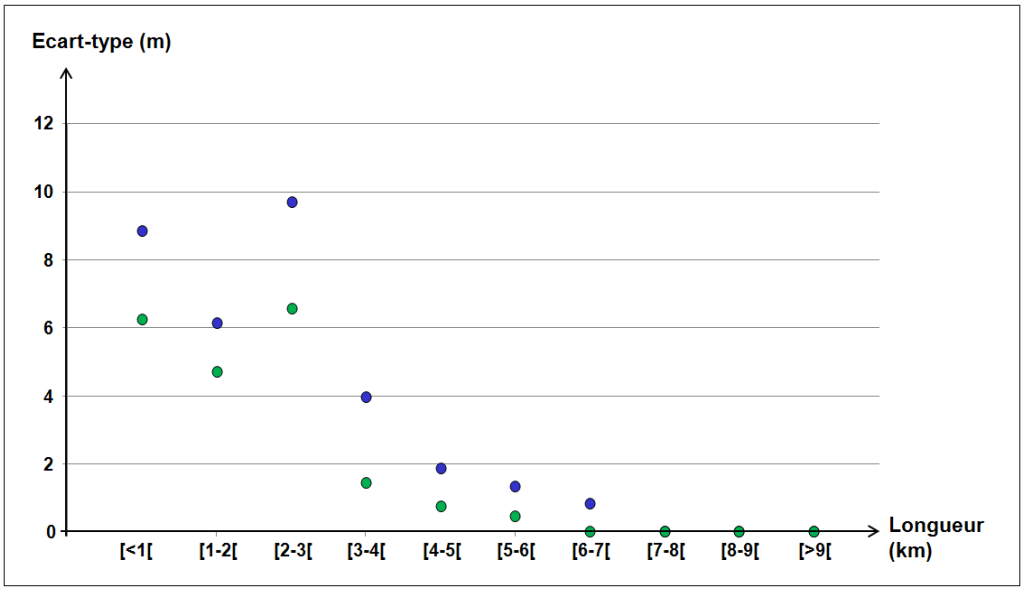


Figure 4.66 – Imprécision de l'erreur relative de longueur des tronçons de route de la BDCARTO en zones rurale (en vert) et de montagne (en bleu)

Cependant, comme le montre l'histogramme de la figure 4.66, plus la longueur des tronçons comparés augmente, plus l'écart-type des erreurs de longueur relatives diminue. Ceci signifie que lorsque l'évaluation porte sur une polyligne de faible longueur, l'imprécision de l'erreur relative de longueur est importante ($\pm\sigma$, voire $\pm 2\sigma$ à $\pm 3\sigma$ autour de l'erreur moyenne). Au contraire, lorsque la longueur de la polyligne évaluée est importante, l'imprécision de l'erreur relative de longueur devient presque nulle, ce qui signifie que l'erreur relative de longueur tend vers l'erreur moyenne.

Ainsi, dans le cadre de l'élaboration du modèle statistique, il est important de s'interroger sur le calibrage de l'intervalle de confiance autour de l'erreur moyenne estimée, en fonction de la longueur étudiée. En effet, si l'erreur moyenne relative de longueur n'est pas sensible à la longueur étudiée, ce n'est pas le cas de l'imprécision de l'erreur de longueur. Dans le cas d'objets présentant une longueur importante, on peut même s'interroger sur l'intérêt de formuler un intervalle de confiance, même de $\pm\sigma$, puisqu'il aurait pour conséquence de rendre la longueur estimée beaucoup plus imprécise. Nous chercherons à répondre à ces questions lors de la validation du modèle d'estimation (cf. section 4.2.3.3), qui nous permettra d'étudier la pertinence d'un intervalle de confiance sur les longueurs estimées.

Dans l'état actuel des choses, nous pouvons déjà proposer une formalisation du modèle statistique d'estimation de l'impact de la généralisation sur les mesures de longueur calculées à partir d'un réseau routier.

Modèle statistique

Les résultats des comparaisons entre données généralisées et données de référence nous ont permis pour l'instant de déterminer :
– l'erreur moyenne relative de longueur,
– l'imprécision de l'erreur relative de longueur.
Ces indicateurs sont relatifs à une classe d'objet (les routes en l'occurrence), une échelle de représentation et un contexte géographique donné. A partir de ces résultats, nous pouvons donc élaborer un modèle statistique permettant d'estimer la longueur d'un tronçon de route de référence à partir d'une longueur généralisée pour chacune des trois bases de données généralisées expérimentées, dans chaque contexte géographique.

Nous présentons ci-dessous, pour chaque échelle étudiée, les 3 fonctions permettant d'estimer une longueur moyenne de référence, comprise au sein d'un intervalle de confiance de $\pm\sigma$ (correspondant à une confiance de 84,1%), à partir d'une mesure de longueur effectuée sur des tronçons de route généralisés. Dans ces équations, L_{Gen} correspond à la longueur généralisée mesurée et L_{Ref} à la longueur de référence estimée.

Estimation de l'imprécision de longueur des routes :

$$\begin{cases} L_{Ref}(Routes, 50000, Montagne) = 1,0085 * L_{Gen} \quad (\pm 0,0811 * L_{Gen}) \\ L_{Ref}(Routes, 50000, Rural) = 1,0053 * L_{Gen} \quad (\pm 0,0595 * L_{Gen}) \\ L_{Ref}(Routes, 50000, Urbain) = 1,0059 * L_{Gen} \quad (\pm 0,0709 * L_{Gen}) \end{cases} \tag{4.12}$$

$$\begin{cases} L_{Ref}(Routes, 100000, Montagne) = 1,0562 * L_{Gen} \quad (\pm 0,1206 * L_{Gen}) \\ L_{Ref}(Routes, 100000, Rural) = 1,0307 * L_{Gen} \quad (\pm 0,0692 * L_{Gen}) \\ L_{Ref}(Routes, 100000, Urbain) = 1,0264 * L_{Gen} \quad (\pm 0,0952 * L_{Gen}) \end{cases} \tag{4.13}$$

$$\begin{cases} L_{Ref}(Routes, 250000, Montagne) = 1,0667 * L_{Gen} \quad (\pm 0,1176 * L_{Gen}) \\ L_{Ref}(Routes, 250000, Rural) = 1,0403 * L_{Gen} \quad (\pm 0,0723 * L_{Gen}) \\ L_{Ref}(Routes, 250000, Urbain) = 1,0566 * L_{Gen} \quad (\pm 0,1334 * L_{Gen}) \end{cases} \tag{4.14}$$

Les tests de comparaisons de longueurs ainsi réalisés sur le réseau routier nous ont permis d'élaborer un modèle statistique permettant d'évaluer l'impact de la généralisation cartographique sur les mesures de longueur, conformément à nos hypothèses de départ, à savoir que l'échelle de représentation et le contexte géographique influent sur la longueur des objets généralisés. Cependant, le modèle ainsi proposé nécessite d'être validé dans le cadre d'une application concrète, qui nous permettra également de calibrer l'intervalle de confiance utilisé pour estimer l'imprécision de la mesure estimée.

4.2.3.3 Validation du modèle

Principe de validation Afin de valider le modèle statistique, une expérimentation a été menée. Cette expérimentation repose sur des calculs d'itinéraires aléatoires réalisés sur une base de données généralisée. Les itinéraires ainsi calculés sont ensuite comparés avec leurs homologues d'une base de données de référence. L'objectif de l'expérimentation est de voir si l'application du modèle statistique d'estimation de la longueur de référence moyenne et des bornes de l'intervalle de confiance correspondent à la longueur calculée sur le réseau routier de référence.

Méthodologie Pour mener cette expérimentation, dix itinéraires, d'une longueur minimale de 20 kilomètres, ont été calculés aléatoirement sur le réseau routier des bases de données TOP100 et de la BDCARTO, dans le département du Var (afin de valider le modèle développé sur le département des Pyrénées-Atlantiques). Pour chaque itinéraire aléatoire tiré, les itinéraires homologues de référence, extraits du réseau routier de la BDTOPO, sont détectés par appariement automatique, en appliquant l'algorithme de Mustière et Devogèle (2008), puis validés visuellement.

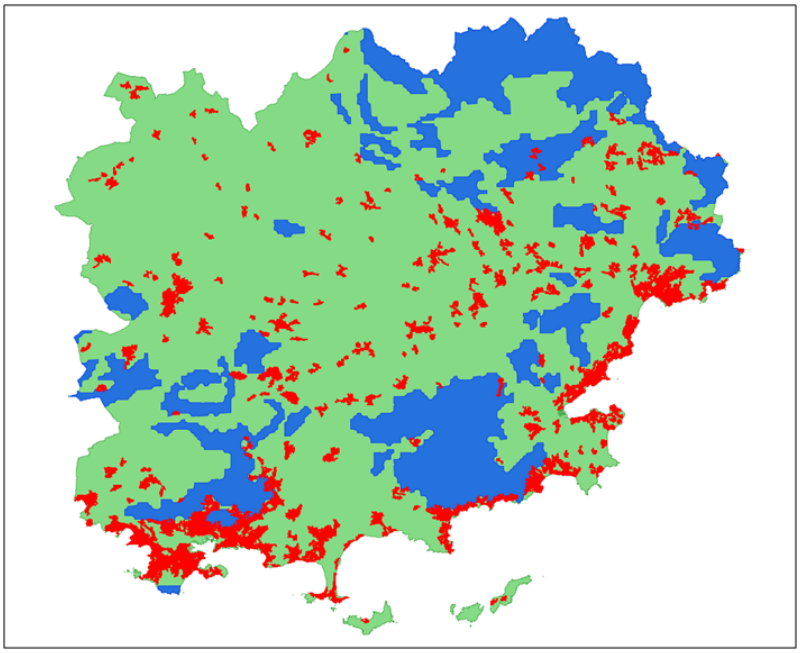


Figure 4.67 – Délimitation des zones de montagne (en bleu), urbaines (en rouge) et rurales (en vert) dans le département du Var

Contexte géographique Afin d'appliquer les fonctions d'estimation du modèle statistique, chaque itinéraire est ensuite découpé en fonction du contexte géographique au sein duquel il est localisé. Pour délimiter les zones de montagne, la méthodologie présentée en section 4.2.2.2 est appliquée, à partir de courbes de niveau (d'une équidistance de 40 mètres) extraites du modèle numérique d'élévation SRTM. Pour les zones urbaines, les classes 111 (*Continuous Urban Areas*) et 112 (*Discontinuous Urban Areas*) de la base de données *CorineLandCover* ont été extraites. Les zones rurales correspondent aux zones qui ne sont ni urbaines ni de montagne (cf. figure 4.67).

Afin de n'attribuer qu'un seul contexte géographique à chaque portion d'itinéraire, un ordre de priorité est appliqué lorsque une même portion d'itinéraire est localisée à la fois en zone de montagne et en zone urbaine. Dans ce cas de figure, l'itinéraire est considéré comme étant localisé en zone urbaine.

Application aux données TOP100

La figure 4.68 présente les dix itinéraires tirés aléatoirement sur le réseau routier des données TOP100 dans le département du Var.

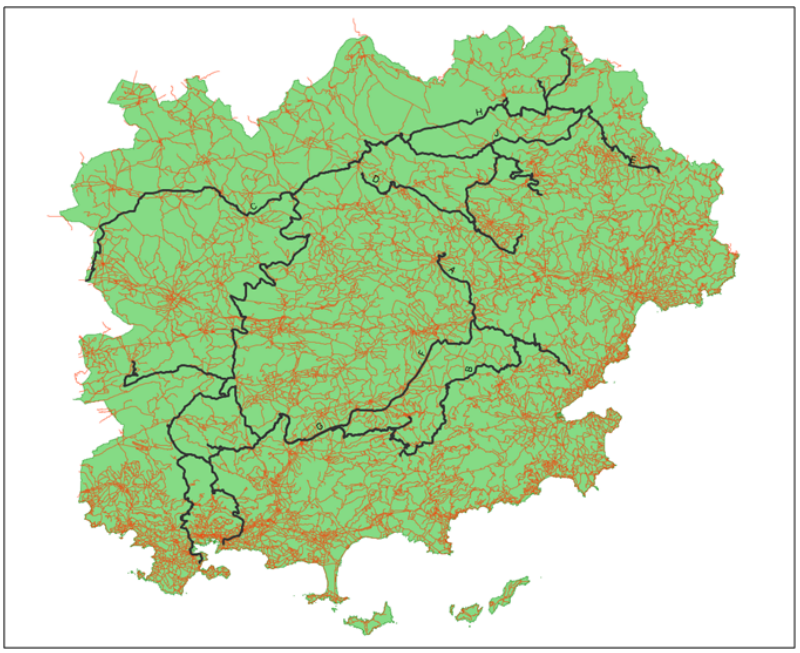


Figure 4.68 – Les dix itinéraires tirés aléatoirement sur le réseau routier des données TOP100 dans le département du Var

La table 4.10 présente la structure des dix itinéraires tirés aléatoirement à partir du réseau routier des données TOP100. Nous observons que malgré leurs longueurs hétérogènes, les dix itinéraires présentent tous des sections localisées en zones de montagne, rurale, et urbaine, bien qu'en grande partie de ces dernières soient très courtes.
La table 4.11 présente quant à elle les longueurs des dix itinéraires extraits des données TOP100, en comparaison avec leurs homologues de référence extraits de la BDTOPO, ainsi que les erreurs associées.

	Longueur (en km)			
	Total	**Montagne**	**Rural**	**Urbain**
A	37,58	0,11	30,46	7,01
B	49,32	18,96	29,57	0,78
C	116,22	26,42	87,48	2,31
D	38,23	2,30	34,60	1,32
E	44,29	7,42	33,32	3,55
F	64,50	17,94	45,53	1,03
G	84,96	34,53	44,60	5,83
H	176,88	37,97	128,59	10,32
I	78,41	34,85	42,87	0,70
J	61,57	19,91	41,56	0,10

Table 4.10 – Contexte géographique des itinéraires extraits des données TOP100

	TOP100 (km)	**BDTOPO** (km)	**Erreur** (km)	**Erreur** (%)
A	37,58	39,30	1,72	4,58
B	49,32	54,40	5,07	10,29
C	116,22	119,81	3,59	3,09
D	38,23	40,19	1,96	5,14
E	44,29	45,90	1,61	3,64
F	64,50	66,09	1,59	2,46
G	84,96	88,25	3,29	3,87
H	176,88	183,16	6,28	3,55
I	78,41	81,59	3,27	4,17
J	61,57	63,13	1,55	2,52

Table 4.11 – Erreurs de longueur des itinéraires extraits des données TOP100

Erreurs de longueur On observe que la longueur des dix itinéraires tirés aléatoirement à partir des données TOP100 est systématiquement sous-estimée, en comparaison avec les longueurs des itinéraires de référence, extraits de la BDTOPO. La sous-estimation moyenne est de l'ordre de 4,3% par rapport à la longueur de référence. L'itinéraire présentant la plus grande erreur est l'itinéraire B, avec une longueur sous-estimée de 10,3% (cf. figure 4.69).

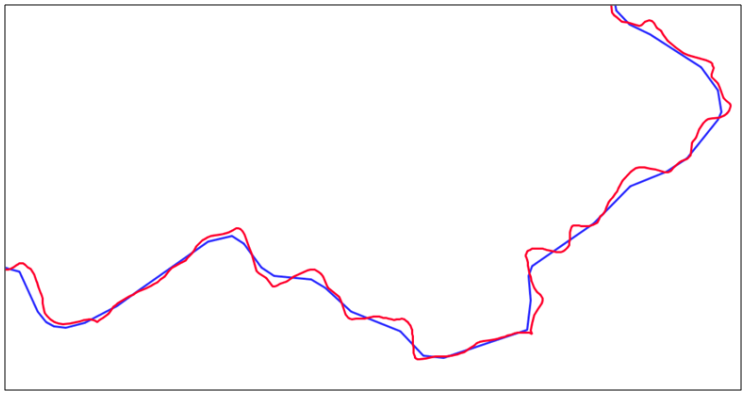


Figure 4.69 – L'itinéraire B extrait de la BDTOPO (en rouge) et des données TOP100 (en bleu)

Application du modèle statistique Le modèle statistique élaboré dans la section précédente est appliqué aux dix itinéraires tirés aléatoirement à partir du réseau routier des données TOP100. L'application du modèle a permis d'estimer une longueur de référence moyenne, ainsi que les bornes inférieures et supérieures d'un intervalle de confiance. Cet intervalle de confiance a été testé en utilisant les valeurs de σ, 2σ et 3σ autour de la longueur moyenne estimée. L'ensemble des estimations est présenté en table 4.12.

	Longueurs(km)								
	Initiales		**Estimées**						
	BDT	**TOP100**	-3σ	-2σ	$-\sigma$	**Moy**	$+\sigma$	$+2\sigma$	$+3\sigma$
A	**39,30**	37,58	30,34	33,13	35,92	**38,71**	41,49	44,28	47,07
B	**54,40**	49,32	38,09	42,50	46,91	**51,32**	55,72	60,13	64,54
C	**119,81**	116,22	92,07	101,53	110,99	**120,45**	129,91	169,37	148,83
D	**40,19**	38,23	31,06	33,86	36,65	**39,45**	42,25	45,05	47,85
E	**45,91**	44,30	35,21	38,75	42,29	**45,83**	49,36	52,90	56,44
F	**66,09**	64,50	50,70	56,11	61,53	**66,94**	72,35	77,76	83,18
G	**88,25**	84,96	65,00	72,81	80,62	**88,42**	96,23	104,03	111,84
H	**183,16**	176,88	139,85	154,31	168,77	**183,23**	197,69	212,15	226,61
I	**81,69**	78,41	60,00	67,23	74,47	**81,71**	88,94	96,18	103,41
J	**63,13**	61,57	48,11	53,40	58,68	**63,97**	69,26	74,55	78,83

Table 4.12 – Longueurs de référence estimées par application du modèle statistique sur les itinéraires extraits des données TOP100

Résultats de l'estimation Les résultats de l'application du modèle statistique montrent que l'estimation de la longueur moyenne de référence à partir de la longueur généralisée constitue une approximation réaliste. En effet, la longueur moyenne estimée sur l'ensemble des itinéraires présente une erreur significativement moins importante par rapport à la mesure initiale. Alors que l'erreur de longueur relative moyenne calculée est initialement de 4,3% sur l'ensemble des 10 itinéraires, l'estimation permet de réduire cette erreur à 0,6%. Seul l'itinéraire B, qui présente une erreur initiale de longueur de 10,29% présente encore une erreur estimée significative (environ 6%).

Etant donné que l'estimation d'une longueur de référence moyenne ne permet pas dans tous les cas de figure d'approximer la longueur réelle, il est intéressant de produire un intervalle de confiance autour de la longueur estimée. Les résultats de l'estimation de l'intervalle de confiance avec les valeurs de σ, 2σ et 3σ autour de la longueur moyenne estimée montrent que l'application de la valeur σ est jugée largement suffisante. Cette calibration permet d'encadrer la longueur de référence pour tous les itinéraires étudiés, notamment l'itinéraire B. Cependant, nous observons que cette calibration a pour conséquence de rendre l'estimation beaucoup plus imprécise dans la grande majorité des cas. Ces résultats nous encouragent donc à ne retenir que la valeur moyenne estimée, et à ne proposer un intervalle de confiance de $\pm\sigma$ qu'à titre indicatif.

Application à la BDCARTO

Une expérimentation similaire à la précédente a été réalisée à partir du réseau routier de la BDCARTO. La figure 4.70 présente les dix itinéraires tirés aléatoirement sur le réseau routier de la BDCARTO.

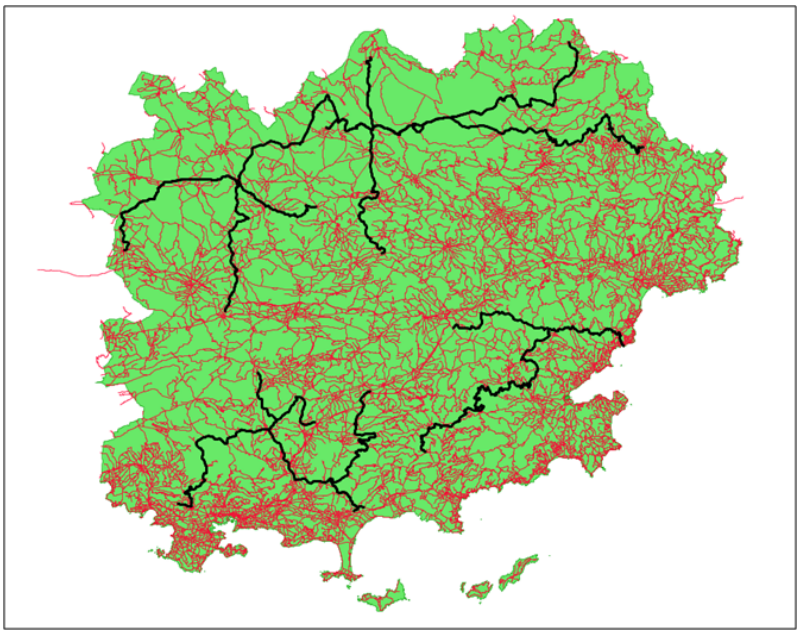


Figure 4.70 – Les dix itinéraires tirés aléatoirement sur le réseau routier de la BDCARTO dans le département du Var

	Longueur (en km)			
	Total	**Montagne**	**Rural**	**Urbain**
A	142,82	21,00	118,96	2,85
B	50,09	22,59	22,58	4,92
C	62,53	12,86	49,08	0,60
D	28,43	0	26,41	2,02
E	44,09	3,15	39,11	1,82
F	41,96	0	41,96	0
G	45,70	0	44,40	1,30
H	66,64	50,47	15,38	0,78
I	37,77	2,82	28,71	0,61
J	43,10	3,85	34,57	3,85

Table 4.13 – Structure des itinéraires extraits de la BDCARTO selon le contexte géographique

La table 4.13 présente la structure des dix itinéraires en fonction de leur contexte géographique, tandis que la table 4.14 présente les longueurs des dix itinéraires extraits de la BDCARTO, en comparaison avec leurs homologues de référence extraits de la BDTOPO, ainsi que les erreurs associées.

	BDCARTO (km)	**BDTOPO** (km)	**Erreur** (km)	**Erreur** (%)
A	142,82	143,47	0,66	0,46
B	50,09	51,36	1,27	2,53
C	62,53	63,02	0,48	0,77
D	28,43	28,68	0,25	0,89
E	44,09	44,46	0,37	0,84
F	41,96	42,66	0,70	1,67
G	45,70	46,23	0,52	1,14
H	66,64	68,02	1,38	2,08
I	37,77	37,96	0,19	0,51
J	43,10	43,10	0	0

Table 4.14 – Erreurs de longueur entre les itinéraires extraits de la BDCARTO et leurs homologues de la BDTOPO

Erreurs de longueur Tout comme pour les itinéraires extraits des données TOP100, on observe que la longueur des dix itinéraires tirés aléatoirement à partir de la BDCARTO est systématiquement sous-estimée par rapport aux itinéraires de référence extraits de la BDTOPO, à l'exception de l'itinéraire J, dont les longueurs sont égales sur les deux bases. La sous-estimation moyenne est de l'ordre de 1,21% par rapport à la longueur de référence.

Application du modèle statistique Le modèle statistique élaboré dans la section précédente est appliqué aux dix itinéraires tirés aléatoirement du réseau routier de la BDCARTO. L'application du modèle permet d'estimer une longueur de référence moyenne, ainsi que les bornes inférieures et supérieures d'un intervalle de confiance, en ne conservant que la valeur σ. L'ensemble des estimations sont présentées en table 4.15.

	Longueurs(km)				
	Initiales		**Estimées**		
	BDT	**BDC**	$-\sigma$	**Moy**	$+\sigma$
A	**143,47**	142,82	134,66	**143,64**	152,63
B	**51,36**	50,09	46,91	**50,43**	53,96
C	**63,02**	62,53	58,90	**62,91**	66,91
D	**44,46**	44,09	41,62	**44,33**	47,04
E	**42,66**	41,96	39,68	**42,18**	44,68
F	**46,23**	45,70	43,21	**45,95**	48,68
G	**68,02**	66,64	62,09	**67,15**	72,22
H	**37,96**	37,77	35,60	**37,98**	40,36
I	**43,10**	43,10	40,64	**43,34**	46,04
J	**63,13**	61,57	48,11	**63,97**	69,26

Table 4.15 – Longueurs de référence estimées par application du modèle statistique sur les itinéraires extraits de la BDCARTO

Résultats de l'estimation Les résultats de l'application du modèle statistique sur les itinéraires extraits de la BDCARTO rejoignent ceux observés sur les données TOP100. La longueur moyenne estimée sur l'ensemble des itinéraires présente une erreur moins importante par rapport à la mesure initiale.

L'erreur moyenne relative calculée à partir des longueurs moyennes estimées est passée à 0,49% alors qu'elle était de 1,21% sur les longueurs intiales. Egalement, nous observons que l'application d'un intervalle de confiance de $\pm\sigma$ rend l'estimation de la longueur de référence beaucoup plus imprécise. Ainsi, dans le cas des objets de la BDCARTO, nous ne jugeons pas pertinent d'exploiter cet intervalle de confiance.

Conclusions Les résultats de cette expérimentation permettent donc de valider, dans le cas du réseau routier, le modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures de longueur, ainsi que la méthodologie globale permettant son élaboration. Les fonctions d'estimation ainsi produites et validées seront donc intégrées au sein du modèle général d'estimation de l'imprécision des mesures géométriques pour les objets routiers. Il aurait été intéressant de développer un modèle statistique similaire pour d'autres classes d'objets, mais ceci ne pu être entrepris dans le cadre de ce travail de thèse. Après avoir étudié l'impact de la généralisation cartographique sur les mesures de longueur, la section suivante aura pour objectif de traiter de cet impact sur les mesures de surface.

Synthèse Section Cette section a permis d'élaborer un modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures de longueur, en utilisant l'exemple d'un réseau routier. Pour concevoir ce modèle statistique, nous avons tout d'abord montré que l'erreur impliquée par la généralisation cartographique n'est pas sensible à la longueur étudiée. Cette observation a permis de mettre au point un processus de comparaison s'affranchissant d'une normalisation géométrique par sections de longueurs de référence identiques. Les comparaisons ainsi réalisées ont permis d'élaborer un modèle permettant d'estimer l'erreur moyenne et l'imprécision relative de longueur en fonction de l'échelle de représentation des données et du contexte géographique. Le modèle ainsi développé a ensuite été validé dans le cadre d'une expérimentation basée sur le tirage aléatoire d'itinéraires routiers.

4.2.4 Impact de la généralisation sur les mesures de surface

Cette section a pour objectif de proposer un modèle d'estimation de l'impact de la généralisation cartographique sur les mesures de surface. Comme pour les mesures de longueur, l'élaboration de ce modèle se base sur les résultats de comparaisons de surfaces entre objets homologues extraits de bases de données généralisées et d'une base de données de référence. Cependant, contrairement à l'étude des erreurs de longueur sur les objets linéaires, nous considérons que l'étude des erreurs de surface implique de prendre en compte les propriétés des objets surfaciques.

En effet, nous pouvons différencier les propriétés des objets surfaciques en deux grandes catégories :

- les objets représentant une partition du monde réel (les limites communales par exemple),
- les objets ne représentant pas une partition du monde réel (les surfaces d'eau par exemple)

Selon les propriétés des objets étudiés, différents processus de généralisation peuvent être appliqués, ou pas (cf. figure 4.71).

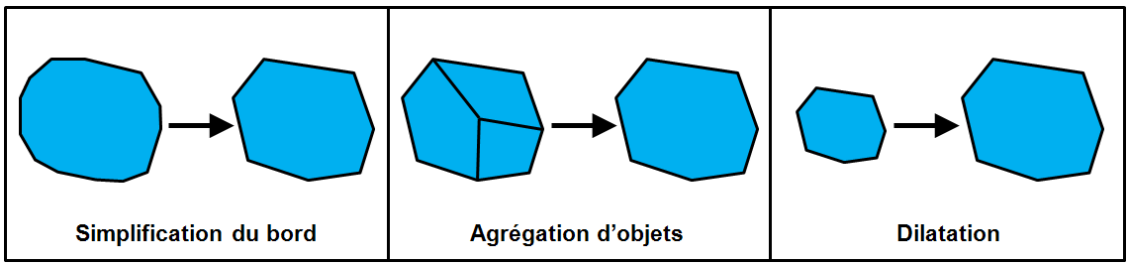


Figure 4.71 – Les opérations de simplification de bord, d'agrégation et de dilatation d'objets surfaciques

Par exemple, pour les objets représentant une partition de l'espace, les opérations classiques de généralisation reposent sur une simplification du contour des objets, par application d'algorithmes de filtrage de la géométrie. Pour les objets ne constituant pas une partition de l'espace, les opérations de généralisation potentiellement applicables sont beaucoup plus variées (simplification des contours, agrégation d'objets, ou encore dilatation d'objets). On sait notamment que pour cette seconde catégorie d'objets surfaciques, les objets de petites surfaces risquent davantage de subir les effets de la généralisation cartographique, afin de conserver leur lisibilité sur la carte.

Nous proposons donc dans cette section d'étudier l'impact de la généralisation cartographique sur les mesures de surface en différenciant ces deux catégories d'objets surfaciques. Pour cela, deux classes d'objets seront utilisées : les limites communales et les surfaces d'eau. Nous présenterons dans un premier temps la méthodologie de comparaison utilisée, puis exposerons les résultats des comparaisons.

4.2.4.1 Méthodologie de comparaison

Nous présentons dans cette section le principe de comparaison mis en œuvre afin d'estimer l'impact de la généralisation cartographique sur les mesures de surface, ainsi que les bases de données utilisées.

Principe de comparaison

Hypothèses Afin de réaliser les comparaisons entre objets surfaciques, nous ne sommes pas partis de l'hypothèse qu'il pouvait y avoir un impact du contexte géographique sur les mesures de surface, contrairement aux objets linéaires, et en particulier les routes, pour lesquelles des modèles spécifiques de généralisation sont développés selon le contexte géographique (Touya, 2011). Nous considérons ici que la répartition des erreurs des surface est homogène sur l'ensemble de la zone d'étude. Ainsi, l'objectif de cette étude consistera à estimer l'erreur de surface en fonction de la classe d'objet étudiée, et de son échelle de représentation (cf. équation 4.15).

$$Erreur = f(Echelle, Classe) \tag{4.15}$$

Appariement des données Pour réaliser les comparaisons sur les objets surfaciques, le processus de pré-traitement des données est également largement simplifié par rapport à celui proposé pour la comparaison des objets linéaires. Ainsi, l'algorithme développé par Bel Hadj Ali (2001) a été utilisé afin d'apparier automatiquement les objets surfaciques homologues entre les bases de données généralisées et la base de données de référence.

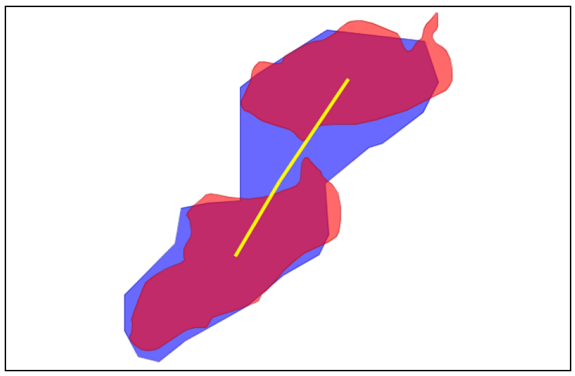


Figure 4.72 – Résultat d'un appariement de surfaces par un lien 1-n, avec l'algorithme de Bel Hadj Ali (2001)

L'algorithme de Bel Hadj Ali (2001), qui repose principalement sur la distance surfacique, suppose que les polygones homologues à apparier se superposent. Cependant ceci n'est pas toujours le cas, ce qui peut générer une perte d'information en termes d'objets à comparer. Cet algorithme offre par contre la possibilité de gérer les liens d'appariement 1-n, ce qui est intéressant dans le cas où un objet généralisé correspond à plusieurs objets de référence (ou l'inverse).

Pré-traitement des données Une seule étape de pré-traitement des géométries est mobilisée dans le processus de comparaison des objets surfaciques. Cette étape concerne uniquement les objets dont le résultat d'appariement suit un lien 1-n. Il est nécessaire dans ce cas de procéder à une fusion des géométries, afin de disposer de liens d'appariement 1-1 (cf. figure 4.73), et ainsi permettre des comparaisons de surfaces.

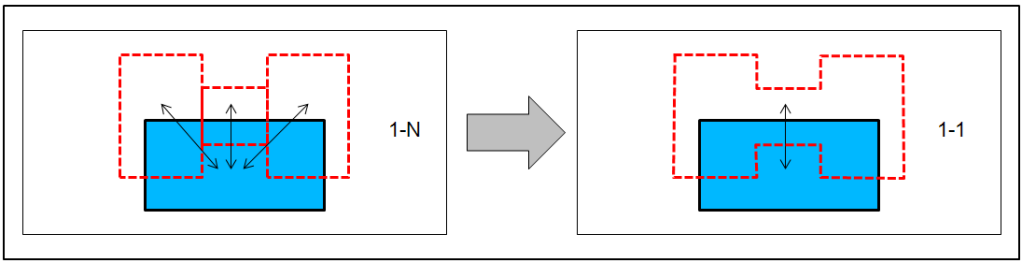


Figure 4.73 – Fusion de polygones adjacents pour la comparaison de surfaces

Normalisation des comparaisons de surfaces Contrairement à la méthodologie de comparaison des objets linéaires, qui nécessite un découpage des extrémités de polylignes afin d'éviter les erreurs de longueur aberrantes, aucun traitement n'est appliqué sur la géométrie des objets surfaciques avant de procéder aux comparaisons de surfaces.

Afin d'élaborer un modèle statistique d'estimation de l'imprécision des mesures de surface, nous proposons d'estimer une **erreur de surface relative** suite aux comparaisons. Ainsi, les erreurs de surface observées entre objets généralisés et objets de référence sont rapportées à la surface de l'objet généralisé. Cependant, les résultats en termes d'**erreur de surface absolue** sont également exploités.

Bases de données utilisées

La principale difficulté pour réaliser ces comparaisons réside dans l'identification de classes d'objets pertinentes, permettant des comparaisons de surfaces à différentes échelles. En effet, plusieurs raisons peuvent empêcher de rendre comparables des classes d'objets à différentes échelles.

Tout d'abord, le processus de généralisation cartographique peut se manifester par une généralisation de modèle, qui implique un changement dans la modélisation du monde réel au niveau du schéma de données. Par exemple, des bâtiments représentés au 1 :10.000 peuvent être transformés en zone bâtie 1 :50.000, comme c'est par exemple le cas entre la BDTOPO et la BDCARTO. Il a donc été nécessaire de trouver des objets n'étant pas impactés par la généralisation de modèle.

Egalement, il est délicat de procéder à des comparaisons de surfaces en se basant sur des objets présentant des limites floues, comme les forêts par exemple, ou laissant une place trop importante à l'interprétation lors de la saisie par un opérateur. De plus, il était important de pouvoir réaliser des comparaisons en prenant en compte les propriétés des objets, c'est-à-dire ceux représentant une partition de l'espace, et ceux n'en représentant pas.

Classes d'objets Deux classes d'objets ont donc pu être exploitées pour réaliser les comparaisons :
- les limites communales, qui représentent une partition de l'espace, et qui portent en plus l'intérêt d'être très souvent utilisées afin de réaliser des mesures de surface (pour des calculs de densité par exemple),
- les surfaces d'eau, qui ne représentent pas une partition de l'espace. Bien que pouvant présenter des limites floues (du fait de leurs variations saisonnières). Cette classe d'objet offre l'intérêt d'être facilement identifiable par un opérateur, et ne suppose donc pas d'interprétations trop importantes.

Bases de données Pour réaliser les comparaisons, uniquement deux bases de données généralisées ont été exploitées, puisque les données utilisées pour produire la Carte Régionale (au 1 :250.000) ne sont pas pourvues de ces deux classes d'objets :
- la BDCARTO, produite à l'échelle 1 :50.000.
- les données TOP100, permettant une représentation cartographique au 1 :100.000.

La BDTOPO a été utilisée comme base de données de référence afin de réaliser ces comparaisons.

4.2.4.2 Elaboration du modèle statistique

Nous présenterons dans cette section les résultats des comparaisons permettant d'élaborer un modèle statistique sur les objets représentant une partition de l'espace (les limites communales), puis sur les objets n'en représentant pas une (les surfaces d'eau).

Application aux limites communales

Un échantillon de 382 emprises communales localisées dans le département des Pyrénées-Atlantiques a été exploité (cf. figure 4.74) pour réaliser les comparaisons de surfaces.

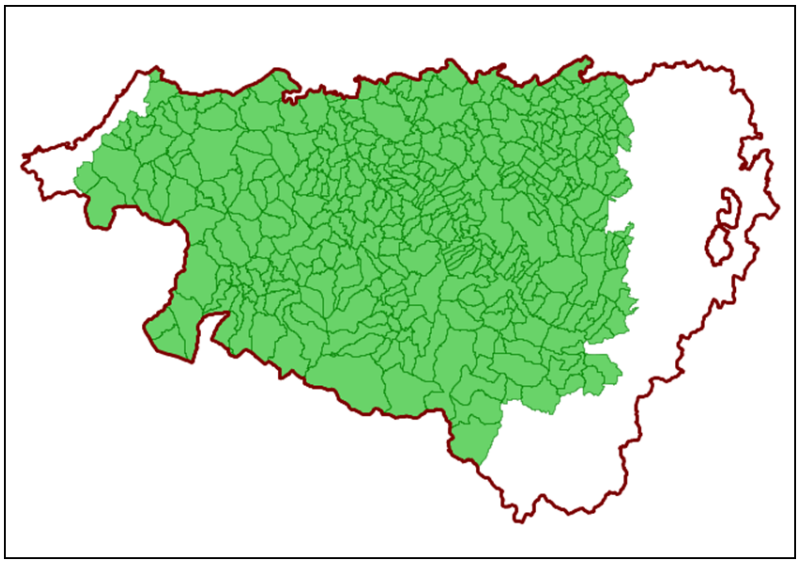


Figure 4.74 – Localisation des 382 emprises communales utilisées pour le test de comparaisons de surfaces

A l'instar de l'étude sur les erreurs de longueur, nous présenterons successivement les résultats des comparaisons en termes d'erreur moyenne, et de dispersion des erreurs (cf. table 4.16).

	Surface		**Périmètre**	
	Moy.(%)	ET.(%)	Moy.(%)	ET.(%)
BDCARTO	0,00	0,83	1,98	2,33
TOP100	0,00	0,94	1,71	2,27

Table 4.16 – Résultat des comparaisons de surfaces sur les limites communales

Erreur moyenne Ces résultats montrent tout d'abord, que contrairement aux mesures de longueur, l'impact de la généralisation cartographique sur les limites communales ne se traduit pas par une diminution systématique des mesures géométriques de surface. En effet, pour les deux comparaisons réalisées au 1 :50.000 et au 1 :100.000 sur les surfaces communales, la moyenne relative des erreurs de surface est considérée comme nulle. Le graphique 4.75 illustre la relation entre les surfaces des communes de la BDTOPO et leurs homologues extraites des données TOP100.

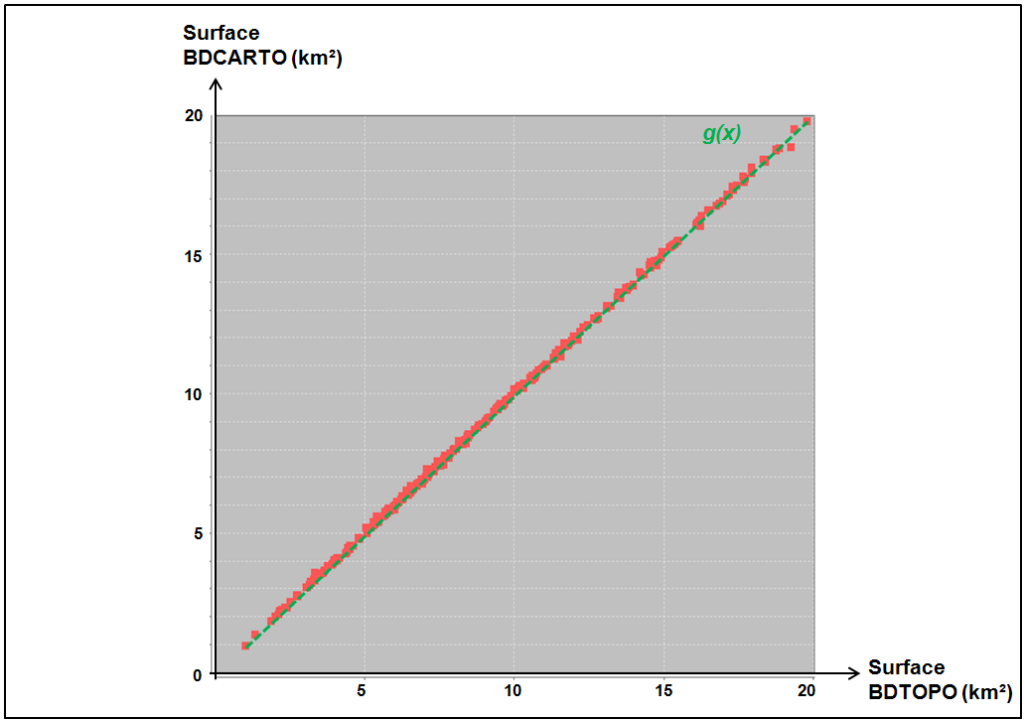


Figure 4.75 – Relation entre les surfaces des communes homologues de la BDTOPO et des données TOP100

Bien que l'erreur moyenne de surface soit nulle, peu importe l'échelle de représentation, l'erreur de périmètre, qui elle est systématiquement positive, montre que l'impact de la généralisation cartographique affecte les mesures géométriques calculées à partir des limites communales. En effet, les périmètres des limites communales généralisées au 1 :50.000 et au 1 :100.000 sont respectivement sous-estimés en moyenne de 1,98% et de 1,71% par rapport au périmètre de référence de la BDTOPO.

Ces résultats s'expliquent principalement par le processus de simplification de contour, utilisé pour généraliser les limites communales à différentes échelles, qui a pour effet de réduire le périmètre du polygone, tout en préservant sa surface.

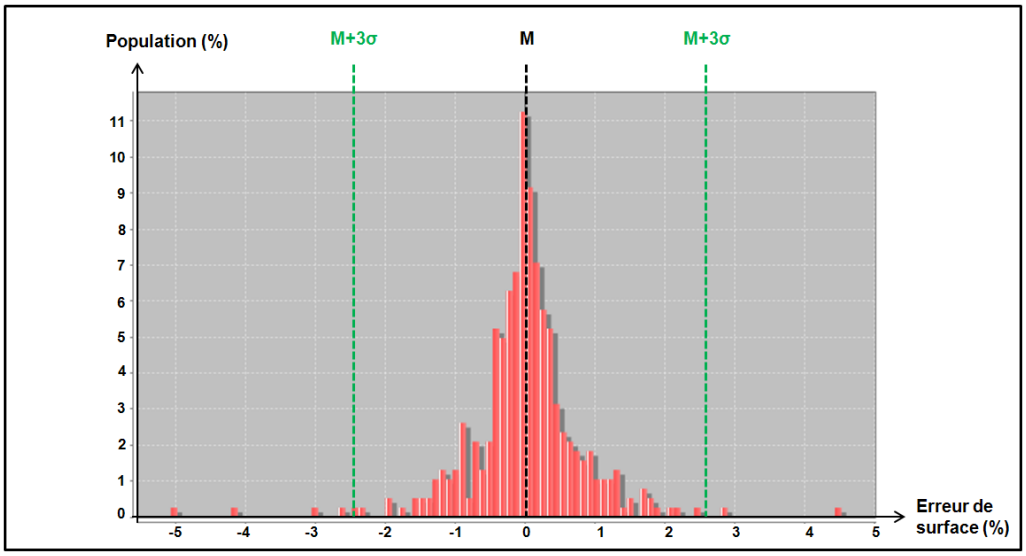


Figure 4.76 – Distribution des erreurs de surface relatives des limites communales de la BDCARTO

Dispersion des erreurs L'étude des écarts-types, permettant d'estimer l'imprécision des erreurs de surface, montre tout d'abord que les erreurs relatives de surface présentent une faible amplitude. En effet, au 1 :50.000 et au 1 :100.000, les écarts-types sont inférieurs à 1% (respectivement 0,83% et 0.94%).

L'histogramme de la figure 4.76 présente la distribution des erreurs de surface relatives des limites communales de la BDCARTO sur l'échantillon de 382 communes étudié. Il montre que la distribution des erreurs relatives de surface présente une forme symétrique, centrée autour d'une moyenne nulle.

Afin d'appliquer un intervalle de confiance à notre estimation, nous avons étudié le pourcentage d'erreurs relatives de surface comprises dans des intervalles de σ, 2σ, et 3σ autour de l'erreur moyenne (cf. table 4.17).

	σ	2σ	3σ
BDCARTO	77,5	94,8	98,4
TOP100	79,8	94,2	98,4

Table 4.17 – Pourcentage d'erreurs de surface relatives comprises dans un intervalle de σ, 2σ, et 3σ autour de l'erreur moyenne

Les résultats présentés en table 4.17 montrent que 98,4% des erreurs relatives de surface sont inscrites dans un intervalle compris entre $M-3\sigma$ et $M+3\sigma$ pour les limites communales de la BDCARTO, comme pour celles extraites des données TOP100, l'écart-type ne variant que de 0,1%.

Ces résultats montrent que pour les limites communales étudiées, l'échelle de représentation influe peu sur l'imprécision des erreurs relatives de surface, qui sont comprises à plus de 98% dans un intervalle de confiance d'environ 2,8% (en moyenne) par rapport à la surface généralisée étudiée. Il est important de relativiser ces résultats, puisque les comparaisons n'ont été effectuées que sur deux bases de données aux échelles de représentation proches (1 :50.000 et 1 :100.000).

Cependant, étant donné que les limites communales étudiées présentent des surfaces hétérogènes, il parait nécessaire d'étudier l'imprécision des erreurs relatives de surface en fonction de la surface étudiée. En effet, la méthodologie de normalisation des erreurs de surface (en les rapportant à la surface généralisée) a pour conséquence de générer des erreurs relatives faibles pour les objets de grandes dimensions, et des erreurs relatives plus importantes sur les objets de petites dimensions. Pour l'illustrer, le graphique de la figure 4.77 rapporte les erreurs relatives de surface observées sur les limites communales de la BDCARTO par rapport à la surface étudiée.

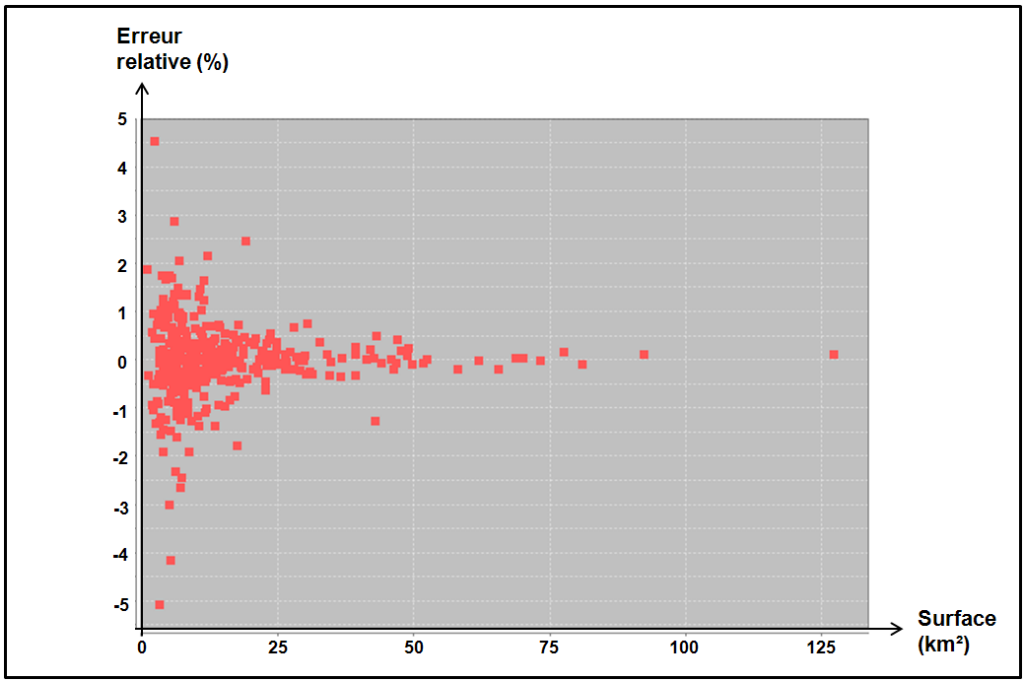


Figure 4.77 – Relation entre l'erreur de surface relative et la surface évaluée sur les limites communales de la BDCARTO

Le graphique montre que plus les surfaces des limites communales étudiées sont petites, plus les erreurs de surface sont dispersées, rendant la mesure plus imprécise. Au contraire, plus les surfaces étudiées sont importantes, plus l'imprécision de mesure est faible. Ces résultats tendent à démontrer que l'estimation de l'imprécision des erreurs de surface sur les limites communales généralisées doit prendre en compte la surface étudiée, comme nous avons pu le constater pour les mesures de longueur. A ce niveau d'avancement, l'élaboration de la fonction d'imprécision des erreurs relatives de surface en fonction de la surface étudiée n'a cependant pu être réalisée.

Modèle statistique Pour estimer l'imprécision des mesures de surface des limites communales, nous proposons ici un modèle statistique simplifié. Dans ce modèle, nous partons du principe que l'erreur de surface relative est indépendante de la surface étudiée. Ce modèle permet ainsi d'estimer une surface de limite communale de référence S_{Ref} à partir d'une surface de limite communale généralisée S_{Gen} conformément à l'équation 4.16.

Estimation de l'imprécision de surface des limites communales :

$$\begin{cases} S_{Ref}(LimitesCom, 50000) = S_{Gen} \quad (\pm 0,025 * S_{Gen}) \\ S_{Ref}(LimitesCom, 100000) = S_{Gen} \quad (\pm 0,028 * S_{Gen}) \end{cases} \tag{4.16}$$

Cependant, nous avons bien conscience que le modèle d'estimation ainsi proposé est réducteur, puisqu'il ne prend pas en compte la surface étudiée pour modéliser l'imprécision de mesure. Des tests complémentaires devront donc être réalisés afin d'élaborer la fonction d'imprécision en fonction de la surface étudiée, et de la valider.

Nous venons de voir que pour les surfaces communales, qui constituent une partition de l'espace, l'impact de la généralisation cartographique ne se traduit pas par une diminution systématique des surfaces étudiées. Ceci s'explique par l'opération de simplification de contour, principale opération utilisée pour généraliser les surfaces communales, qui a pour effet de préserver les surfaces tout en diminuant légèrement leur périmètre.

Cependant, nous avons observé que l'imprécision de l'erreur de surface relative dépend de la surface étudiée. Plus les limites communales sont de petites dimensions, plus leurs surfaces sont imprécises. A défaut de pouvoir proposer un modèle statistique réaliste, ces résultats nous encouragent à approfondir cette étude de l'impact de la généralisation cartographique sur les mesures de surface, en nous intéressant à des objets de petites dimensions ne constituant pas une partition de l'espace, pouvant être affectés par d'autres opérations de généralisation. Pour cela, nous proposons de réaliser des tests sur la classe d'objets des surfaces d'eau.

Application aux surfaces d'eau

Une seconde série de comparaisons de surfaces a été réalisée sur les surfaces d'eau extraites de la BDCARTO et des données TOP100 dans le département des Pyrénées-Atlantiques. Cette classe d'objets a été sélectionnée étant donné qu'elle est l'une des seules classes d'objets surfaciques à la fois présente dans la BDTOPO, mais également dans la BDCARTO et les données TOP100 (sous forme de classe d'occupation du sol).

Une sélection des objets pertinents a malgré tout dû être réalisée préalablement, afin de ne conserver que les objets homologues et comparables. Il en résulte un échantillon de 25 triplets d'objets homologues, ce qui n'est pas suffisant pour en dégager un modèle statistique.

Cet échantillon permet cependant d'étudier la sensibilité des mesures de surface d'objets de petites dimensions, et ne constituant pas une partition de l'espace. Les résultats des comparaisons sur les surfaces d'eau sont présentés dans la table 4.18.

	Surface	
	Moyenne(%)	Ecart-Type(%)
BDCARTO	-2,77	25,83
TOP100	-42,5	51,76

Table 4.18 – Résultats des comparaisons de surfaces sur les surfaces d'eau

Erreur moyenne Les résultats des comparaisons sur les surfaces d'eau montrent que ces objets géographiques (de petite dimension et ne représentant pas une partition de l'espace) sont beaucoup plus sensibles aux effets de la généralisation cartographique. Les résultats montrent en effet que l'échelle de représentation impacte fortement l'erreur moyenne de mesure. Ceci est particulièrement observable sur les surfaces d'eau extraites des données TOP100 (au 1 :100.000), où les mesures de surface sont sur-estimées en moyenne de 42,5% par rapport à la surface de référence de la BDTOPO.

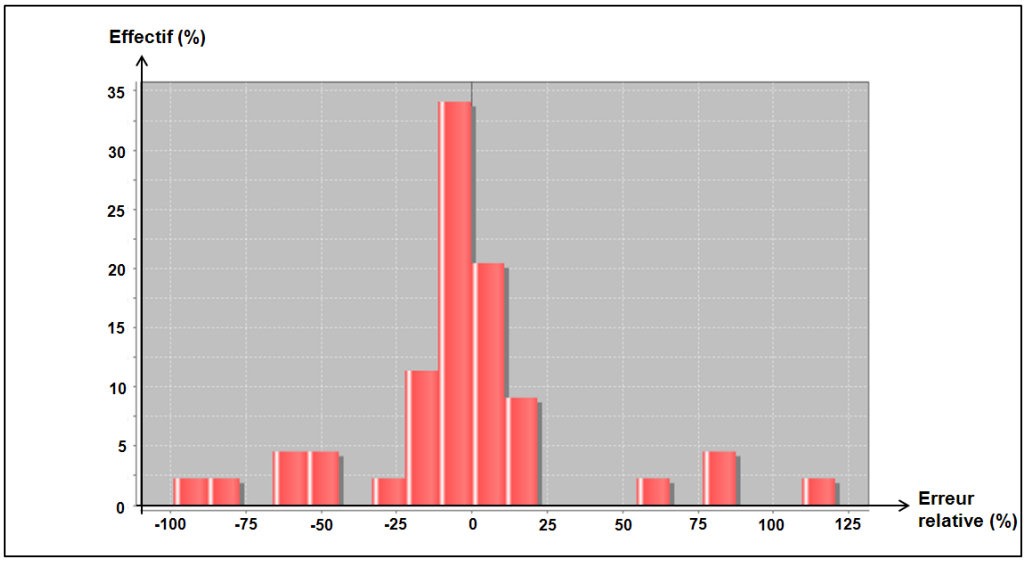


Figure 4.78 – Distribution des erreurs de surface relatives des surfaces d'eau de la BDCARTO

L'histogramme de la figure 4.78 présente la dispersion des erreurs relatives des surfaces d'eau extraites de la BDCARTO, par comparaison avec leurs homologues de la BDTOPO. On y voit que dans l'ensemble la distribution des erreurs relatives de surface est relativement centrée, bien que des sous-estimations et des sur-estimations de surfaces soient observées.

En comparaison, la dispersion des erreurs de surface relatives des surfaces d'eau extraites des données TOP100 présente clairement une tendance à la sur-estimation des surfaces (cf. figure 4.79). En effet, la grande majorité des erreurs de surface est négative, ce qui signifie que la surface des objets extraits des données TOP100 est supérieure à son homologue de la BDTOPO (sur les objets conservés dans les deux bases).

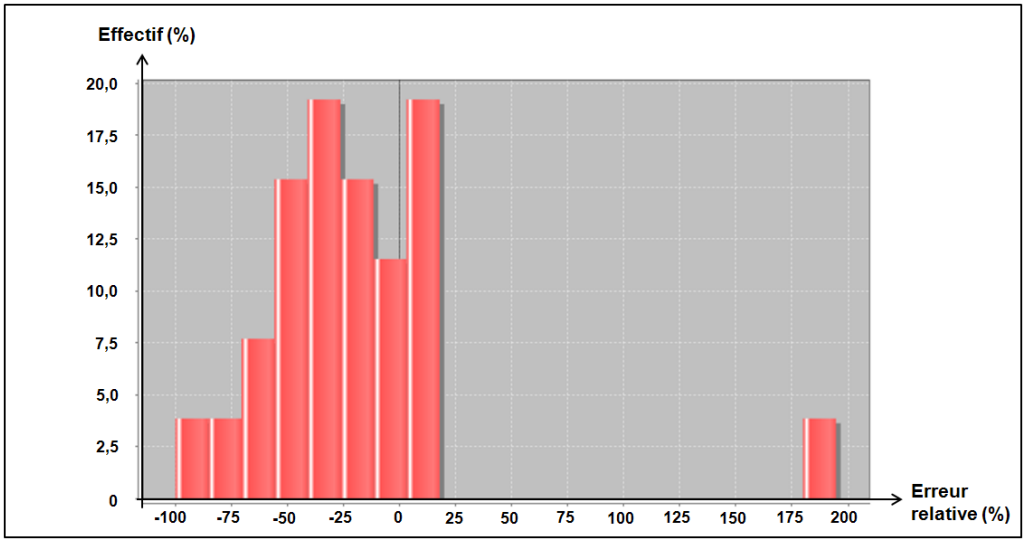


Figure 4.79 – Distribution des erreurs de surface relatives des surfaces d'eau des données TOP100

Cette sur-estimation s'explique par l'application de certaines opérations de généralisation cartographique, en particulier celle d'amplification, qui ont pour effet de dilater la géométrie de certains objets remarquables (comme peuvent l'être les surfaces d'eau), lorsqu'ils sont trop petits pour être visibles sur la carte à l'échelle de représentation (cf. figure 4.80).

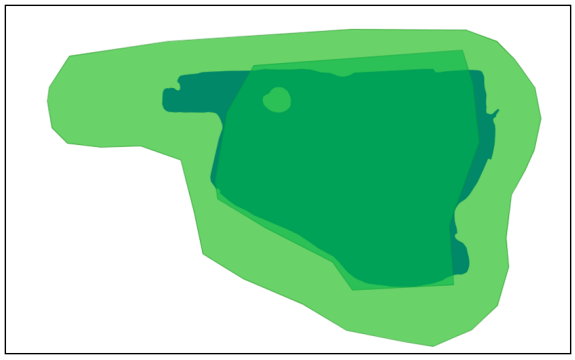


Figure 4.80 – Une surface d'eau représentée par la BDTOPO, la BDCARTO et les données TOP100 (du vert foncé au vert clair)

Les surfaces d'eau ne constituant pas une partition de l'espace, il est possible de leur appliquer des opérations d'amplification, au contraire des limites communales. Leurs surfaces étant de plus beaucoup plus petites, l'erreur relative de surface est par conséquent beaucoup plus importante.

Dispersion des erreurs Les écarts-types des erreurs relatives de surface sont également très importants (25,8% au 1 :50.000 et 51,8% au 1 :100.000), montrant une hétérogénéité des effets de la généralisation cartographique sur la géométrie des objets, et donc sur les mesures de surface. Certaines surfaces d'eau sont en effet fortement affectés (en particulier par des phénomènes d'amplification) et d'autres beaucoup moins.

Comme nous avons pu le remarquer sur les limites communales, nous observons également que plus la surface d'eau est petite, plus sa surface est imprécise. Pour l'illustrer, le graphique de la figure 4.81 présente les erreurs de surface relatives des surfaces d'eau de la BDCARTO, en fonction de la surface étudiée.

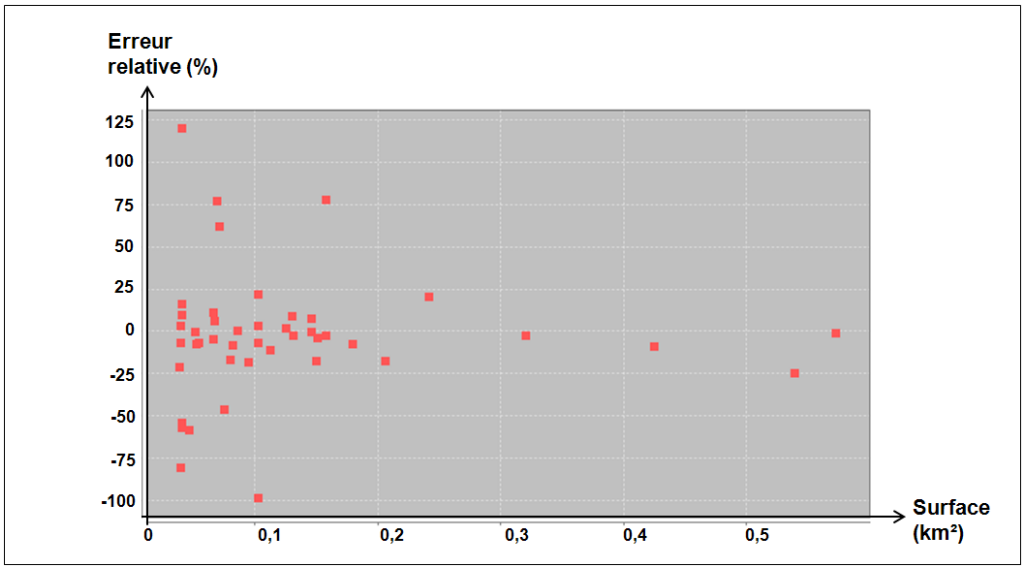


Figure 4.81 – Relation entre l'erreur de surface relative et la surface évaluée sur les surfaces d'eau de la BDCARTO

Vers un modèle statistique Etant donné la petite taille de l'échantillon testé (25 triplets d'objets homologues), il n'est pas possible de dégager un modèle statistique permettant d'estimer, pour les surfaces d'eau, l'impact de la généralisation cartographique sur les mesures géométriques de surface. Des comparaisons complémentaires sur de plus grands échantillons devront donc être réalisées afin de pouvoir en dégager des tendances significatives. Toutefois, les résultats des comparaisons sur les surfaces d'eau montrent bien que la généralisation cartographique affecte significativement les objets de petites dimensions, ne représentant pas une partition de l'espace.

Cependant, nous pouvons constater que l'impact de la généralisation cartographique sur les surfaces d'eau se manifeste de manière hétérogène sur les objets étudiés. En effet, nous pouvons constater que sur les objets de grandes dimensions, les surfaces calculées demeurent relativement précises, puisque ces objets sont principalement affectés par des opérations de simplification de contour, qui préservent relativement les formes. Au contraire, si les objets sont de petites dimensions, il peuvent également être affectés par des opérations d'amplification, qui ont pour conséquence d'impacter significativement leurs géométries. Le graphique de la figure 4.82 permet d'illustrer cette observation.

Sur ce graphique, les surfaces des objets de la BDCARTO sont mises en comparaison avec leurs homologues de la BDTOPO. On y voit que deux fonctions sont nécessaires pour modéliser la relation de surface entre les objets des deux bases : une fonction $f(x)$ pour les objets de petites dimensions et une fonction $g(x)$ pour les objets de plus grandes dimensions.

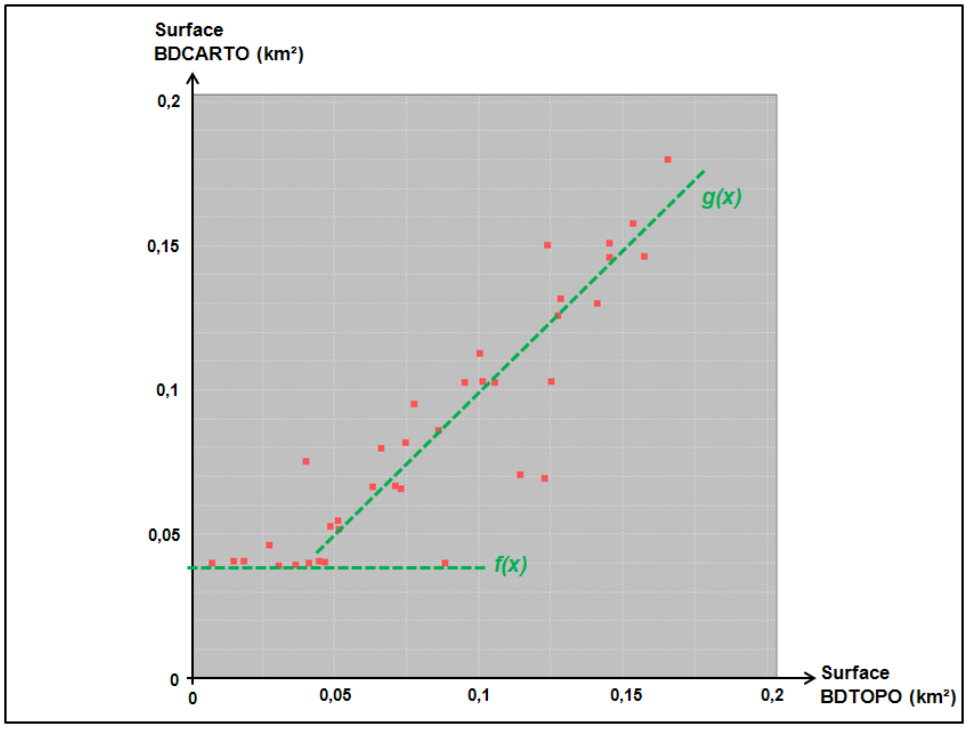


Figure 4.82 – Relation entre les surfaces d'eau de la BDTOPO et de la BDCARTO

La fonction $f(x)$ montre que les surfaces d'eau de la BDCARTO ne sont pas représentables en dessous d'une surface minimale d'environ 0,04 km^2, soit 40 hectares. Ainsi, pour les objets de la BDTOPO de dimension inférieure à cette valeur seuil, nous pouvons supposer que des opérations d'amplification sont mises en œuvre. La fonction $g(x)$ montre que pour les surfaces d'eau de plus grandes dimensions, les rapports de surfaces demeurent relativement identiques, bien qu'une certaine imprécision de mesure soit constatée.

La surface minimale observée des surfaces d'eau de la BDCARTO est de 40 hectares, ce qui est légèrement supérieur aux spécifications (IGN, 2011b), qui précisent que seules les surfaces d'eau d'une superficie supérieure à 25 hectares sont saisies. Cependant, étant donné que la BDCARTO est historiquement saisie à partir de cartes au 1 :50.000, nous pouvons supposer que des opérations d'amplification avaient déjà été appliquées sur les surfaces d'eau représentées dans la carte source.

Ces observations nous amènent par conséquent à envisager l'élaboration d'un modèle statistique sous forme de règles, prenant en compte la surface initiale des objets en entrée. Si la surface initiale des objets est supérieure à un seuil donnée (qui est fonction de l'échelle de représentation), il est possible que ceux-ci n'aient pas subi d'opérations d'amplification. Par conséquent l'imprécision de leur mesure demeure relativement faible. Au contraire, si la surface des objets est inférieure à ce même seuil, il est possible que des opérations d'amplification soient intervenues, générant une imprécision de mesure plus importante. Des expérimentations complémentaires sur de plus grands échantillons d'objets doivent être entreprises afin de fixer la surface seuil en fonction de l'échelle de représentation, et d'estimer les différentes fonctions d'estimation de l'imprécision des mesures de surface.

Une fois ces comparaisons réalisées, un modèle statistique organisé conformément au pseudo code ci-dessous pourra être envisagé, où T est la surface seuil, et où $SmallPolygonModel$ est le modèle statistique appliqué aux objets de petites dimensions, et $LargePolygonModel$ le modèle statistique appliqué aux objets de grandes dimensions.

```
1  for each polygon :dbPolygon
2     if polygon.area < T then
3          polygon.getImprecision() = SmallPolygonModel.estimate(polygon)
4     else if polygon.area > T then
5          polygon.getImprecision() = LargePolygonModel.estimate(polygon)
6     end if
7  end for
```

Conclusions Les comparaisons réalisées entre objets surfaciques ont permis de montrer que pour estimer l'imprécision des mesures de surface, il est tout d'abord nécessaire de différencier les objets surfaciques selon leurs propriétés, notamment s'ils représentent une partition de l'espace ou non, du fait des opérations de généralisation mobilisées. Pour les objets représentant une partition de l'espace, nous avons vu que les erreurs relatives de surface sont faibles, ce qui est logique étant donné que les opérations de généralisation utilisées préservent les surfaces. Pour les objets ne représentant pas une partition de l'espace, nous avons vu que les mesures de surface sont beaucoup plus imprécises. Ceci s'explique par les opérations d'amplification dont ces objets peuvent être affectés, notamment lorsqu'ils présentent de petites dimensions. Nous avons ainsi pu voir qu'il est nécessaire de prendre en compte la surface des objets évalués afin de quantifier l'imprécision de leur mesure. Ainsi, ces premières conclusions offrent des perspectives quant à l'élaboration d'un modèle d'estimation de l'imprécision des mesures de surface. Pour cela, des tests supplémentaires doivent être réalisés sur de plus grands échantillons et sur d'autres classes d'objets.

Synthèse Section Nous avons présenté dans cette section les résultats de comparaisons permettant d'estimer l'impact de la généralisation cartographique sur les mesures géométriques de surface. Les comparaisons ont été réalisées sur les classes d'objets des limites communales et des surfaces d'eau. Nous avons tout d'abord montré que les impacts de la généralisation cartographique varient en fonction des propriétés des objets surfaciques, en particulier s'ils représentent une partition de l'espace ou non, étant donné que les opérations de généralisation ne sont pas les mêmes. Pour les limites communales, l'erreur de surface moyenne est pratiquement nulle, du fait de l'application d'opérations de simplification de contour qui préservent les surfaces, tandis qu'une tendance à la surestimation est observée sur les surfaces d'eau généralisées, du fait des opérations d'amplification mobilisées. Egalement, ces comparaisons ont permis de montrer que l'imprécision des mesures de surface est dépendante de la surface étudiée, les objets de petites dimensions étant plus sensibles que ceux de grandes dimensions. Si l'élaboration d'un modèle statistique prenant en compte la surface étudiée et les propriétés des objets évalués n'a pu être envisagée, des pistes relatives à son élaboration ont malgré tout été proposées.

Conclusion de la section

Cette section a permis d'étudier l'impact de la généralisation cartographique, en tant que processus de production de données, sur les mesures géométriques de longueur et de surface. L'objectif de ce travail visait à élaborer des modèles statistiques permettant d'estimer l'imprécision d'une mesure géométrique calculée à partir de données généralisées. Etant donné la diversité des impacts géométriques pouvant potentiellement faire suite à un processus de généralisation cartographique, nous avons décidé d'élaborer ces modèles statistiques par le biais d'une approche par tests, faisant l'objet de comparaisons préalables avec des données de référence. Des méthodes de comparaisons automatiques ont donc été élaborées afin d'estimer l'imprécision des mesures géométriques de longueur et de surface sur des objets généralisés.

Pour les objets linéaires, nous sommes partis de l'hypothèse que l'impact de la généralisation cartographique sur les mesures de longueur pouvait varier en fonction de la classe d'objets étudiée, du contexte géographique, et de l'échelle de représentation. Ces hypothèses ont été confirmées à travers les comparaisons réalisées sur les objets d'un réseau routier. Les tests de comparaisons ont permis d'élaborer un modèle statistique permettant d'estimer une longueur de référence et son imprécision, à partir d'une longueur généralisée. Nous avons également observé que l'erreur moyenne de longueur impliquée par la généralisation cartographique était indépendante de la longueur étudiée.

Pour les objets surfaciques, l'hypothèse de l'influence du contexte géographique sur les mesures de longueur et de surface n'a pas été prise en compte. Nous nous sommes uniquement concentré sur les impacts de l'échelle de représentation et de la classe d'objets considérée. Des comparaisons ont été menées sur des objets représentant une partition de l'espace (les limites communales) et n'en représentant pas (les surfaces d'eau). Les résultats des comparaisons ont notamment montré que l'imprécision des mesures de surface dépend à la fois de la propriété des objets étudiés et de la surface étudiée, qui conditionnent les opérations de généralisation mises en œuvre. Nous avons montré que les objets ne représentant pas une partition de l'espace et présentant de petites dimensions peuvent être affectés par des opérations d'amplification, qui ont pour conséquence de rendre les mesures géométriques de surface très imprécises. Ces conclusions ne nous ont pas permis d'élaborer un modèle statistique d'estimation de l'imprécision des mesures de surface formalisé, mais ont cependant ouvert quelques pistes permettant de l'élaborer.

Les différents résultats ont permis de montrer que l'impact de la généralisation cartographique sur les mesures géométriques se manifeste de manière très hétérogène en fonction des classes d'objets concernées, mais également de l'échelle de représentation ou du contexte géographique. Cependant, nous avons également observé que les mesures géométriques sont également sensibles à d'autres paramètres, comme la nature des objets et leur taille. Cela suppose donc l'élaboration de méthodologies spécifiques pour chacune des deux mesures, afin d'élaborer des modèles statistiques réalistes.

Dans le cadre de l'élaboration du modèle d'estimation de l'imprécision des mesures géométriques, seul le modèle statistique d'estimation de l'imprécision des mesures de longueur d'un réseau routier sera donc intégré. Celui-ci pourra malgré tout être enrichi d'autres modèles élaborés sur d'autres classes d'objets.

Enfin, ces travaux soulignent l'intérêt d'estimer l'impact des différentes opérations de généralisation cartographique sur les mesures géométriques, afin de mieux qualifier les bases de données cartographiques produites par dérivation. De plus, nous ne nous sommes penché dans cette section que sur l'impact de la généralisation cartographique sur des objets homologues conservés aux différents niveaux d'échelle étudiés. Il serait également intéressant d'étudier l'impact de la généralisation de modèle, impliquant la présence ou non d'objets à certaines échelles de représentation cibles. Son impact doit en effet pouvoir être pris en compte dans des applications mobilisant des mesures géométriques, comme les calculs d'itinéraires par exemple.

Conclusion du chapitre

Ce chapitre a permis de présenter les méthodes d'estimation de l'impact des processus de production sur les mesures géométriques de longueur et de surface. Deux impacts des processus de production ont été traités : l'erreur de pointé et la généralisation cartographique. Afin de modéliser ces impacts, deux approches ont été mobilisées. Une approche par simulations pour modéliser l'impact de l'erreur de pointé, et une approche par tests pour la généralisation cartographique.

Comme pour les règles de représentation, l'estimation de ces deux impacts implique la mise en place d'un processus d'acquisition des connaissances. Ces connaissances se manifestent sous forme de données externes (délimitation du contexte géographique) ou par le biais d'informations extraites des données évaluées (échelles de représentation ou de saisie). Différentes méthodes ont été proposées afin de faciliter ce processus d'acquisition des connaissances nécessaire à l'estimation des deux impacts.

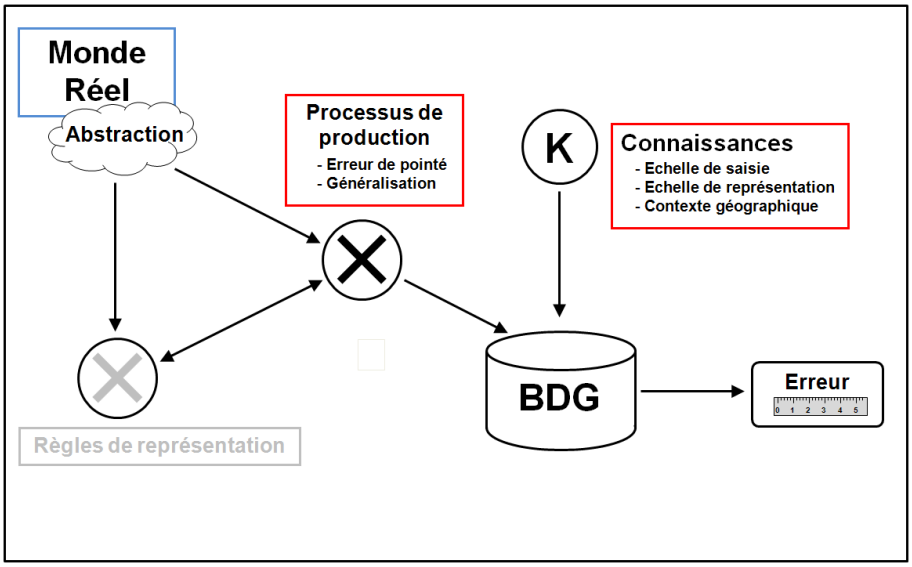


Figure 4.83 – Modélisation de l'impact des processus de production au sein du processus global de génération d'erreur de mesure

Suite aux expérimentations, nous avons pu observer que les ordres de grandeur impliqués par ces deux impacts sont très différents. La généralisation cartographique a notamment pour conséquence de se manifester par des erreurs de mesure très importantes.

Bien que ces impacts aient été traités de manière indépendante, nous avons bien conscience qu'ils s'accumulent dans la pratique (lorsque les données traitées sont généralisées). De plus, les évaluations effectuées au cours de ce chapitre sont également affectées par les impacts des règles de représentation étudiés dans le chapitre précédent. Cependant, nous avons fait le choix de traiter chacun de ces impacts de manière indépendante jusqu'à présent, afin d'étudier leurs effets respectifs. Nous traiterons de la combinaison de ces impacts dans le chapitre 5, ce qui permettra par la suite de mettre en œuvre le modèle général d'estimation de l'imprécision des mesures géométriques.

Chapitre 5

Conception du modèle d'estimation de l'imprécision des mesures géométriques

Introduction

Nous avons présenté dans les chapitres précédents différentes méthodes permettant de modéliser les impacts de règles de représentation et de processus de production sur les mesures géométriques de longueur et de surface. Nous allons décrire dans ce chapitre comment sont intégrées ces méthodes au sein d'un modèle général, permettant à un utilisateur d'estimer l'imprécision des mesures géométriques calculées à partir des objets d'une base de données.

Pour procéder à cette estimation, nous proposons un modèle général qui intègre l'ensemble des méthodes d'estimation des impacts des règles de représentation et des processus de production. A ce stade, nous avons vu qu'il était possible d'estimer chacun de ces impacts de manière indépendante. L'estimation des impacts respectifs constitue le premier objectif en termes de communication de l'information sur l'imprécision des mesures géométriques. Cependant, l'objectif final de ce travail est de fournir à l'utilisateur une estimation globale de l'imprécision de la mesure géométrique calculée. Pour cela, il est nécessaire de combiner les impacts estimés. Nous proposerons donc dans la première section de ce chapitre plusieurs méthodes de combinaison des impacts, qui seront testées et évaluées afin d'identifier la plus performante.
Chacun des modèles d'estimation des impacts des règles de représentation et des processus de production constitue une composante du modèle général d'estimation. Nous détaillerons dans la seconde section de ce chapitre les choix de modélisation retenus afin d'intégrer chacune de ces composantes au sein du modèle général.
Enfin, la troisième section de ce chapitre sera consacrée au développement du prototype EstIM (Estimation de l'Imprécision des Mesures). Nous y décrirons son principe de fonctionnement et ses interfaces homme-machine, et illustrerons sa mise en œuvre en nous appuyant sur un cas d'utilisation concret.

5.1 Combinaison des impacts sur les mesures géométriques

Cette section a pour objectif de présenter la démarche employée afin de produire une estimation globale de l'imprécision des mesures géométriques à partir des impacts des règles de représentation et des processus de production sélectionnés. Nous nous sommes attaché jusqu'à présent à modéliser ces différents impacts de manière indépendante. Des méthodes permettant de quantifier ces impacts ont ainsi été développées, ce qui constituait notre premier objectif en termes d'évaluation (cf. chapitres 3 et 4).

Cependant, le second objectif de ce travail consiste à produire une estimation globale de l'imprécision des mesures géométriques. Cette estimation doit pouvoir être formulée de la manière suivante : "*La mesure calculée est de X mètres, cependant nous estimons que cette mesure est probablement de (X+A) mètres, comprise entre (X+B) mètres et (X+C) mètres*". La production de cette estimation globale de l'imprécision soulève donc des questions relatives à la combinaison des impacts sélectionnés.

Nous savons que les impacts des règles de représentation et des processus de production s'accumulent et génèrent une erreur de mesure géométrique. Cependant nous ne savons pas comment s'accumulent ces impacts, ni même s'il est possible de modéliser les inter-actions entre ces différentes composantes de l'erreur finale de mesure.
Ainsi, pour produire une estimation globale de l'imprécision des mesures, nous proposons dans cette section plusieurs méthodes de combinaison des impacts, qui, malgré leurs limites ou leurs imperfections, peuvent nous aider à atteindre cet objectif. Nous présenterons dans un premier temps les méthodes de combinaison des impacts envisagées dans le cadre de ce travail de thèse, puis les mettrons en application afin de comparer leurs performances respectives. Nous chercherons notamment à identifier les méthodes de combinaison des impacts qui permettent de minimiser l'erreur de mesure calculée, par comparaison avec des données de référence. Enfin, des pistes complémentaires relatives à d'autres méthodes de combinaison des impacts seront abordées en guise de perspectives.

5.1.1 Méthodes de combinaison des impacts

Avant de présenter les méthodes de combinaison des impacts retenues dans le cadre de ce travail de thèse, nous exposerons tout d'abord la difficulté que constitue une telle entreprise.

5.1.1.1 La combinaison des impacts : une démarche complexe

Produire une estimation globale de l'imprécision des mesures géométriques par combinaison des impacts constitue à plus d'un titre une démarche complexe. Différents arguments sont ici exposés afin de souligner la difficulté et les limites d'une telle entreprise.

Une réalité complexe à modéliser Nous avons déjà souligné que les impacts des règles de représentation et des processus de production s'accumulent dans la pratique, mais nous ne savons pas de quelle manière. Il est donc légitime de se demander s'il est mathématiquement possible de combiner ces impacts. En effet, les modèles développés dans le cadre de ce travail de thèse ont pour objectif de quantifier un impact, de manière indépendante.

Cependant, nous savons que l'estimation d'un impact donné par un modèle donné, "absorbe" dans le même temps les erreurs impliquées par un autre impact. Par exemple, l'estimation de l'impact de la non-prise en compte du terrain est déjà affectée par l'impact de la projection cartographique (puisque les objets évalués sont déjà projetés). De la même manière, l'impact de la généralisation cartographique est lui-même affecté de l'impact de l'erreur de pointé, bien que les ordres de grandeur soient très différents.

Ces observations montrent qu'il est très difficile de quantifier par des modèles isolés, des impacts qui se combinent de manière complexe les uns avec les autres.

Limites des modèles proposés Si la réalité des interactions entre impacts est délicate à appréhender, il est également nécessaire d'avoir conscience des limites des modèles élaborés dans le cadre de cette thèse. En effet, les modèles d'impact proposés dans les chapitres 3 et 4 permettent de produire une estimation d'un impact, avec toutes les imperfections des choix de modélisation retenus. Par exemple, nous nous appuyons sur des MNT (ou bien des MNE) pour quantifier l'impact de la non-prise en compte du terrain, en ayant bien conscience que l'information altimétrique qu'ils contiennent présente elle aussi une imprécision. Egalement, l'impact de la généralisation cartographique est élaboré par comparaison avec des données de référence qui portent elles-mêmes de l'imprécision.

Chacune des estimations produites par les modèles d'impact porte donc une imprécision, qu'il est parfois difficile d'estimer.

Hétérogénéité interne des objets Les modèles développés au cours de ce travail de thèse ne permettent de traiter les objets géographiques que dans leur globalité, mais ne permettent pas d'appréhender leur hétérogénéité interne ou contextuelle. Par exemple, les méthodes proposées pour estimer l'impact de l'approximation polygonale des courbes sont élaborées pour traiter une collection d'objets entière ou une sélection d'objets sinueux. Mais dans la réalité, les objets routiers peuvent présenter certaines sections sinueuses et d'autres non. Egalement, l'impact de la généralisation cartographique peut se manifester localement par des sous-estimations ou des sur-estimations de mesures, en fonction du contexte des objets, comme nous avons pu le voir avec l'étude des surfaces d'eau.

Les modèles proposés permettent donc de quantifier des tendances d'erreurs sur des objets, mais ne permettent pas d'appréhender l'hétérogénéité interne des objets, qui impacte pourtant les mesures géométriques.

Quelle méthode de combinaison ? A travers ces différentes remarques, nous voyons bien qu'il est complexe d'estimer globalement l'imprécision des mesures géométriques en cherchant à combiner les impacts affectant la géométrie des objets géographiques. Cependant, ceci constitue un besoin pour un utilisateur de données géographiques. Ainsi, différentes méthodes de combinaison des impacts peuvent être envisagées, malgré leurs imperfections. Nous en proposerons deux dans le cadre de ce travail de thèse, par agrégation ou sélection des estimations produites indépendamment par chacun des modèles. D'autres méthodes de combinaison (par composition de modèles ou par apprentissage supervisé) seront également évoquées en guise de perspectives (cf. section 5.1.3).

5.1.1.2 Proposition de méthodes de combinaison

Bien qu'il soit délicat de combiner les impacts potentiels affectant les mesures géométriques, pour les raisons que nous venons d'évoquer, nous chercherons malgré tout à estimer la performance de deux méthodes de combinaison simples, qui seront mises en application dans la suite de ce travail.

Principe général L'approche proposée réside dans une combinaison des estimations calculées de manière indépendante par chacun des modèles d'impact des règles de représentation et des processus de production. Cette approche suppose que chaque modèle soit activé de manière indépendante à partir du jeu de données à évaluer (cf. figure 5.1).

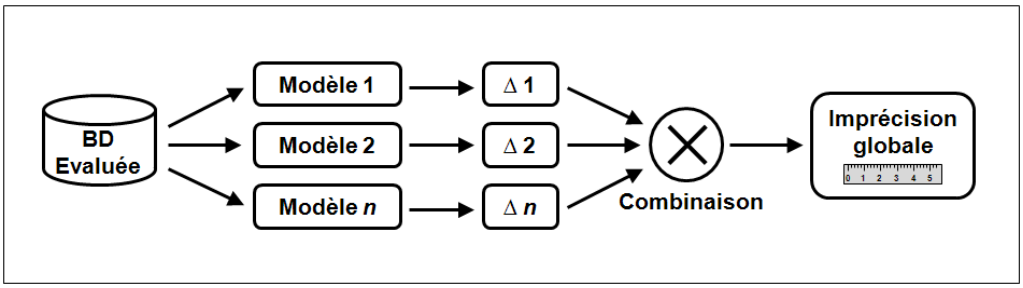


Figure 5.1 – Principe général de l'approche par combinaison des estimations

Différentes méthodes peuvent ensuite être utilisées afin de combiner les estimations calculées par chacun des modèles, et ainsi produire une estimation globale de l'imprécision des mesures. Deux méthodes de combinaison sont ici proposées : l'agrégation et la sélection des impacts.

Agrégation des impacts

Le principe d'agrégation proposé ici repose sur une addition des estimations des impacts calculées indépendamment par chacun des modèles sélectionnés pour l'évaluation. Cependant, deux méthodes d'agrégation seront testées dans le cadre de ce travail : l'agrégation simple et l'agrégation quadratique des erreurs.

Agrégation simple des erreurs La première méthode repose sur une addition simple des impacts estimés de manière indépendante. L'équation 5.1 permet de formaliser ce principe d'agrégation simple des impacts Δ_1, Δ_2 ... Δ_i en une erreur globale Δ_T.

$$\Delta_T = \Delta_1 + \Delta_2 + ... + \Delta_i \tag{5.1}$$

Cette première méthode constitue une solution aisée à mettre en œuvre afin de combiner les impacts modélisés de manière indépendante. Cependant, cette méthode pose le risque de majorer l'erreur estimée, puisque nous supposons que chaque impact estimé intègre une part de l'erreur impliquée par les autres impacts. C'est pourquoi une seconde méthode d'agrégation est proposée.

Agrégation des erreurs quadratiques La seconde méthode d'agrégation proposée repose sur une addition des erreurs quadratiques. L'intérêt de cette méthode est qu'elle permet de majorer la contribution des impacts les plus significatifs, et dans le même temps de minorer la part des impacts les moins significatifs. L'équation 5.2 permet de formaliser ce principe d'agrégation des erreurs quadratiques des impacts Δ_i en une erreur globale Δ_T.

$$\Delta_T = \sqrt{\sum_{i=1}^{n} (\Delta_i)^2} \tag{5.2}$$

Les deux méthodes d'agrégation des impacts proposées constituent donc des solutions simples, bien qu'incorrectes, pour produire une estimation globale de l'imprécision des mesures géométriques. En effet, les composantes de l'erreur de mesure s'accumulent de manière plus complexe que par une simple addition, ce qui rend ces méthodes relativement réductrices.

Nous supposons également que la méthode par agrégation simple majore l'erreur globale estimée. Cette majoration, bien que pessimiste, peut dans le même temps s'avérer sécurisante dans le but de produire une estimation globale de l'imprécision des mesures. C'est pourquoi nous considérons malgré tout intéressant d'étudier la pertinence de cette méthode de combinaison.

Sélection des impacts

Si l'agrégation simple des impacts a pour principal risque de majorer l'erreur de mesure estimée, nous proposons dans ce cas de ne sélectionner que l'impact le plus significatif, afin d'estimer une erreur globale de mesure. En effet, certaines composantes de l'erreur finale de mesure, comme la généralisation cartographique par exemple, peuvent avoir comme principal effet d'absorber les erreurs impliquées par d'autres impacts. Selon cette hypothèse, retenir l'impact le plus significatif peut permettre d'approximer l'erreur réelle de mesure, sans trop majorer l'estimation.

Ainsi, à partir des impacts Δ_1, Δ_2 ... Δ_i modélisés de manière indépendante, le principe de sélection permet d'estimer l'erreur de mesure globale Δ_T conformément à l'équation 5.3.

$$\Delta_T = max(\Delta_1, \Delta_2, ..., \Delta_i) \tag{5.3}$$

Bien que ces deux méthodes de combinaison des impacts ne permettent pas véritablement de modéliser la complexité des relations entre les différents impacts, elles offrent l'avantage d'être aisées à mettre en œuvre dans le cadre d'une estimation globale de l'imprécision des mesures géométriques. Nous proposons donc de les mettre en application dans la section suivante, afin d'étudier leurs performances respectives.

Synthèse Section La production d'une estimation globale de l'imprécision des mesures géométriques, par combinaison des différents impacts impliqués, est une démarche complexe. Les modèles d'erreurs proposés ne permettent d'appréhender que de manière partielle la complexité des interactions entre les différentes composantes de l'erreur finale. Malgré ces imperfections, deux méthodes de combinaison des impacts sont proposées pour estimer l'imprécision globale des mesures géométriques : l'agrégation et la sélection des impacts modélisés de manière indépendante.

5.1.2 Application des méthodes de combinaison

Cette section a pour objectif de mettre en application les méthodes de combinaison des impacts présentées précédemment, pour produire une estimation globale de l'imprécision des mesures géométriques. Nous chercherons principalement à identifier la méthode de combinaison offrant la meilleure estimation possible de l'erreur de mesure géométrique, mais nous nous intéresserons également à la complémentarité des différentes méthodes. Pour cela, nous proposons d'appliquer les modèles élaborés dans les chapitres 3 et 4 sur un échantillon de données, et de comparer les résultats produits par les méthodes de combinaison des impacts, avec une mesure de référence.
Cette analyse ne concerne ici que l'étude des mesures de longueur. En effet, nous avons souligné précédemment que de nombreux travaux complémentaires nécessitent d'être réalisés au sujet des mesures de surface (en particulier sur l'impact de la généralisation cartographique).
Nous présenterons tout d'abord l'expérimentation réalisée, puis comparerons les résultats produits par les méthodes de combinaison des impacts, et terminerons par une analyse de leurs performances pour une estimation globale.

5.1.2.1 Présentation de l'expérimentation

Afin d'étudier les méthodes de combinaison des impacts, nous proposons de réaliser une expérimentation basée sur un échantillon d'itinéraires routiers tirés aléatoirement sur plusieurs bases de données. Les mesures de longueur observées sont comparées avec leurs homologues extraites d'une base de données de référence.

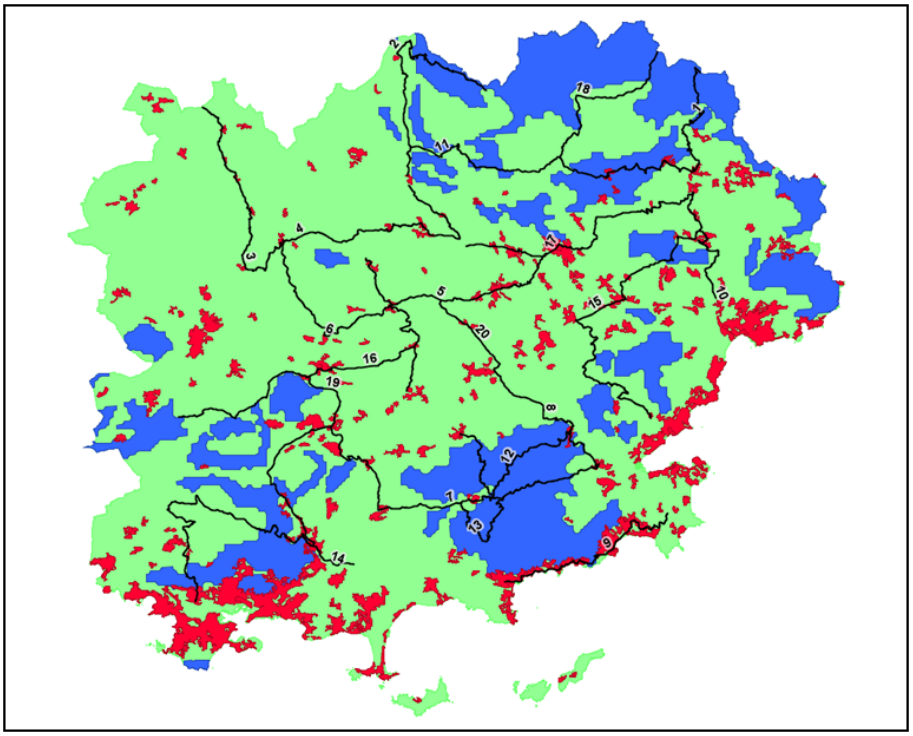


Figure 5.2 – Les 20 itinéraires au sein de leur contexte géographique (montagne en bleu, urbain en rouge, et rural en vert)

Données utilisées

Les données utilisées dans le cadre de cette expérimentation sont différentes de celles utilisées pour la validation du modèle d'impact de la généralisation cartographique, excepté pour la délimitation du contexte géographique (cf. section 4.2.3.3).

Bases de données Un échantillon de 20 itinéraires routiers (cf. figure 5.2), tirés aléatoirement à partir de la BDCARTO et des données TOP100 dans le département du Var, est exploité dans le cadre de cette expérimentation. En guise de référence, les itinéraires homologues extraits de la BDTOPO sont utilisés. Chacun de ces itinéraires représente la même entité du monde réel dans les trois bases de données. Pour procéder à l'appariement des itinéraires homologues dans les trois bases, l'algorithme de Mustière et Devogèle (2008) est utilisé, suivi d'une validation visuelle.

Contexte géographique La même délimitation du contexte géographique que celle utilisée en section 4.2.3.3 est exploitée afin d'étudier l'impact de la généralisation cartographique. Un ordre de priorité est également appliqué pour les portions d'itinéraires localisées à la fois en zone de montagne et en zone urbaine. Dans ce cas de figure, nous considérons qu'elles sont localisées en zone urbaine. La structure des itinéraires en fonction du contexte géographique est présentée en table 5.1.

	Longueurs (en km)							
	BDCARTO				**TOP100**			
	Total	**Mont.**	**Urbain**	**Rural**	**Total**	**Mont.**	**Urbain**	**Rural**
1	**49,768**	27,955	7,310	14,503	**47,935**	26,722	5,741	15,472
2	**43,805**	16,774	0,285	26,745	**42,055**	15,350	0,296	26,409
3	**54,680**	0	4,439	50,241	**52,063**	0	4,054	48,009
4	**40,619**	0	3,759	36,860	**39,396**	0	3,733	35,663
5	**46,461**	0	9,851	36,610	**45,425**	0	8,654	36,771
6	**30,845**	0	3,134	27,711	**29,616**	0	2,999	26,617
7	**52,749**	14,387	6,835	31,527	**51,036**	13,232	5,978	31,826
8	**38,880**	18,179	4,818	15,883	**36,154**	17,083	3,942	15,129
9	**36,187**	3,068	22,219	10,901	**34,389**	3,097	19,731	11,560
10	**22,847**	0,254	5,149	17,444	**21,715**	0,306	4,613	16,795
11	**28,959**	5,303	0	23,657	**27,788**	5,115	0	22,673
12	**21,928**	21,235	0,226	0,468	**20,269**	19,394	0,318	0,557
13	**22,018**	20,394	0	1,625	**19,786**	18,352	0	1,434
14	**62,274**	12,264	9,892	40,119	**58,610**	11,304	9,129	38,178
15	**56,819**	4,430	6,642	45,747	**53,700**	4,707	6,575	42,418
16	**40,683**	1,334	6,594	32,755	**39,887**	1,503	6,155	32,229
17	**32,364**	0	3,376	28,988	**30,634**	0	3,349	27,285
18	**26,496**	6,820	0	19,675	**25,792**	6,949	0	18,843
19	**35,033**	4,333	8,083	22,617	**34,582**	4,256	7,340	22,986
20	**22,570**	0	0,970	21,600	**22,418**	0	1,212	21,206

Table 5.1 – Structure des itinéraires extraits de la BDCARTO et des données TOP100 selon le contexte géographique

Erreurs de longueur initiales

Méthodologie de calcul La quantification des erreurs de longueur suppose une normalisation géométrique des données initiales afin de les rendre comparables. Cette normalisation géométrique fait référence aux impacts de l'approximation polygonale des courbes et de la projection cartographique.

Tout d'abord, nous considérons que les 20 itinéraires représentent tous des entités courbes du monde réel. De ce fait, si des géométries courbes sont reconstruites sur les itinéraires extraits de la BDCARTO et des données TOP100, il est également nécessaire de l'appliquer sur les itinéraires homologues de la BDTOPO. En effet, bien que le degré d'approximation polygonale soit beaucoup plus faible sur ces derniers, il est nécessaire, d'un point de vue méthodologique, de reconstruire des géométries courbes sur ces objets.
Egalement, l'impact de la projection cartographique ne sera pas intégré dans les comparaisons de longueurs. En effet, les itinéraires extraits des trois bases de données étant tous projetés dans le même système de projection (à savoir la projection Lambert93), nous considérons que cet impact agit de manière (pratiquement) identique sur les trois bases de données.

Ainsi, la quantification des erreurs initiales est réalisée en comparant les longueurs en 2D des itinéraires extraits de la BDCARTO et des données TOP100, avec les longueurs courbes en 2D5 de leurs homologues extraits de la BDTOPO (notées "2D5C"). La table 5.2 permet de résumer ces erreurs de longueur sur chacun des 20 itinéraires.

	BDTOPO		**BDCARTO**			**TOP100**		
	Long. (km)		**Long. (km)**	**Erreur**		**Long. (km)**	**Erreur**	
	2D5	**2D5C**	**2D**	**km**	*%*	**2D**	**km**	*%*
1	50,030	**50,055**	**49,768**	-0,287	*-0,58*	**47,935**	-2,120	*-4,42*
2	43,886	**43,903**	**43,805**	-0,099	*-0,23*	**42,055**	-1,849	*-4,40*
3	55,001	**55,019**	**54,680**	-0,339	*-0,62*	**52,063**	-2,956	*-5,68*
4	40,545	**40,556**	**40,619**	0,063	*0,16*	**39,396**	-1,160	*-2,94*
5	46,737	**46,770**	**46,461**	-0,309	*-0,67*	**45,425**	-1,345	*-2,96*
6	30,896	**30,919**	**30,845**	-0,073	*-0,24*	**29,616**	-1,303	*-4,40*
7	53,002	**53,041**	**52,749**	-0,292	*-0,55*	**51,036**	-2,005	*-3,93*
8	39,016	**39,049**	**38,880**	-0,169	*-0,43*	**36,154**	-2,894	*-8,01*
9	36,734	**36,765**	**36,187**	-0,578	*-1,60*	**34,389**	-2,376	*-6,91*
10	22,891	**22,911**	**22,847**	-0,065	*-0,28*	**21,715**	-1,197	*-5,51*
11	29,041	**29,051**	**28,959**	-0,092	*-0,32*	**27,788**	-1,263	*-4,54*
12	22,375	**22,406**	**21,928**	-0,477	*-2,18*	**20,269**	-2,137	*-10,54*
13	22,567	**22,597**	**22,018**	-0,579	*-2,63*	**19,786**	-2,811	*-14,21*
14	63,221	**63,271**	**62,274**	-0,996	*-1,60*	**58,610**	-4,661	*-7,95*
15	58,406	**58,464**	**56,819**	-1,645	*-2,89*	**53,700**	-4,764	*-8,87*
16	40,986	**41,008**	**40,683**	-0,324	*-0,80*	**39,887**	-1,120	*-2,81*
17	33,091	**33,116**	**32,364**	-0,751	*-2,32*	**30,634**	-2,482	*-8,10*
18	26,434	**26,442**	**26,496**	0,054	*0,20*	**25,792**	-0,650	*-2,52*
19	35,051	**35,059**	**35,033**	-0,027	*-0,08*	**34,582**	-0,477	*-1,38*
20	22,545	**22,550**	**22,570**	0,020	*0,09*	**22,418**	-0,132	*-0,59*

Table 5.2 – Erreurs de longueur des vingt itinéraires comparés

Analyse des erreurs La table 5.2 montre qu'au total, les longueurs des itinéraires extraits de la BDCARTO et des données TOP100 sont sous-estimées respectivement de 0,9% et de 5,4%. Cependant, nous pouvons remarquer que ces erreurs sont très hétérogènes sur l'ensemble des itinéraires traités.

En effet, trois des itinéraires de la BDCARTO sont plus longs que leurs homologues de la BDTOPO. Cette situation s'explique principalement par l'impact de la généralisation cartographique, qui peut avoir comme conséquence d'élargir certains virages, et par conséquent d'augmenter la longueur de la route.

Au contraire de la BDCARTO, les itinéraires extraits des données TOP100 présentent tous des longueurs sous-estimées en comparaison de leurs homologues de la BDTOPO. Nous constatons que l'erreur de longueur peut atteindre 14,2% pour l'itinéraire 13 (cf. figure 5.3), qui est principalement localisé en zone de montagne, et a fortement subi l'impact de la généralisation cartographique.

Figure 5.3 – Exemple de l'itinéraire 13, représenté par la BDTOPO (en rouge) et les données TOP100 (en vert)

D'une manière générale, les erreurs de longueur varient énormément d'un itinéraire à l'autre au sein d'une même base de données. Ces variations peuvent s'expliquer par différents facteurs, en particulier le contexte géographique au sein duquel les tronçons de route sont inscrits, mais également les caractéristiques internes des routes. En effet, nous avons proposé en section 4.2.3.2 un modèle d'estimation de l'impact de la généralisation sur les erreurs de longueur des tronçons de route qui ne différencie pas les objets routiers selon leur classification. Or il est possible que ces erreurs varient selon ces caractéristiques.

Suite à la quantification des erreurs de référence, nous proposons dans la section suivante d'estimer les impacts affectant les mesures de longueur sur les vingt itinéraires, puis de procéder à leur combinaison pour une estimation globale.

5.1.2.2 Estimation et combinaison des impacts

L'analyse de la performance des méthodes de combinaison des impacts implique tout d'abord de les quantifier de manière indépendante. Nous présenterons dans un premier temps les résultats des estimations calculées par chaque modèle d'impact, puis opèrerons à une combinaison en utilisant les différentes méthodes proposées.

Estimation des impacts

Nous présenterons ici la méthodologie de calcul et l'estimation des impacts des règles de représentation et des processus de production calculés sur les vingt itinéraires expérimentés.

Méthodologie de calcul Comme nous l'avons signalé, l'impact de la projection cartographique n'est pas estimé dans cette expérimentation, étant donné qu'il affecte également (de manière quasi-identique) les itinéraires de référence.
Ainsi, le modèle numérique d'élévation SRTM (Farr *et al.*, 2007) est utilisé pour estimer l'impact de la non-prise en compte du terrain (cf. section 3.2). L'ensemble des itinéraires étudiés sont considérés comme des entités courbes, supposant une reconstruction de géométrie courbe par interpolation polynomiale (cf. section 3.3). L'impact de l'erreur de pointé est estimé suite à une estimation des échelles moyennes de saisie (par exploration des distances inter-points, cf. section 4.1.2.2) approximées au 1 :27.000 pour la BDCARTO et au 1 :40.000 pour les données TOP100. Enfin, les échelles de représentation ont été fixées au 1 :50.000 pour la BDCARTO et au 1 :100.000 pour les données TOP100 pour estimer l'impact de la généralisation cartographique (cf. section 4.2).

Etant donné que les itinéraires expérimentés présentent des longueurs hétérogènes, nous proposons de normaliser les erreurs de longueur en les rapportant à la longueur de l'itinéraire étudié. Ceci nous permet d'estimer, pour chaque contribution, une estimation relative de chaque impact (en pourcentage).
Les tables 5.3 et 5.4 présentent respectivement les estimations absolues (en kilomètres) et relatives (en pourcentage de la longueur initiale) de chaque impact calculé à partir des itinéraires extraits de la BDCARTO et des données TOP100.

Analyse des impacts Les impacts des règles de représentation et des processus de production calculés à l'aide des méthodes proposées aux chapitres 3 et 4 apportent des contributions très variables en fonction de la base de données expérimentée, et des caractéristiques des itinéraires traités.

Pour les itinéraires extraits de la BDCARTO, les impacts les plus significatifs concernent la généralisation cartographique et la non-prise en compte en terrain. L'impact de la généralisation cartographique (en nous basant sur l'estimation de l'erreur moyenne estimée), tel que nous l'avons modélisé, constitue l'impact majeur sur 70% des itinéraires étudiés. Cependant, pour certains itinéraires, l'impact de la non-prise en compte du terrain peut s'affirmer de manière très significative, comme dans l'itinéraire 8 notamment, où l'erreur estimée est d'environ 702 mètres, soit 1,8% de la mesure initiale. Dans l'ensemble, l'impact de l'approximation polygonale des courbes et de l'erreur de pointé interviennent de manière secondaire par rapport aux deux impacts précédents.

		Impacts BDCARTO (en km.)							
	Longueur	**Terrain**	**App.Poly.**	**Erreur de pointé**			**Généralisation**		
ID		-	-	Min	Moy	Max	Min	Moy	Max
1	**49,768**	0,366	0,179	-0,027	0,120	0,239	-3,291	0,358	4,006
2	**43,805**	0,641	0,127	-0,009	0,104	0,214	-2,686	0,286	3,258
3	**54,680**	0,209	0,117	0,027	0,108	0,210	-3,012	0,292	3,597
4	**40,619**	0,079	0,056	-0,016	0,057	0,125	-2,242	0,218	2,677
5	**46,461**	0,120	0,071	-0,018	0,055	0,128	-2,625	0,252	3,129
6	**30,845**	0,116	0,060	-0,022	0,042	0,111	-1,706	0,165	2,036
7	**52,749**	0,576	0,152	-0,016	0,102	0,208	-3,198	0,330	3,857
8	**38,880**	0,702	0,164	-0,020	0,112	0,225	-2,494	0,267	3,028
9	**36,187**	0,271	0,107	-0,013	0,078	0,170	-2,258	0,215	2,688
10	**22,847**	0,076	0,052	-0,029	0,048	0,122	-1,299	0,125	1,549
11	**28,959**	0,106	0,046	-0,032	0,040	0,111	-1,667	0,170	2,008
12	**21,928**	0,380	0,106	-0,011	0,065	0,180	-1,582	0,184	1,950
13	**22,018**	0,231	0,137	-0,027	0,079	0,202	-1,569	0,182	1,933
14	**62,274**	0,298	0,190	-0,015	0,131	0,254	-3,708	0,375	4,458
15	**56,819**	0,215	0,160	-0,008	0,110	0,223	-3,233	0,319	3,871
16	**40,683**	0,074	0,063	-0,029	0,047	0,112	-2,301	0,224	2,748
17	**32,364**	0,103	0,076	-0,031	0,058	0,132	-1,791	0,174	2,138
18	**26,496**	0,105	0,046	-0,020	0,035	0,098	-1,562	0,162	1,886
19	**35,033**	0,448	0,028	-0,006	0,035	0,082	-2,066	0,204	2,475
20	**22,570**	0,038	0,019	-0,011	0,025	0,075	-1,234	0,120	1,474

		Impacts BDCARTO (en %)							
	Longueur	**Terrain**	**App.Poly.**	**Erreur de pointé**			**Généralisation**		
ID	-	-	-	Min	Moy	Max	Min	Moy	Max
1	**49,768**	*0,736*	*0,360*	*-0,054*	*0,240*	*0,481*	*-6,612*	*0,719*	*8,049*
2	**43,805**	*1,463*	*0,291*	*-0,020*	*0,237*	*0,489*	*-6,132*	*0,653*	*7,437*
3	**54,680**	*0,381*	*0,215*	*0,048*	*0,198*	*0,384*	*-5,508*	*0,535*	*6,577*
4	**40,619**	*0,193*	*0,138*	*-0,040*	*0,140*	*0,309*	*-5,520*	*0,536*	*6,591*
5	**46,461**	*0,258*	*0,153*	*-0,039*	*0,118*	*0,275*	*-5,649*	*0,543*	*6,734*
6	**30,845**	*0,376*	*0,193*	*-0,071*	*0,135*	*0,359*	*-5,530*	*0,536*	*6,602*
7	**52,749**	*1,092*	*0,288*	*-0,030*	*0,193*	*0,394*	*-6,062*	*0,625*	*7,312*
8	**38,880**	*1,806*	*0,423*	*-0,052*	*0,288*	*0,579*	*-6,414*	*0,687*	*7,788*
9	**36,187**	*0,749*	*0,296*	*-0,037*	*0,216*	*0,470*	*-6,239*	*0,594*	*7,427*
10	**22,847**	*0,331*	*0,228*	*-0,126*	*0,210*	*0,535*	*-5,684*	*0,547*	*6,778*
11	**28,959**	*0,366*	*0,158*	*-0,109*	*0,138*	*0,382*	*-5,757*	*0,589*	*6,934*
12	**21,928**	*1,732*	*0,484*	*-0,051*	*0,297*	*0,822*	*-7,213*	*0,840*	*8,894*
13	**22,018**	*1,049*	*0,624*	*-0,124*	*0,361*	*0,919*	*-7,124*	*0,826*	*8,777*
14	**62,274**	*0,479*	*0,305*	*-0,024*	*0,210*	*0,408*	*-5,954*	*0,603*	*7,159*
15	**56,819**	*0,378*	*0,281*	*-0,015*	*0,194*	*0,393*	*-5,690*	*0,562*	*6,814*
16	**40,683**	*0,181*	*0,156*	*-0,072*	*0,117*	*0,274*	*-5,655*	*0,550*	*6,756*
17	**32,364**	*0,320*	*0,236*	*-0,094*	*0,178*	*0,407*	*-5,533*	*0,536*	*6,605*
18	**26,496**	*0,396*	*0,175*	*-0,075*	*0,132*	*0,370*	*-5,894*	*0,612*	*7,118*
19	**35,033**	*1,278*	*0,079*	*-0,017*	*0,099*	*0,234*	*-5,897*	*0,583*	*7,064*
20	**22,570**	*0,169*	*0,086*	*-0,050*	*0,110*	*0,334*	*-5,466*	*0,533*	*6,532*

Table 5.3 – Estimations absolues et relatives des impacts sur les itinéraires extraits de la BDCARTO

		Impacts TOP100 (en km.)							
	Longueur	**Terrain**	**App.Poly.**	**Erreur de pointé**			**Généralisation**		
ID	-	-	-	Min	Moy	Max	Min	Moy	Max
1	**47,935**	0,394	0,109	-0,014	0,080	0,179	-2,712	2,128	6,968
2	**42,055**	0,784	0,073	-0,004	0,071	0,149	-2,026	1,681	5,388
3	**52,063**	0,201	0,080	0,001	0,066	0,144	-2,127	1,581	5,289
4	**39,396**	0,069	0,040	-0,023	0,037	0,128	-1,630	1,193	4,017
5	**45,425**	0,130	0,055	-0,016	0,041	0,115	-2,011	1,357	4,726
6	**29,616**	0,110	0,045	-0,016	0,032	0,089	-1,231	0,896	3,024
7	**51,036**	0,541	0,090	-0,012	0,071	0,144	-2,489	1,878	6,246
8	**36,154**	0,638	0,083	-0,009	0,064	0,143	-1,954	1,529	5,011
9	**34,389**	0,299	0,080	-0,030	0,059	0,132	-2,002	1,050	4,102
10	**21,715**	0,066	0,045	-0,040	0,028	0,087	-0,984	0,655	2,293
11	**27,788**	0,113	0,031	-0,027	0,032	0,088	-1,202	0,984	3,169
12	**20,269**	0,227	0,045	-0,012	0,045	0,115	-1,292	1,115	3,523
13	**19,786**	0,252	0,103	-0,024	0,067	0,161	-1,237	1,075	3,388
14	**58,610**	0,276	0,122	0,000	0,089	0,196	-2,826	2,048	6,922
15	**53,700**	0,161	0,091	-0,012	0,070	0,145	-2,389	1,740	5,869
16	**39,887**	0,075	0,042	-0,033	0,026	0,082	-1,761	1,236	4,234
17	**30,634**	0,114	0,055	-0,018	0,039	0,110	-1,281	0,926	3,133
18	**25,792**	0,095	0,029	-0,015	0,022	0,061	-1,173	0,969	3,111
19	**34,582**	0,458	0,022	-0,022	0,018	0,064	-1,664	1,139	3,941
20	**22,418**	0,039	0,015	-0,027	0,011	0,050	-0,900	0,683	2,266

		Impacts TOP100 (en %)							
	Longueur	**Terrain**	**App.Poly.**	**Erreur de pointé**			**Généralisation**		
ID	-	-	-	Min	Moy	Max	Min	Moy	Max
1	**47,935**	*0,823*	*0,228*	*-0,029*	*0,166*	*0,372*	*-5,657*	*4,440*	*14,537*
2	**42,055**	*1,863*	*0,175*	*-0,010*	*0,170*	*0,354*	*-4,817*	*3,998*	*12,812*
3	**52,063**	*0,387*	*0,153*	*0,002*	*0,127*	*0,278*	*-4,086*	*3,037*	*10,159*
4	**39,396**	*0,176*	*0,103*	*-0,059*	*0,093*	*0,324*	*-4,137*	*3,029*	*10,196*
5	**45,425**	*0,285*	*0,120*	*-0,036*	*0,091*	*0,254*	*-4,427*	*2,988*	*10,403*
6	**29,616**	*0,371*	*0,151*	*-0,052*	*0,108*	*0,301*	*-4,157*	*3,026*	*10,210*
7	**51,036**	*1,061*	*0,177*	*-0,023*	*0,139*	*0,283*	*-4,876*	*3,681*	*12,238*
8	**36,154**	*1,764*	*0,229*	*-0,024*	*0,176*	*0,394*	*-5,404*	*4,228*	*13,860*
9	**34,389**	*0,869*	*0,232*	*-0,086*	*0,172*	*0,385*	*-5,822*	*3,053*	*11,928*
10	**21,715**	*0,303*	*0,209*	*-0,184*	*0,129*	*0,399*	*-4,530*	*3,015*	*10,559*
11	**27,788**	*0,405*	*0,113*	*-0,099*	*0,116*	*0,318*	*-4,327*	*3,539*	*11,406*
12	**20,269**	*1,118*	*0,221*	*-0,059*	*0,222*	*0,568*	*-6,376*	*5,503*	*17,382*
13	**19,786**	*1,272*	*0,519*	*-0,119*	*0,339*	*0,812*	*-6,252*	*5,435*	*17,123*
14	**58,610**	*0,470*	*0,208*	*0,000*	*0,151*	*0,334*	*-4,821*	*3,495*	*11,811*
15	**53,700**	*0,300*	*0,169*	*-0,022*	*0,130*	*0,269*	*-4,448*	*3,241*	*10,930*
16	**39,887**	*0,187*	*0,104*	*-0,084*	*0,066*	*0,204*	*-4,415*	*3,100*	*10,615*
17	**30,634**	*0,373*	*0,181*	*-0,059*	*0,129*	*0,359*	*-4,181*	*3,023*	*10,227*
18	**25,792**	*0,369*	*0,112*	*-0,057*	*0,085*	*0,238*	*-4,548*	*3,757*	*12,062*
19	**34,582**	*1,323*	*0,065*	*-0,063*	*0,051*	*0,184*	*-4,812*	*3,293*	*11,397*
20	**22,418**	*0,174*	*0,067*	*-0,119*	*0,050*	*0,224*	*-4,014*	*3,047*	*10,107*

Table 5.4 – Estimations absolues et relatives des impacts sur les itinéraires extraits des données TOP100

Pour les itinéraires extraits des données TOP100, la contribution de l'impact de la généralisation cartographique se manifeste de manière hégémonique, noyant totalement les estimations des trois autres impacts. La contribution moyenne de la généralisation cartographique sur l'ensemble des itinéraires correspond à 3,53% de la longueur totale initiale. Les impacts de la non-prise en compte du terrain, de l'approximation polygonale et de l'erreur de pointé interviennent par conséquent de manière secondaire.

En comparant les erreurs relatives sur les itinéraires homologues extraits de la BDCARTO et des données TOP100, nous constatons que les impacts de l'approximation polygonale des courbes et de l'erreur (moyenne) de pointé sont plus importants sur les itinéraires de la BDCARTO que sur ceux extraits des données TOP100. Nous pourrions pourtant supposer que ces impacts sont relativement identiques pour un itinéraire homologue dans les deux bases de données, mais ce n'est pas le cas.
Ceci s'explique principalement par le fait que les itinéraires extraits de la BDCARTO sont moins généralisés que ceux des données TOP100, et sont par conséquent constitués de plus de points. Or, les modèles de simulation de l'erreur de pointé, et de reconstruction de géométries courbes que nous proposons sont sensibles à la densité de points de la poyligne. Ainsi, plus la densité de points est importante, plus les impacts de l'erreur de pointé ou de l'approximation polygonale des courbes augmentent. Ceci explique pourquoi ces deux erreurs sont plus importantes sur les itinéraires de la BDCARTO que sur les données TOP100.

Nous constatons également que les itinéraires présentant les erreurs de référence les plus importantes (en particulier les itinéraires 12, 13, 15 et 17) le sont à la fois dans la BDCARTO et dans les données TOP100. Les itinéraires 12 et 13 des données TOP100, qui présentent des erreurs de référence respectivement de -10,% et -14,2% ont pour caractéristiques d'être localisés dans des zones de montagnes. Ils sont donc impactés significativement par les impacts de la généralisation cartographique et la non-prise en compte du terrain. Egalement, l'itinéraire 13 présente une forte sinuosité, ce qui se traduit par une contribution importante de l'approximation polygonale des courbes, en plus de la généralisation cartographique (qui peut impliquer des suppressions de virages).
Ces observations montrent que pour les données très généralisées, comme peuvent l'être les données TOP100, les impacts du contexte géographique et du caractère sinueux des routes ont pour conséquence d'accumuler les impacts, générant ainsi une erreur de longueur importante.

Nous venons donc de voir qu'en fonction du niveau de généralisation des bases de données, le poids des contributions apportées par chaque impact varie de manière significative. Si l'on peut affirmer que globalement pour les données TOP100, l'impact de la généralisation cartographique agit comme l'impact explicatif de l'erreur de mesure initiale, ce n'est pas le cas pour les itinéraires extraits de la BDCARTO, où les impacts de la non-prise en compte du terrain, voire de l'approximation polygonale apportent des contributions significatives.

Combinaison des impacts

Avant de mettre en œuvre les méthodes de combinaison des impacts (par agrégation et sélection) sur l'échantillon d'itinéraires expérimentés, la méthodologie de combinaison utilisée est tout d'abord présentée.

Méthodologie de combinaison Pour étudier les performances des méthodes de combinaison des impacts proposées dans le cadre de ce travail de thèse, les contributions relatives (à la longueur initiale) de chaque impact sont utilisées, étant donné que les vingt itinéraires étudiés présentent des longueurs hétérogènes. Les estimations globales sont par conséquent exprimées en pourcentage de la longueur initiale évaluée.

Pour les deux méthodes de combinaison proposées, par agrégation et par sélection, les impacts suivants ont été utilisés :
- prise en compte du terrain,
- approximation polygonale des courbes,
- erreur de pointé moyenne estimée,
- erreur moyenne de généralisation cartographique.

Nous n'avons pas souhaité agréger les bornes minimales et maximales de l'impact de la généralisation cartographique, étant donné qu'elles ont pour conséquence de rendre la mesure estimée beaucoup trop imprécise. Nous étudierons malgré tout l'intérêt de les utiliser après analyse des estimations globales.

Etude des erreurs résiduelles Pour étudier la performance des méthodes de combinaison des impacts, les estimations de l'erreur globale, notées Δ_T, sont rapportées à l'erreur de référence de chaque itinéraire (par comparaison avec la longueur courbe en 2D5 de la BDTOPO), notée Δ_{Ref}. Ceci nous permet d'estimer une erreur résiduelle estimée R_e (cf. équation 5.4).

$$R_e = \Delta_T + \Delta_{Ref} \tag{5.4}$$

L'étude des erreurs résiduelles a pour objectif d'évaluer si les estimations globales d'erreurs, par agrégation et par sélection des impacts, permettent de réduire l'erreur de référence.

Les résultats des trois méthodes de combinaison pour les itinéraires extraits de la BDCARTO et des données TOP100 sont présentés dans les tables 5.5 et 5.6. Les estimations totales y sont comparées à l'erreur de référence, afin d'exprimer un résidu estimé pour chaque itinéraire. Ces résidus estimés sont représentés graphiquement aux figures 5.4 et 5.5, permettant ainsi de les comparer visuellement avec les erreurs de référence (correspondant à une erreur de 0%). Dans ces graphiques, l'erreur de référence est représentée avec un trait noir, tandis que les résidus estimés sont représentés par un trait rouge pour la méthode de combinaison par agrégation simple, par un trait vert pour la méthode par agrégation quadratique, et par un trait bleu pour la méthode par sélection de l'impact maximal.

La section suivante nous permettra d'analyser ces résultats, et de comparer les performances respectives des méthodes de combinaison des impacts.

	Erreur Ref.	**Agrég. Simple**		**Agrég. Quad.**		**Sélect. Max.**	
	-	Total	Résidu	Total	Résidu	Total	Résidu
1	**-0,577**	2,054	**1,478**	1,116	**0,539**	0,736	**0,159**
2	**-0,226**	2,644	**2,418**	1,646	**1,420**	1,463	**1,238**
3	**-0,619**	1,329	**0,710**	0,719	**0,100**	0,535	**-0,084**
4	**0,156**	1,007	**1,163**	0,602	**0,758**	0,536	**0,691**
5	**-0,665**	1,072	**0,406**	0,631	**-0,034**	0,543	**-0,123**
6	**-0,238**	1,241	**1,003**	0,696	**0,459**	0,536	**0,299**
7	**-0,554**	2,198	**1,644**	1,305	**0,751**	1,092	**0,538**
8	**-0,434**	3,204	**2,771**	1,999	**1,565**	1,806	**1,372**
9	**-1,596**	1,856	**0,260**	1,024	**-0,572**	0,749	**-0,847**
10	**-0,283**	1,316	**1,033**	0,711	**0,427**	0,547	**0,264**
11	**-0,316**	1,250	**0,934**	0,724	**0,408**	0,589	**0,273**
12	**-2,177**	3,353	**1,176**	2,007	**-0,170**	1,732	**-0,445**
13	**-2,628**	2,860	**0,232**	1,518	**-1,110**	1,049	**-1,579**
14	**-1,600**	1,597	**-0,003**	0,854	**-0,746**	0,603	**-0,998**
15	**-2,895**	1,415	**-1,479**	0,758	**-2,136**	0,562	**-2,333**
16	**-0,797**	1,004	**0,207**	0,611	**-0,186**	0,550	**-0,247**
17	**-2,322**	1,269	**-1,053**	0,691	**-1,631**	0,536	**-1,786**
18	**0,203**	1,314	**1,518**	0,761	**0,965**	0,612	**0,816**
19	**-0,076**	2,040	**1,964**	1,411	**1,335**	1,278	**1,202**
20	**0,089**	0,897	**0,986**	0,576	**0,664**	0,533	**0,621**

Table 5.5 – Combinaison des impacts sur les itinéraires de la BDCARTO

	Erreur Ref.	**Agrég. Simple**		**Agrég. Quad.**		**Sélect. Max.**	
	-	Total	Résidu	Total	Résidu	Total	Résidu
1	**-4,423**	5,656	**1,233**	4,524	**0,101**	4,440	**0,017**
2	**-4,396**	6,206	**1,810**	4,417	**0,022**	3,998	**-0,398**
3	**-5,678**	3,704	**-1,974**	3,068	**-2,610**	3,037	**-2,641**
4	**-2,943**	3,401	**0,458**	3,038	**0,094**	3,029	**0,086**
5	**-2,960**	3,485	**0,524**	3,005	**0,045**	2,988	**0,028**
6	**-4,398**	3,656	**-0,742**	3,055	**-1,343**	3,026	**-1,372**
7	**-3,929**	5,057	**1,128**	3,837	**-0,092**	3,681	**-0,249**
8	**-8,005**	6,397	**-1,608**	4,590	**-3,415**	4,228	**-3,777**
9	**-6,910**	4,326	**-2,584**	3,187	**-3,723**	3,053	**-3,857**
10	**-5,512**	3,655	**-1,857**	3,040	**-2,472**	3,015	**-2,497**
11	**-4,544**	4,173	**-0,371**	3,566	**-0,978**	3,539	**-1,004**
12	**-10,543**	7,065	**-3,478**	5,624	**-4,919**	5,503	**-5,040**
13	**-14,207**	7,565	**-6,643**	5,616	**-8,591**	5,435	**-8,772**
14	**-7,952**	4,324	**-3,628**	3,536	**-4,416**	3,495	**-4,457**
15	**-8,871**	3,840	**-5,031**	3,262	**-5,609**	3,241	**-5,630**
16	**-2,809**	3,457	**0,648**	3,108	**0,299**	3,100	**0,291**
17	**-8,102**	3,706	**-4,396**	3,054	**-5,048**	3,023	**-5,079**
18	**-2,519**	4,323	**1,804**	3,778	**1,259**	3,757	**1,238**
19	**-1,379**	4,731	**3,352**	3,549	**2,170**	3,293	**1,914**
20	**-0,588**	3,338	**2,750**	3,053	**2,465**	3,047	**2,459**

Table 5.6 – Combinaison des impacts sur les itinéraires extraits des données TOP100

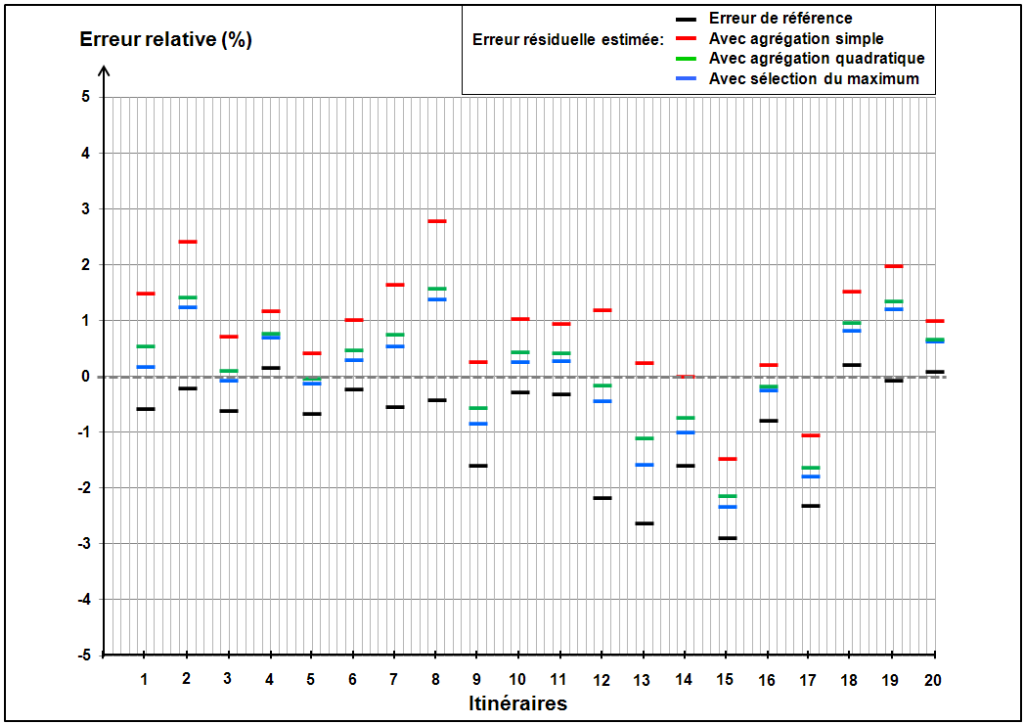


Figure 5.4 – Erreurs relatives de référence et résiduelles estimées sur les itinéraires extraits de la BDCARTO

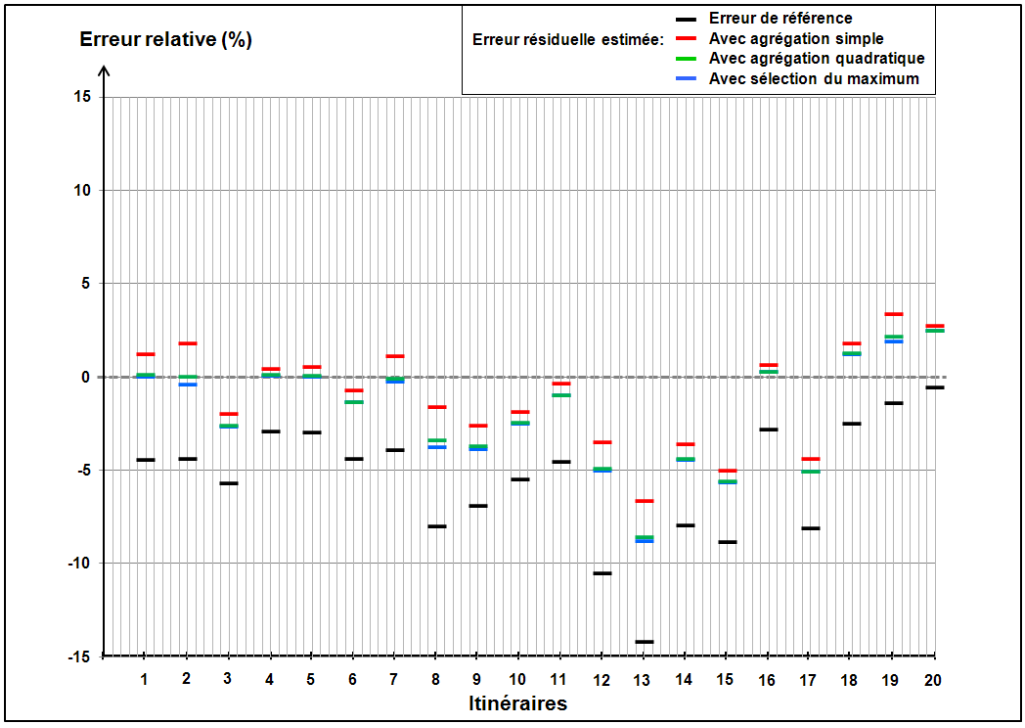


Figure 5.5 – Erreurs relatives de référence et résiduelles estimées sur les itinéraires extraits des données TOP100

5.1.2.3 Comparaison des méthodes de combinaison des impacts

Nous proposons dans cette section d'analyser et de comparer les résultats des estimations globales des erreurs de longueur, calculées par agrégations simple ou quadratique, et par sélection de l'impact le plus significatif. L'objectif de ce travail est d'évaluer la méthode de combinaison la plus performante en fonction des données utilisées, mais également de voir si ces méthodes peuvent s'avérer complémentaires pour une estimation globale de l'erreur de mesure.

Méthodologie de comparaison Pour évaluer la méthode de combinaison des impacts la plus performante, une analyse par moindres carrés des erreurs (initiales ou résiduelles estimées) est proposée sur l'ensemble des itinéraires traités à partir des deux bases de données. L'objectif de cette analyse est de chercher la méthode de combinaison qui permet de minimiser l'erreur résiduelle estimée. Ainsi, pour chaque erreur Δ, correspondant à l'erreur de référence ou au résidu estimé suite à la combinaison des impacts, l'erreur moyenne quadratique totale est exprimée conformément à l'équation 5.5.

$$EMQ = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\Delta)^2} \tag{5.5}$$

En plus de cette estimation globale des erreurs de référence et résiduelles estimées, nous regarderons également pour chaque itinéraire, si l'erreur résiduelle estimée permet de minimiser l'erreur de référence.

Analyse comparative des estimations globales

La table 5.7 présente les erreurs moyennes quadratiques résiduelles estimées sur l'échantillon de vingt itinéraires extraits de la BDCARTO et des données TOP100, en comparaison avec l'erreur moyenne quadratique de référence.

	Erreur	**Résidus**		
	Référence	**Agrég. Simple**	**Agrég. Quadra.**	**Sélection Max**
BDCARTO	**1,288**	1,329	0,975	1,001
TOP100	**6,416**	2,831	3,377	3,439

Table 5.7 – Erreurs moyennes quadratiques de référence et résiduelles estimées

Itinéraires BDCARTO Pour les itinéraires extraits de la BDCARTO, les résultats montrent que seules deux des trois méthodes de combinaison étudiées permettent de minimiser l'erreur de référence. En effet, nous constatons qu'avec la méthode par agrégation simple, l'erreur résiduelle estimée globale est supérieure à l'erreur de référence (1,329% contre 1,288%). Au contraire, la méthode par agrégation quadratique des impacts est celle qui permet de minimiser le plus l'erreur résiduelle estimée globale, devant la méthode par sélection de l'impact le plus significatif, avec respectivement des erreurs de 0,975% et 1,001%.

Cependant, en analysant ces résultats sur chaque itinéraire, la méthode de combinaison par sélection des impacts permet de minimiser l'erreur de référence sur 13 des 20 itinéraires étudiés, tandis la méthode par agrégation quadratique ne réduit cette erreur que sur 10 itinéraires sur 20. Il est important de noter que trois des itinéraires de la BDCARTO présentent une longueur légèrement plus importante que leurs homologues de la BDTOPO, impliquant une maximisation de l'erreur par toutes les méthodes de combinaison. Enfin, la méthode par agrégation simple sur-estime la longueur de référence sur 17 itinéraires, contre 12 pour la méthode par agrégation quadratique, et 11 pour la méthode par sélection de l'impact maximal.

Itinéraires TOP100 Pour les itinéraires extraits des données TOP100, c'est au contraire la méthode par agrégation simple des impacts qui permet de minimiser l'erreur résiduelle estimée globale, par comparaison avec les méthodes par agrégation quadratique et par sélection des impacts (avec respectivement 2,831% contre 3,380% et 3,439%).

Il est cependant important de noter que les résultats des deux méthodes de combinaison sont très proches, étant donné le poids de la contribution de l'impact de la généralisation cartographique. Sur 18 des 20 itinéraires étudiés, l'ensemble des méthodes de combinaison étudiées permettent de minimiser l'erreur de référence, seuls les itinéraires 19 et 20 présentant des résidus estimés supérieurs à l'erreur de référence. Nous observons cependant que seulement dans 9 des 20 itinéraires étudiés, la méthode par agrégation simple sur-estime la longueur de référence.

Ces résultats montrent, que de manière globale, les différentes méthodes de combinaison des impacts permettent de réduire l'erreur de mesure de référence. Cependant, nous voyons bien qu'en fonction des données utilisées, la méthode de combinaison des impacts la plus performante diffère. Nous allons à présent voir comment exploiter ces résultats, afin de produire une estimation globale de l'imprécision des mesures géométriques de longueur.

Exploitation des résultats pour une estimation globale

A la vue des résultats observés précédemment, il est difficile de dégager une tendance forte étant donné l'hétérogénéité des objets et des erreurs traitées. Cependant, dans le cadre de l'élaboration du modèle général, deux scénarios d'estimation globale de l'imprécision des mesures géométriques peuvent être proposés : l'un concernant les données faiblement généralisées (représentées à une échelle supérieure ou égale au 1 :50.000), l'autre concernant les données fortement généralisées (représentées à une échelle inférieure au 1 :50.000). Nous chercherons à formuler l'imprécision globale de la mesure géométrique, évaluée à l'aide d'un intervalle de confiance estimé grâce aux méthodes de combinaison des impacts utilisées précédemment.

Données faiblement généralisées Pour les données faiblement généralisées, comme c'est le cas de la BDCARTO, nous avons vu que la méthode de combinaison par agrégation quadratique des impacts est la plus performante pour minimiser globalement l'erreur de longueur résiduelle. Par conséquent, cette mesure peut être utilisée pour dégager une estimation centrale.

Cependant, pour définir les bornes supérieure et inférieure d'un intervalle de confiance, nous avons vu que la tâche était relativement délicate, puisque :
- 3 itinéraires sur 20, soit 15% de l'échantillon, présentent une erreur positive,
- 3 itinéraires sur 20, soit 15% de l'échantillon, présentent une erreur résiduelle négative suite à l'agrégation simple des impacts.

Ceci signifie que globalement, les méthodes de combinaison des impacts proposées dans cette étude nous permettent d'estimer une mesure de référence avec une confiance de 70%.

Ainsi, dans le cadre de la mise en œuvre du modèle général, pour les données faiblement généralisées, l'estimation globale de l'imprécision des mesures de longueur sera exprimée à l'aide de la mesure initiale M_i, et des erreurs obtenues par agrégation simple Δ_t et quadratique Δ_q, avec une confiance de 70%, comme suit :

Estimation globale pour les données faiblement généralisées :
- Longueur minimale estimée = M_i
- Longueur estimée = $M_i + \Delta_q$
- Longueur maximale estimée = $M_i + \Delta_t$

Données fortement généralisées Pour les données représentées à plus petites échelles, ayant subi un fort impact de la généralisation cartographique, nous avons vu que la méthode par agrégation simple des impacts est la plus performante, et qu'elle permet de minimiser l'erreur de référence dans 90% des cas traités. Nous avons également montré que les longueurs portées par les objets fortement généralisés sous-estiment systématiquement la longueur de référence. Ceci nous permet de fixer aisément la borne inférieure de l'intervalle de confiance.

Par contre, seuls 9 itinéraires sur 20, soit 45% de l'échantillon présentent une erreur résiduelle positive suite à l'agrégation des impacts. Ceci signifie que dans 55% des cas, la longueur de référence est sous-estimée. Ceci n'est pas étonnant, puisque nous avons vu dans la section 4.2.3.3 que l'impact de la généralisation cartographique peut se manifester par des erreurs de longueur très importantes, impliquant la formulation d'un intervalle de confiance très imprécis. Afin de s'assurer de bien encadrer la mesure de référence à partir des mesures calculées sur des données fortement généralisées, la borne supérieure de l'intervalle de confiance de l'impact de la généralisation cartographique peut donc être exploitée.

Ainsi, pour les données fortement généralisées, nous proposons de formuler l'estimation globale de l'imprécision des mesures de longueur à l'aide de la mesure initiale M_i, de l'erreur agrégée totale simple Δ_t, et de l'impact maximal de la généralisation cartographique Δ_g, comme suit :

Estimation globale pour les données fortement généralisées :
- Longueur minimale estimée = M_i
- Longueur estimée = $M_i + \Delta_t$
- Longueur maximale estimée = $M_i + \Delta_g$

Conclusion sur les méthodes de combinaison Les méthodes de combinaison des impacts proposées, malgré leurs imperfections, offrent une solution simple afin de produire une estimation globale de l'imprécision des mesures géométriques de longueur. Même si ces estimations demeurent relativement imprécises, elles permettent expérimentalement de réduire globalement l'erreur de référence.

Cependant, nous venons de voir que la formulation d'un intervalle de confiance n'est pas aisée à produire à l'aide de ces méthodes simples de combinaison, en particulier pour les données faiblement généralisées. Nous avons en effet pu constater que les estimations globales produites ne permettent pas de prendre en compte le cas de figure où la longueur évaluée est plus importante que la longueur de référence. Pour les données fortement généralisées, la prise en compte de la borne maximale de l'impact de la généralisation cartographique peut par contre être exploitée, étant donné que les erreurs de longueurs observées peuvent s'avérer très importantes.

Ainsi, nous venons également de montrer que les deux méthodes d'agrégation des impacts peuvent se montrer complémentaires dans le cadre d'une estimation globale. Ces deux méthodes, malgré leurs limites et leurs imperfections, seront donc exploitées dans le cadre du développement et de la mise en œuvre du modèle général d'estimation de l'imprécision des mesures géométriques. Cependant, nous allons voir dans la section suivante que d'autres méthodes de combinaison, plus ou moins complexes, peuvent être utilisées afin d'estimer une imprécision globale. Ces méthodes seront donc proposées en guise de perspectives.

Synthèse Section Cette section avait pour objectif de comparer les performances des différentes méthodes de combinaison des impacts, par agrégation et sélection de l'impact le plus significatif, dans le but de produire une estimation globale de l'imprécision des mesures géométriques de longueur. Pour cela, un échantillon de 20 itinéraires, extraits de la BDCARTO et des données TOP100 a été utilisé, en comparaison avec les itinéraires homologues de référence de la BDTOPO (enrichie de l'impact de l'approximation polygonale des courbes). Les résultats montrent que la méthode de combinaison par agrégation quadratique des impacts est jugée comme étant la plus performante pour minimiser l'erreur résiduelle estimée sur les itinéraires de la BDCARTO, tandis que pour les données TOP100, il s'agit de la méthode par agrégation simple. Ces résultats ont également permis de formuler l'expression d'un intervalle de confiance sur les mesures calculées, montrant que ces méthodes, bien qu'imparfaites, peuvent s'avérer complémentaires pour produire une estimation globale de l'imprécision des mesures géométriques.

5.1.3 Perspectives pour une estimation globale

Nous avons proposé dans la section précédente des méthodes de combinaison simples des impacts modélisés de manière indépendante, afin d'estimer l'imprécision globale des mesures géométriques. Cependant, nous avons bien conscience que ces méthodes de combinaison des impacts sont réductrices, puisqu'elles ne permettent pas d'appréhender la complexité des interactions entre les différentes composantes de l'erreur de mesure. Nous proposons donc dans cette section d'évoquer d'autres méthodes de combinaison des impacts, qui ne seront pas mises en œuvre dans le périmètre de ce travail de thèse, mais qui peuvent potentiellement offrir des alternatives intéressantes en vue de quantifier l'imprécision globale des mesures géométriques.
Deux approches seront présentées dans cette section. Nous nous intéresserons tout d'abord aux méthodes par apprentissage automatique supervisé, et ensuite à une approche par composition de modèles.

5.1.3.1 Apprentissage automatique supervisé

L'apprentissage automatique supervisé est une technique permettant de produire automatiquement des règles à partir d'une base de données d'apprentissage. Il constitue un des champs d'étude en intelligence artificielle. Différentes méthodes d'apprentissage automatique supervisé peuvent être utilisées en fonction du problème à résoudre. Nous en étudierons deux dans le cadre de l'étude de la combinaison des impacts : les arbres de décision et les réseaux de neurones.

Arbres de décision

L'arbre de décision est un outil très populaire dans les domaines de l'aide à la décision et de l'exploration de données. Après avoir présenté le principe de classification par arbres de décision, nous présenterons l'intérêt de cette méthode dans le cadre de l'estimation de l'imprécision des mesures géométriques.

Définition Un arbre de décision est un outil d'aide à la décision permettant de prédire, sous forme d'arbre, l'appartenance d'un individu à un groupe en fonction de ces caractéristiques et attributs. Un arbre de décision permet de classer les individus à l'aide de questions, correspondant à un ensemble de règles de décision. Chaque nœud de l'arbre correspond à une question portant sur un attribut (une variable explicative), permettant d'orienter le cheminement vers sa résolution. L'évaluation démarre à la racine de l'arbre et termine sur une feuille, qui représente la résolution du problème.

Différents algorithmes de fouille de données (ID3, CART, C4.5, C5 par exemple) permettent de produire automatiquement des arbres de décision, afin de répartir une population d'individus, selon un ensemble de variables discriminantes, et en fonction d'un objectif souhaité. Pour bâtir automatiquement un arbre de décision, il est nécessaire de disposer d'une base de données d'apprentissage, de taille suffisamment conséquente, ainsi que de la solution au problème étudié pour chaque individu de la population, c'est-à-dire la classe à laquelle il appartient.

Application à la combinaison d'erreurs Dans le contexte de l'estimation de l'imprécision des mesures géométriques, les arbres de décision constituent une solution intéressante afin de produire une estimation globale.

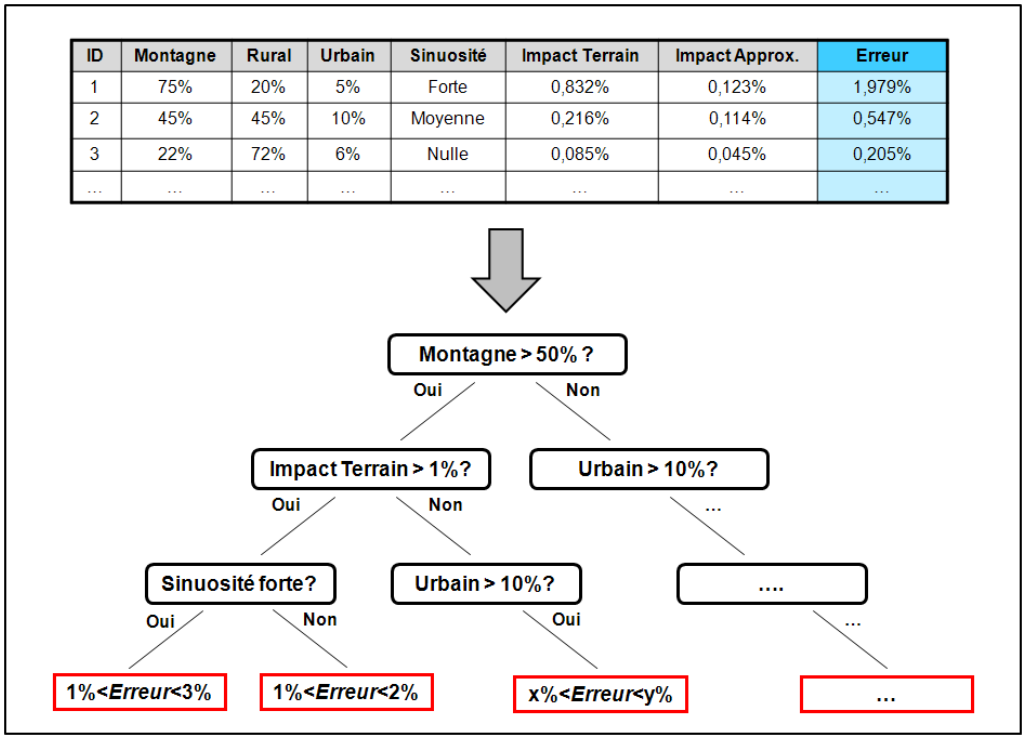


Figure 5.6 – Exemple fictif d'application des arbres de décision à l'estimation de l'imprécision des mesures

L'intérêt des arbres de décision réside dans la possibilité d'estimer un "intervalle d'erreur de mesure" sur des objets évalués, à partir d'un ensemble de variables, qui peuvent être les impacts estimés par les différents modèles, mais également des connaissances sur les données évaluées (cf. figure 5.6). En effet, l'efficacité d'un arbre de décision repose sur l'identification des variables explicatives pertinentes afin de segmenter le plus efficacement possible la population étudiée.

L'identification de ces variables explicatives implique un processus d'acquisition des connaissances et de qualification des données, tel que nous avons pu l'entreprendre au cours de ce travail de thèse (estimation des échelles de saisie et de représentation, contexte géographique des objets, qualification de la sinuosité des routes...).

Cependant, une réflexion doit être entreprise afin d'enrichir les données d'indicateurs explicatifs supplémentaires, pour mettre en œuvre de façon efficiente une évaluation par arbres de décision. De plus, une base de données d'apprentissage suffisamment conséquente et représentative doit pouvoir être mobilisée afin de générer un arbre de décision pertinent par rapport au problème traité.

Réseaux de neurones

La seconde méthode par apprentissage automatique supervisé que nous proposons pour estimer l'imprécision globale des mesures géométriques repose sur les réseaux de neurones artificiels.

Définition Un réseau de neurones artificiels est un modèle de calcul dont la conception est inspirée des neurones biologiques. Le réseau est construit à partir de neurones formels, qui sont des unités de calcul élémentaires inter-connectées. Ces neurones formels reçoivent des signaux en entrée et renvoient une information suite à l'application d'une fonction de transfert.

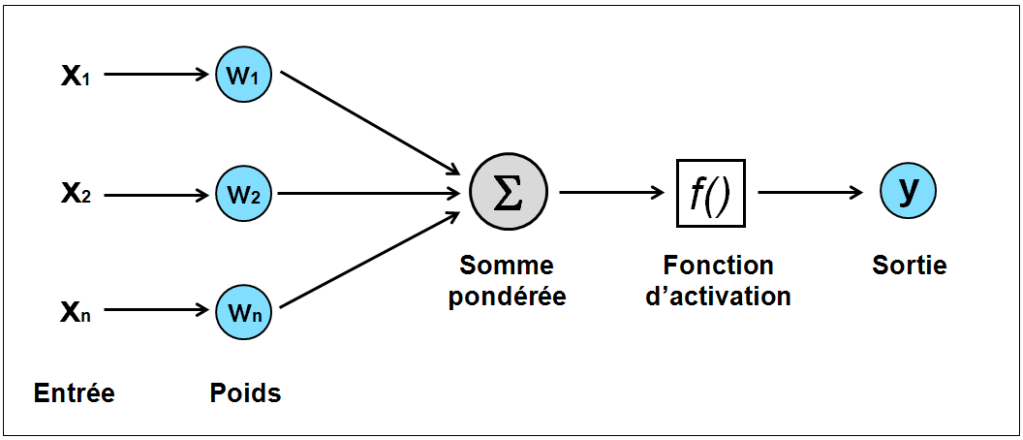


Figure 5.7 – Principe de fonctionnement d'un neurone formel

La figure 5.7 permet de représenter un neurone recevant en entrée des valeurs réelles x_1, $x_2 ... x_n$ et renvoyant en sortie un réel y compris entre 0 et 1. Un neurone formel réalise la somme pondérée des poids w_1, $w_2 ... w_n$ (appelés coefficients synaptiques) des entrées x_1, $x_2 ... x_n$ qu'il reçoit du reste du réseau. La fonction de transfert $f()$ calcule la sortie y en fonction de la somme pondérée. Cette fonction de transfert peut être activée si elle supérieure à un seuil donné.

La détermination d'un réseau de neurones consiste à calculer le poids des coefficients synaptiques w_i appliqué à chaque entrée x_i, à partir d'une base de données d'apprentissage. L'objectif est de minimiser l'erreur quadratique sur les sorties y_i sur l'ensemble des exemples testés.

Application à la combinaison d'erreurs Les réseaux de neurones peuvent constituer une solution intéressante afin de combiner les impacts estimés de manière indépendante, en vue de produire une estimation globale de l'imprécision des mesures géométriques.
Dans le cadre de la phase d'apprentissage automatique, les valeurs x_i correspondraient aux impacts relatifs estimés, et la sortie y à l'erreur de mesure observée par comparaison avec la base de données de référence. La détermination des coefficients synaptique w_i devant permettre d'estimer une erreur de mesure globale, exprimée en termes d'erreur de mesure relative.

Cependant, comme pour l'application des arbres de décision, l'application des réseaux de neurones implique de disposer d'une base de données d'apprentissage suffisamment conséquente et représentative, pour que les estimations calculées soient pertinentes. De la même manière, l'enrichissement par des indicateurs complémentaires des données en entrée du réseau de neurones peut être envisagé. Ceci suppose d'enrichir le processus de qualification des données à l'aide de mesures pertinentes.

Nous venons de proposer deux méthodes, relevant de l'apprentissage automatique supervisé, afin de combiner les impacts estimés de manière indépendante. Une dernière approche de combinaison peut également être proposée pour produire une estimation globale de l'imprécision des mesures géométriques, il s'agit de l'approche par composition de modèles.

5.1.3.2 Composition de modèles

La seconde approche proposée pour produire une estimation globale de l'imprécision des mesures géométriques repose sur une composition ordonnée des modèles d'estimation. Jusqu'à présent, nous avons proposé d'activer chaque modèle de manière indépendante, puis de combiner les estimations produites. Chaque modèle prend donc en entrée les données initiales pour produire une estimation. L'approche par composition de modèles suppose quant à elle que les différents modèles utilisés soient activés de manière ordonnée.

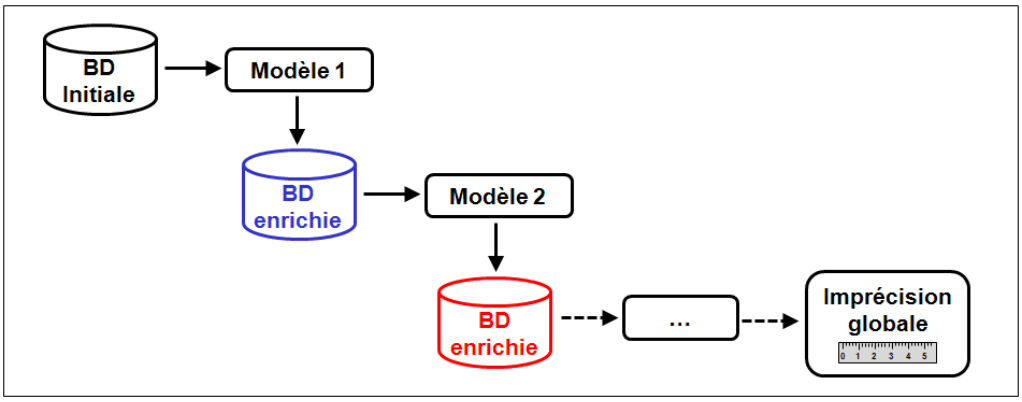


Figure 5.8 – Principe général de l'approche par composition ordonnée de modèles

Comme le montre la figure 5.8, un premier modèle prend en entrée les données initiales et produit des données enrichies, qui serviront en entrée d'un second modèle, et ainsi de suite. Une fois l'ensemble des modèles activés successivement, l'estimation produite par le dernier modèle permet de produire une estimation globale, puisque l'ensemble des impacts sont pris en compte dans le processus d'estimation.

L'équation 5.6 permet de formaliser cette approche, où Gi est la géométrie initiale, $f()$, $g()$ et $h()$ les modèles utilisés pour produire une géométrie enrichie Ge.

$$G_e = f(g(h(G_i))) \tag{5.6}$$

L'imprécision globale est calculée par la différence entre la mesure sur les données enrichies $M(G_e)$ avec la mesure calculée sur les données initiales $M(G_i)$.

La composition de modèles suppose de définir préalablement l'ordonnancement des modèles à activer. Cette question sera discutée dans un premier temps concernant les modèles relatifs aux impacts des règles de représentation, puis dans un second temps sur ceux relatifs aux processus de production.

Intégration de l'impact des règles de représentation

Pour mettre en œuvre la composition des modèles d'impact des règles de représentation, nous proposons dans un premier temps d'enrichir la géométrie des objets évalués à l'aide des méthodes utilisées pour estimer les impacts de la non-prise en compte du terrain et de l'approximation polygonale des courbes. Dans un second temps, le modèle d'impact de la projection cartographique est appliqué sur les objets enrichis.

Enrichissement géométrique Pour mettre en œuvre l'approche par composition de modèles, nous proposons successivement d'enrichir la géométrie initiale des objets évalués :
- d'une information altimétrique (1), si la géométrie des objets n'en est pas pourvue,
- d'une reconstruction de géométrie courbe (2) sur les objets enrichis d'une altitude, si les objets évalués représentent une entité courbe du monde réel.

A ce stade, il est possible de calculer une mesure géométrique de longueur sur une géométrie courbe en 2D5. Nous allons maintenant présenter comment intégrer le modèle d'impact de la projection cartographique.

Projection cartographique La prise en compte de l'impact de la projection cartographique doit être adaptée aux nouvelles caractéristiques des objets, dotés à présent d'information altimétrique. Jusqu'à présent, nous nous sommes contenté de réaliser des mesures sur ellipsoïde (c'est à dire à la surface de l'ellipsoïde) pour prendre en compte l'impact de la projection cartographique.
Cependant, suite à l'enrichissement géométrique, nous ne cherchons plus à calculer la distance sur ellipsoïde D_0, ni même la distance horizontale D_h, mais bien la distance selon la pente D_z, corrigée des effets de la projection cartographique (cf. figure 5.9).

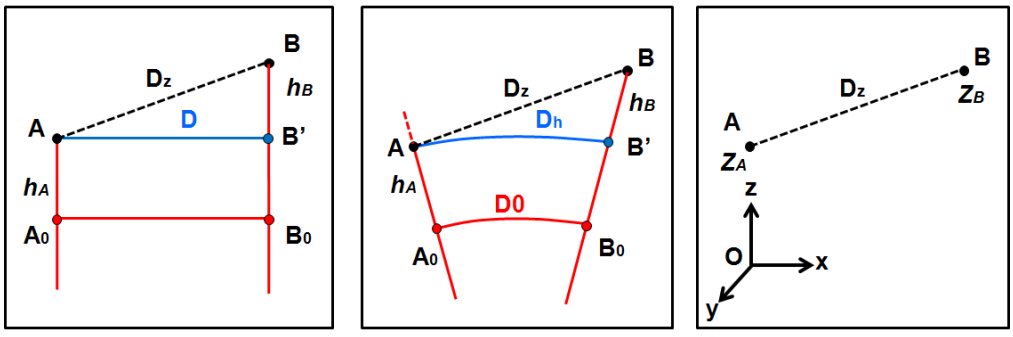


Figure 5.9 – Calcul de la distance entre les points A et B en projection (à gauche), sur ellipsoïde (au centre) et en coordonnées cartésiennes (à droite)

Pour cela, il est nécessaire de convertir les coordonnées projetées (Easting, Northing, Altitude) de chaque point de la géométrie enrichie, en coordonnées cartésiennes géocentriques (X, Y, Z). La conversion du système de coordonnées permet de s'abstraire de l'impact de la projection cartographique pour réaliser des mesures géométriques en 2D5.

Ainsi, nous proposons de formaliser la composition des modèles d'impact des règles de représentation dans l'équation 5.7, où G_i est la géométrie initiale, et G_e la géométrie enrichie par application des méthodes $p()$, $t()$, $c()$ associées respectivement aux modèles d'impacts de la projection cartographique, de la non-prise en compte du terrain, et de l'approximation polygonale.

$$G_e = p(c(t(G_i))) \tag{5.7}$$

Nous allons maintenant voir comment intégrer les modèles d'impact des processus de production au sein de l'approche par composition de modèles.

Intégration de l'impact des processus de production

L'intégration des modèles d'impact des processus de production s'avère plus problématique que pour les règles de représentation, du fait de l'hétérogénéité des méthodes d'estimation mobilisées. Il convient donc de discuter, pour chacun des processus de production impliqués, de la faisabilité de l'intégration des modèles en fonction de la méthode utilisée.

Erreur de pointé L'approche par simulations, utilisée pour modéliser l'impact de l'erreur de pointé, offre l'avantage de produire des géométries simulées, en intégrant un bruit en chaque point de la géométrie initiale. Chaque réalisation d'une simulation présente donc une réalité physique, à travers une géométrie simulée. Il parait donc envisageable, pour chaque réalisation de la simulation, d'appliquer les méthodes d'enrichissement impliquées par les règles de représentation présentées précédemment.
Ainsi, à partir d'une géométrie initiale G_i, une géométrie simulée enrichie G_s est produite pour chaque réalisation de la méthode de simulation de l'erreur de pointé $d()$ et par application des méthodes $p()$, $t()$, $c()$ associées aux impacts des règles de représentation (cf. équation 5.8).

$$G_s = p(c(t(d(G_i)))) \tag{5.8}$$

La figure 5.10 illustre l'intégration du modèle de simulation de l'erreur de pointé dans le processus de composition de modèles, sur un tronçon de route extrait de la BDCARTO.

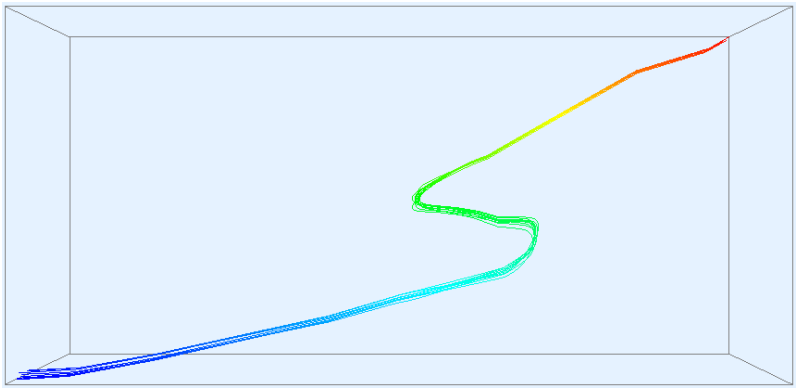


Figure 5.10 – Géométries simulées enrichies d'un tronçon de route de la BDCARTO

Pour chaque géométrie simulée enrichie G_s, une mesure géométrique M_s est calculée. A partir de l'ensemble des réalisations simulées enrichies, l'imprécision globale de mesure peut être exprimée par le biais des mesures géométriques moyenne, minimale et maximale.

Il semble donc envisageable d'intégrer le modèle de simulation de l'erreur de pointé au sein de l'approche par composition de modèles. Nous allons discuter à présent de la faisabilité de l'intégration du modèle d'impact de la généralisation cartographique.

Généralisation cartographique Au contraire de l'approche par simulations, pour laquelle chaque réalisation représente une réalité physique, l'approche par tests, utilisée pour estimer l'impact de la généralisation cartographique, ne se matérialise que sous forme de modèle statistique. Par conséquent, aucune réalité physique n'est associée au processus d'évaluation. Il parait donc difficilement envisageable dans l'état actuel des choses, d'intégrer cette approche dans un scénario de composition de modèles.

Nous venons de présenter les principes régissant l'estimation globale de l'imprécision des mesures géométriques, par le biais d'une approche par composition des modèles mobilisés pour chaque impact. D'un point de vue théorique, la composition des modèles d'impact des règles de représentation et du modèle de simulation de l'erreur de pointé semble envisageable, ce qui n'est pas le cas du modèle d'impact de la généralisation cartographique. Cette approche ne convient donc qu'à des données n'ayant pas subi l'impact de la généralisation cartographique. Afin d'évaluer des données généralisées, une solution serait de combiner les estimations produites par composition des modèles d'un côté, et par le modèle d'impact de la généralisation cartographique de l'autre.

Malgré le principe relativement réaliste de l'approche par composition de modèles, le problème de la sur-estimation de l'impact global peut se poser. En effet, nous savons que les modèles relatifs à la non-prise en compte du terrain et à l'approximation polygonale ont systématiquement tendance à augmenter les mesures calculées. De la même manière, nous avons vu que le modèle de simulation de l'impact de l'erreur de pointé présente un biais positif. Ceci laisse penser que les erreurs cumulées peuvent se multiplier au sein d'un processus par composition de modèles, les estimations risquant par conséquent d'être surestimées. Ces hypothèses nécessitent d'être validées par des expérimentations, qui ne seront pas réalisées dans le périmètre de ce travail de thèse.

Synthèse Section D'autres approches de combinaison des impacts peuvent être envisagées pour produire une estimation globale de l'imprécision des mesures géométriques. Deux méthodes par apprentissage automatique supervisé, les arbres de décision et les réseaux de neurones, on été proposées. Ces méthodes impliquent de disposer d'une base de données d'apprentissage conséquente et représentative du phénomène à observer. Egalement, l'application de ces méthodes suppose d'identifier les attributs pertinents afin de discrétiser efficacement les données traitées, et de bien les classifier. Une seconde approche, par composition de modèles, a également été proposée, permettant d'enchainer les modèles d'impacts de manière ordonnée afin de produire une estimation globale de l'imprécision des mesures. Ces méthodes n'ont pas été mises en œuvre dans le cadre de ce travail de thèse, mais représentent des perspectives de recherche intéressantes dans le domaine de l'estimation de l'imprécision des mesures géométriques.

Conclusion de la section

Afin de communiquer à l'utilisateur une estimation globale de l'imprécision des mesures géométriques calculées à partir de la géométrie des objets évalués, nous avons étudié dans cette section la possibilité de combiner les impacts des règles de représentation et des processus de production, estimés jusqu'à présent de manière indépendante.

Aprés avoir expliqué la difficulté d'entreprendre une telle démarche, nous avons proposé des méthodes simples de combinaison des impacts (par agrégation et par sélection des impacts) permettant de produire une estimation globale de l'imprécision des mesures géométriques.

Ces méthodes ont ensuite été mises en application dans le cadre d'une expérimentation basée sur l'étude du réseau routier. L'objectif de cette expérimentation était d'identifier la méthode de combinaison des impacts la plus performante pour réduire l'erreur de longueur de référence. Les résultats ont montré que pour les données faiblement généralisées, la méthode de combinaison par agrégation quadratique s'avère être la plus performante pour minimiser l'erreur résiduelle estimée, tandis que pour les données fortement généralisées, il s'agit de la méthode par agrégation totale. Cette étude a également montré la complémentarité de ces méthodes de combinaison afin de délimiter un intervalle de confiance à l'estimation globale.

Bien qu'imparfaites, ces méthodes de combinaison des impacts seront donc utilisées dans le cadre du développement du modèle général d'estimation de l'imprécision des mesures géométriques. D'autres méthodes de combinaison des impacts (par apprentissage supervisé et par composition de modèles) ont également été évoquées en perspectives.

Les deux sections suivantes traiterons de la conception du modèle général d'estimation de l'imprécision des mesures géométriques, à travers son principe de modélisation (cf. section 5.2), ainsi que le développement du prototype EstIM (cf. section 5.3).

5.2 Principe de modélisation

Aprés avoir testé et identifié les méthodes de combinaison des impacts retenues dans le cadre de l'élaboration du modèle général d'estimation de l'imprécision des mesures géométriques, cette section aura pour objectif de présenter son principe de modélisation. Le modèle proposé s'organise comme un modèle général intégrant différents modèles, permettant chacun d'estimer les impacts des règles de représentation et des processus de production présentés précédemment. Le diagramme de classes de la figure 5.11, qui avait été proposé initialement en section 2.2.3, présente ce modèle général.

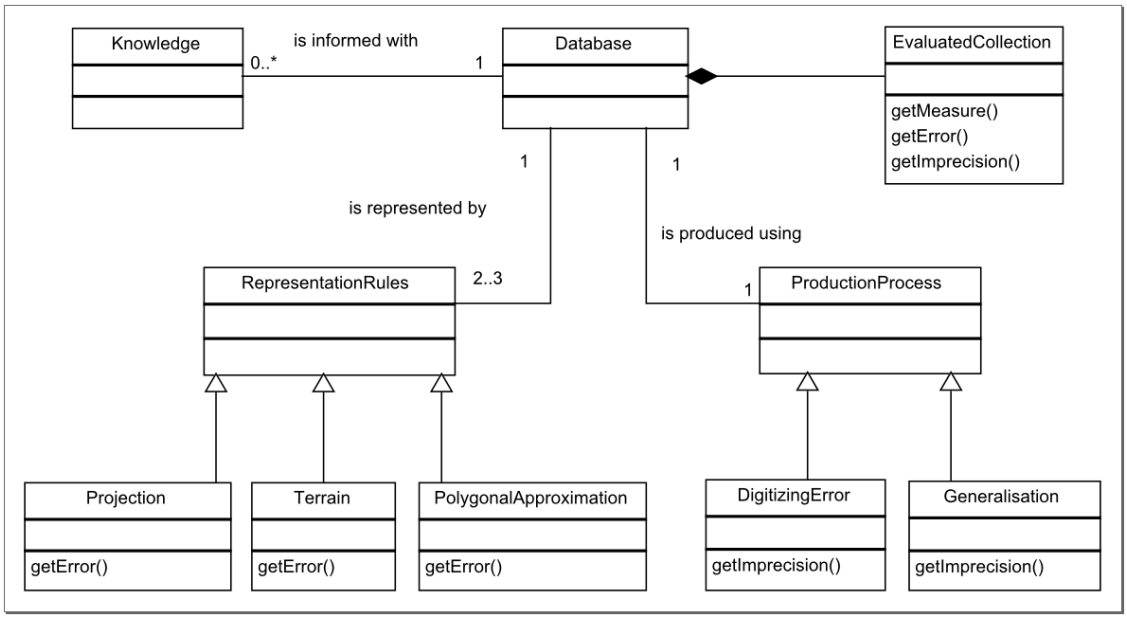


Figure 5.11 – Diagramme de classes du modèle d'estimation de l'imprécision des mesures géométriques

Nous allons décrire dans cette section chacune des composantes de ce modèle, permettant d'estimer les différents impacts étudiés dans le cadre de ce travail de thèse. Préalablement à cela, les principes généraux de la modélisation des géométries des objets géographiques dans notre modèle seront présentés.

5.2.1 Modélisation des géométries

La plate-forme GeOxygene Le développement du modèle d'estimation de l'imprécision des mesures géométriques a été réalisé sur la plateforme open source GeOxygene, issue des développements menés au sein du laboratoire COGIT de l'IGN depuis 2002 (voir Bucher *et al.* (2009), Grosso *et al.* (2012), et Brasebin (2009) pour le module GeOxygene3D).

Le projet GeOxygene s'appuie sur les standards préconisés par l'ISO et l'OGC, notamment en ce qui concerne la modélisation des géométries. La figure 5.12 présente le diagramme des classes relatives à la modélisation des géométries dans GeOxygene. La classe *GM_ Object* en constitue la classe mère.

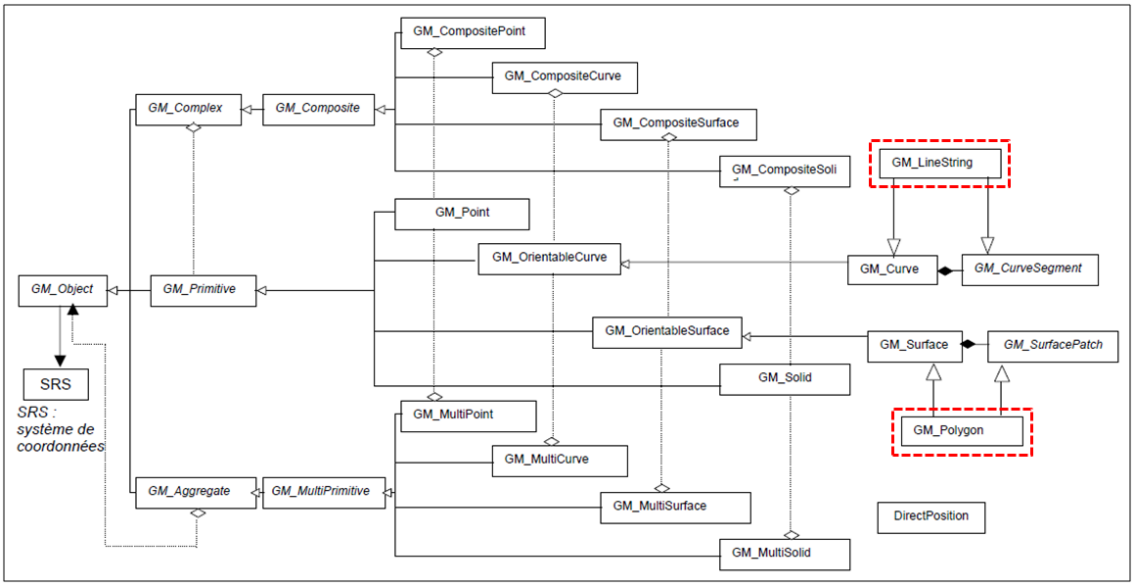


Figure 5.12 – Diagramme de classes relatif à la modélisation des géométries dans GeOxygene

Modélisation des géométries Bien que la modélisation des géométries dans GeOxygene s'appuie sur des standards internationaux, quelques simplifications ont été réalisées. C'est par exemple le cas des classes *GM_ LineString* et *GM_ Polygon*. En effet, les *GM_ LineString* (c'est-à-dire les polylignes) sont directement instanciables dans la librairie GeOxygene, alors que dans la norme ISO 19107 (ISO19107, 2003), elles constituent des *GM_ CurveSegment* particulières. Ceci s'explique par le fait que la majorité des logiciels SIG et SGBD actuels n'offrent pas beaucoup d'autres capacités de stockage pour les géométries linéaires. Pour des raisons de commodité, cette simplification de la norme permet de travailler directement sur la classe *GM_ LineString*. C'est également le cas des *GM_ Polygon*, qui constituent dans la norme ISO 19107 des *GM_ SurfacePatch* particuliers.

Dans le cadre de ce travail de thèse, tous les modèles d'estimation proposés reposent sur une modélisation géométrique des objets à évaluer à partir de ces deux classes. La géométrie des objets linéaires est modélisée sous forme de *GM_ LineString* et celle des objets surfaciques sous forme de *GM_ Polygon*. Dans certains cas, les géométries linéaires des objets évalués peuvent être modélisées sous forme de *GM_ MultiCurve* (un ensemble de *GM_ LineString*) et les géométries surfaciques sous forme de *GM_ MultiSurface* (un ensemble de *GM_ Polygon*), ce qui oblige simplement à les décomposer en un ensemble d'objets *GM_ LineString* ou *GM_ Polygon* pour réaliser les estimations.

Objets à évaluer Les objets géographiques dont nous souhaitons évaluer l'imprécision des mesures géométriques sont modélisés dans la classe *EvaluatedFeature*, spécialisation de la classe *FT_ Feature*, qui est la classe mère des objets géographiques dans GeOxygene. Les objets *FT_ Feature* disposent d'une géométrie (classe *GM_ Object*), qui est soit de type *GM_ Polygon*, soit de type *GM_ LineString*. Les objets à évaluer s'agrègent en collections d'objets à évaluer, dans la classe *EvaluatedCollection*, qui est une spécialisation de la classe *FT_ FeatureCollection* (cf. figure 5.13).

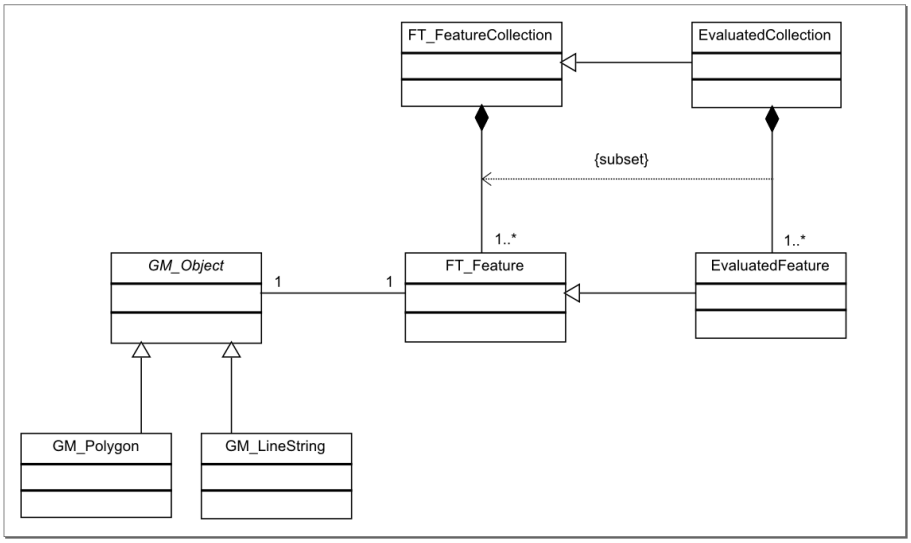


Figure 5.13 – Diagramme de classes relatif à la modélisation de la géométrie des objets à évaluer

Maintenant que les principes de modélisation géométrique ont été décrits, nous allons présenter dans la section suivante le modèle général d'estimation de l'imprécision des mesures géométriques, à travers ses différentes composantes, permettant de modéliser les impacts des règles de représentation et des processus de production.

5.2.2 Description des composantes du modèle d'estimation

Dans le modèle que nous proposons, nous avons décidé de représenter le plus d'information possible sous la forme d'objets et d'attributs d'objets, afin de faciliter la compréhension du modèle par autrui, et assurer son évolution au cours du temps. Ce modèle est en effet destiné à être utilisé au delà de cette thèse de doctorat. Ainsi, le schéma de données proposé pourra être facilement enrichi par de nouvelles causes d'erreurs de mesure.

5.2.2.1 Règles de représentation

Les impacts de trois règles de représentation ont été modélisés dans le cadre de ce travail de thèse (la projection cartographique utilisée, la non-prise en compte du terrain, ainsi que l'approximation polygonale des courbes). Nous allons détailler chacun des modèles permettant de modéliser leurs impacts.

a. Projection cartographique

Pour estimer l'impact de la projection cartographique sur les mesures géométriques, rappelons tout d'abord que cet impact correspond à l'écart entre la mesure projetée et la mesure sur ellipsoïde. La mesure sur ellipsoïde n'est ici pas calculée directement, mais en utilisant des grilles d'altération linéaire existantes, qui permettent d'enrichir la géométrie des objets vectoriels à évaluer (cf. section 3.1.2). Une classe *AlterationGrid* permet de stocker ces informations.

Enrichissement géométrique Pour procéder à l'enrichissement de la géométrie d'une collection d'objets évalués (classe *EvaluatedCollection*) par une information d'altération (stockés dans l'attribut *Alteration* de la classe *AlterationPoint*), une spécialisation des objets évalués (classe *EvaluatedFeature*) sous forme d'objets enrichis (classe *EnrichedFeature*) est proposée. L'enrichissement des objets s'effectue par le biais de la méthode *getAlteration()*.
La méthode *getAlteration()* permet, par le biais d'une méthode d'interpolation IDW, de récupérer l'information d'altération contenue dans la grille d'altération (classe *AlterationGrid*), et de la calculer en tout point. Dans le cas des géométries linéaires, la valeur d'altération linéaire est donc calculée en chaque point de la géométrie. Dans le cas des géométries surfaciques, un suréchantillonnage préalable de la géométrie initiale est effectué (en se basant sur la grille d'altération), l'information d'altération étant calculée au centroïde de chacun des polygones ainsi créés.

Estimation de l'impact L'estimation de l'impact de de la projection cartographique s'effectue par le biais des méthodes suivantes :
- *getEllipsoideMeasure()* : calcule une mesure ellipsoïdale sur un objet enrichi,
- *getMeasure()* : calcule une mesure projetée sur un objet évalué,
- *getProjectionImpact()* : calcule l'impact de la projection cartographique sur la collection d'objets évaluée, par la somme des différences entre mesures projetées et mesures sur ellipsoïde de chaque objet évalué.

L'ensemble de ces concepts est présenté dans le diagramme de classes de la figure 5.14.

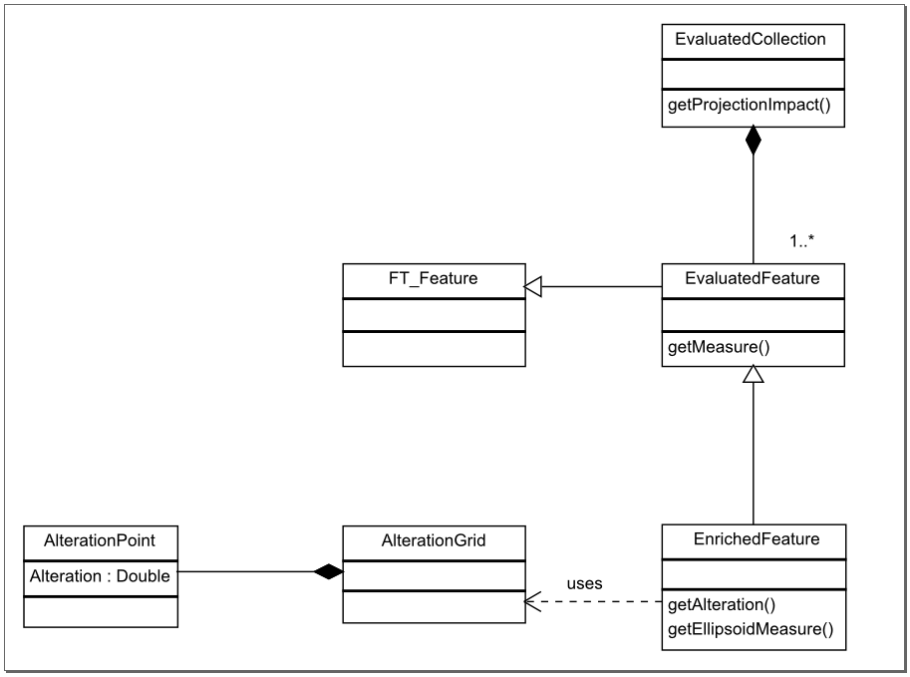


Figure 5.14 – Diagramme de classes du modèle d'estimation de l'impact de la projection cartographique

b. Prise en compte du terrain

L'estimation de l'impact de la non-prise en compte du terrain correspond à la différence entre mesures géométriques en 2D5 et en 2D. Cette estimation suppose également un enrichissement de la géométrie des objets évalués, en leur affectant une information altimétrique provenant d'un Modèle Numérique de Terrain (cf. section 3.2.2).

Enrichissement géométrique Une spécialisation des objets évalués (classe *EvaluatedFeature*) est également proposée, sous forme d'objets enrichis (classe *EnrichedFeature*). La méthode *getAltitude()* permet de récupérer les informations altimétriques stockées dans les cellules du Modèle Numérique de Terrain (dans l'attribut *Altitude* de la classe *DTMCells*), suite à une triangulation préalable du Modèle Numérique de Terrain (classe *DTM*). Un suréchantillonnage de la géométrie des objets évalués est réalisé, à partir de la triangulation du MNT. L'affectation des altitudes est effectuée en chacun des points composant la géométrie des objets suréchantillonnés.

Estimation de l'impact L'estimation de l'impact de la non-prise en compte du terrain s'effectue par le biais des méthodes suivantes :
- *get2D5Measure()* calcule une longueur géométrique de longueur ou de surface en 2D5 sur chaque objet enrichi,
- *getMeasure()* : calcule une mesure en 2D sur chaque objet évalué,
- *getTerrainImpact()* calcule l'impact de la non-prise en compte du terrain sur la collection d'objets évalués, par la somme des différences entre mesures en 2D5 et en 2D.

La figure 5.15 permet de représenter l'ensemble de ces concepts.

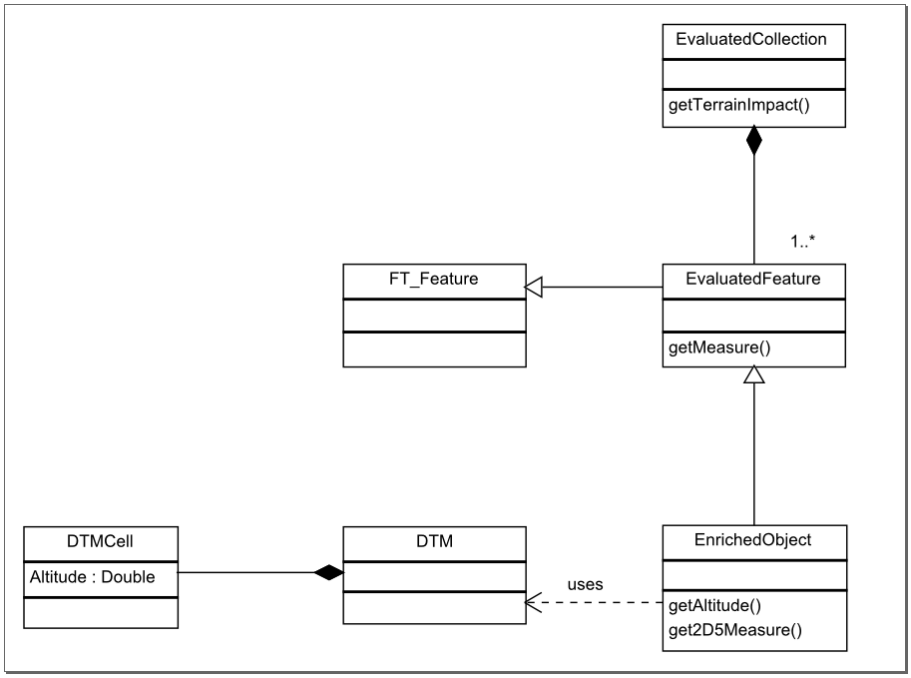


Figure 5.15 – Diagramme de classes du modèle d'estimation de l'impact de la non-prise en compte du terrain

c. Approximation polygonale des courbes

L'estimation de l'impact de l'approximation polygonale des objets courbes n'est effectuée que sur des géométries linéaires dans le cadre de ce travail de thèse. Cet impact correspond à la différence entre la longueur calculée sur la géométrie courbe avec la longueur calculée sur la ligne polygonale initiale (cf. section 3.3.2). Ainsi, pour estimer l'impact de l'approximation polygonale des courbes sur les mesures géométriques de longueur d'une collection d'objets évalués, une spécialisation des objets évalués sous forme d'objets courbes (classe *CurveFeature*) est proposée.

Création des géométries courbes Pour créer une géométrie courbe, plusieurs méthodes sont utilisées :

- la méthode *getGranularity()* calcule la longueur médiane de l'ensemble des segments de la collection d'objets évalués (classe *EvaluatedCollection*). Cette longueur est utilisée pour paramétrer le suréchantillonnage des objets évalués afin de reconstruire des géométries courbes en évitant les exagérations locales.
- la méthode *isCurve()* permet de définir si une géométrie courbe est instanciable, ou non, sur un objet évalué. Cette méthode qualifie si un objet évalué représente une entité courbe du monde réel, en fonction de la classe d'objets concernée. Par exemple, pour les objets routiers, nous proposons de détecter les objets au caractère sinueux (cf. section 3.3.2.3).

Pour chaque objet représentant une entité courbe (*isCurve()* = *true*), une géométrie courbe de type *GM_ CubicSpline* est calculée sur la polyligne enrichie.

Estimation de l'impact L'estimation de l'impact de l'approximation polygonale sur les mesures de longueur s'effectue par le biais des méthodes suivantes :

- *getCurveMeasure()* : calcule la longueur géométrique de chaque objet courbe,
- *getMeasure()* : calcule la longueur de la polyligne initiale,
- *getCurvatureImpact()* calcule l'impact de l'approximation polygonale des courbes sur la collection d'objets évalués, par la somme des différences entre mesures de longueur polygonale et mesures de longueur courbe sur chaque objet évalué présentant une géométrie courbe.

La figure 5.16 présente l'ensemble des concepts relatifs à l'estimation de l'impact de l'approximation polygonale.

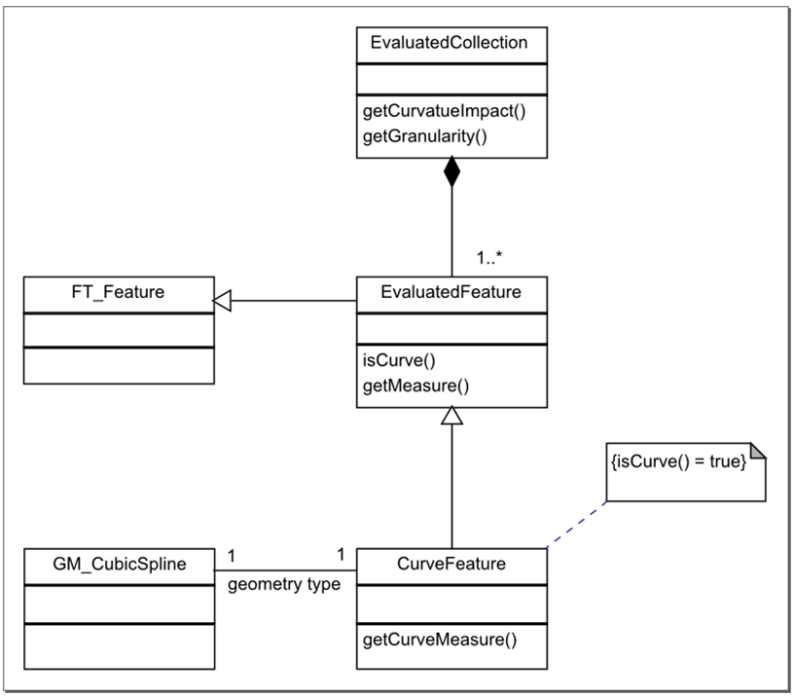


Figure 5.16 – Diagramme de classes du modèle d'estimation de l'impact de l'approximation polygonale des courbes

5.2.2.2 Processus de production

Nous allons présenter dans cette section la modélisation de l'estimation des impacts de l'erreur de pointé et de la généralisation cartographique.

a. Erreur de pointé

Pour modéliser l'impact de l'erreur de pointé, une approche par simulations est utilisée (cf. section 4.1). Cette approche consiste à simuler l'erreur de pointé de l'opérateur sur les objets évalués par le biais d'un bruitage des points composant leur géométrie, afin d'étudier la sensibilité de la mesure géométrique initiale.

Simulation de l'erreur de pointé Pour mettre en œuvre une simulation sur une collection d'objets évalués (classe *EvaluatedCollection*), des objets simulés (classe *SimulatedFeature*) sont créés à partir de chaque objet évalué (classe *EvaluatedFeature*) par le biais d'un bruitage de la géométrie. L'ensemble des objets simulés compose une collection d'objets simulés (classe *SimulatedCollection*).
L'amplitude du bruit est définie par la connaissance de l'échelle de saisie de la base de données. Si l'échelle de saisie des données n'est pas connue, la méthode *getScale()* permet de l'estimer (par exploration des distances inter-points de la collection d'objets évalués). Le principe d'estimation de l'échelle de saisie utilisé pour cette méthode est décrit dans la section 4.1.2.2.

Estimation de l'imprécision Pour estimer l'imprécision des mesures géométriques impliquée par l'erreur de pointé, la collection d'objets simulés (classe *SimulatedCollection*) est exploitée. Trois méthodes sont utilisées pour exprimer l'imprécision de mesure géométrique :
– *getMeanMeasure()* : calcule la mesure géométrique moyenne des objets simulés,
– *getMinMeasure()* : calcule la mesure géométrique minimale des objets simulés.
– *getMaxMeasure()* : calcule la mesure géométrique maximale des objets simulés.

La méthode *getDigitizingImpact()* de la classe *EvaluatedCollection* permet d'estimer l'impact de l'erreur de pointé sur la collection d'objets évalués, à partir des trois méthodes précédentes, en comparant les mesures simulées avec la mesure géométrique initiale.

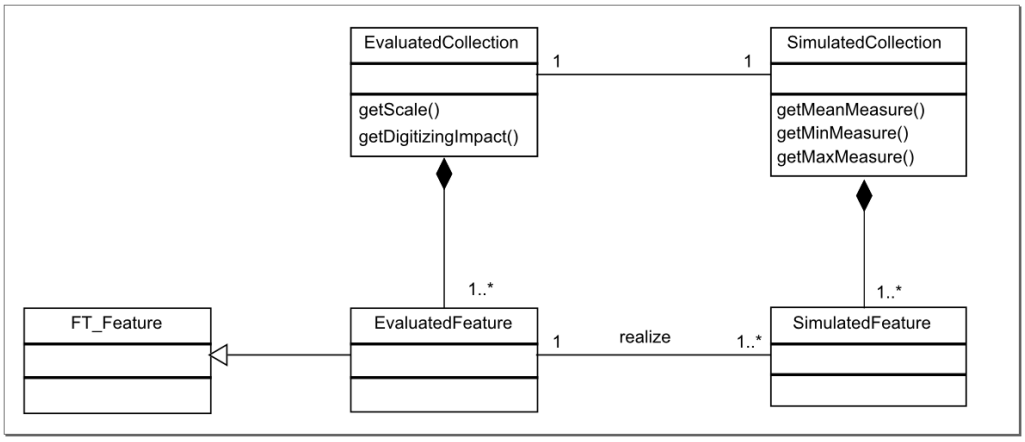


Figure 5.17 – Diagramme de classes du modèle d'estimation de l'impact de l'erreur de pointé

b. Généralisation cartographique

Pour modéliser l'impact de la généralisation cartographique sur les mesures géométriques, une approche par tests a été suivie (cf. section 4.2.1.3). Cette approche consiste à comparer les mesures géométriques réalisées sur des données généralisées avec celles réalisées sur des données de référence, afin d'établir un modèle statistique d'imprécision des mesures, en fonction de différents paramètres (classe d'objets, échelle de représentation, contexte géographique). Le modèle statistique ainsi établi est ensuite exploité afin d'estimer l'imprécision des mesures géométriques calculées sur une collection d'objets à évaluer.

Elaboration du modèle statistique Le modèle statistique d'estimation de l'impact de la généralisation cartographique sur les mesures géométriques s'appuie sur des comparaisons entre des objets généralisés (classe *GeneralisedFeature*) et des objets de référence (classe *ReferenceFeature*). Ces objets composent respectivement des collections d'objets généralisées (classe *GeneralisedCollection*) et des collections d'objets de référence (classe *ReferenceCollection*). Les objets (généralisés et de référence) sont localisés dans un contexte géographique (classe *GeographicalContext*) qui peut être urbain, rural, ou de montagne.

Les comparaisons de mesures géométriques (classe *Comparisons*) entre objets généralisés et de référence sont effectuées suite à une procédure d'appariement et de pré-traitement des objets (afin de permettre des comparaisons normalisées). Les comparaisons permettent l'élaboration du modèle statistique (classe *StatisticalModel*).

Le modèle statistique permet d'estimer :

- l'erreur moyenne de mesure, avec la méthode *getMeanError()*,
- les bornes inférieures et supérieures de l'intervalle de confiance, par les méthodes *getErrorMin()* et *getErrorMax()*.

en fonction de :

- l'échelle de représentation (attribut *scale*),
- la classe d'objets (attribut *objectType*),
- le contexte géographique (attribut *geographicalContext*).

Avec cette modélisation, il est possible de créer différents modèles statistiques stockés, qui seront utilisés en fonction des besoins en termes d'évaluation.

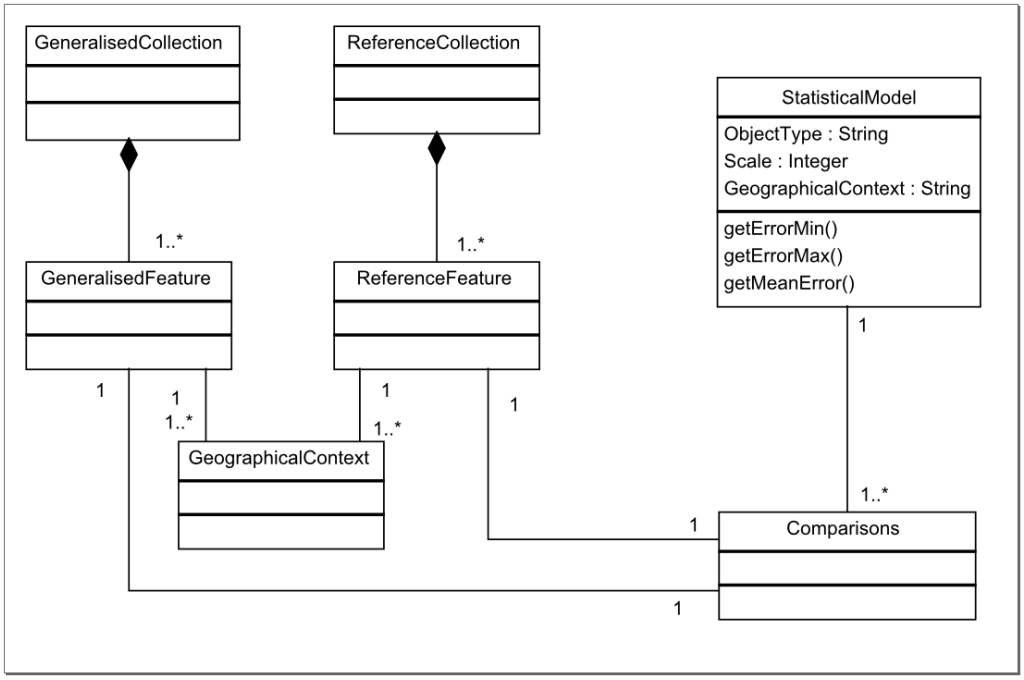


Figure 5.18 – Diagramme de classes du modèle de comparaison des mesures géométriques entre données généralisées et données de référence

Application du modèle statistique Pour estimer l'impact de la généralisation cartographique sur les mesures géométriques, un modèle statistique (classe *StatisticalModel*) est appliqué sur une collection d'objets évalués (classe *EvaluatedFeature*).

La collection d'objets évalués représente une classe d'objets géographiques particulière (attribut *ObjectType*), comme des tronçons de route par exemple, et est caractérisée par une échelle de représentation (attribut *scale*). Pour estimer l'échelle de représentation, si celle-ci est inconnue, une méthode *getScale()* a été créée dans la classe *EvaluatedCollection*. Le principe d'estimation de l'échelle de représentation est décrit en section 4.2.2.1.
Chaque objet à évaluer (classe *EvaluatedFeature*) composant la collection d'objets évalués, est localisé dans un contexte géographique particulier. Pour délimiter le contexte géographique (urbain, rural ou de montagne), si celui-ci est inconnu, des méthodes automatiques de délimitation sont proposées dans la section 4.2.2.2.

L'estimation de l'impact de la généralisation cartographique sur les mesures géométriques est donc réalisée à l'aide de la méthode *getGeneralisationImpact()*, par application d'un modèle statistique, en fonction des paramètres utiles à l'évaluation. Pour les géométries linéaires, nous avons retenu l'échelle de représentation, la classe d'objets et le contexte géographique. Pour les géométries surfaciques, le contexte géographique n'est pas pris en compte car il n'a pas semblé discriminant.

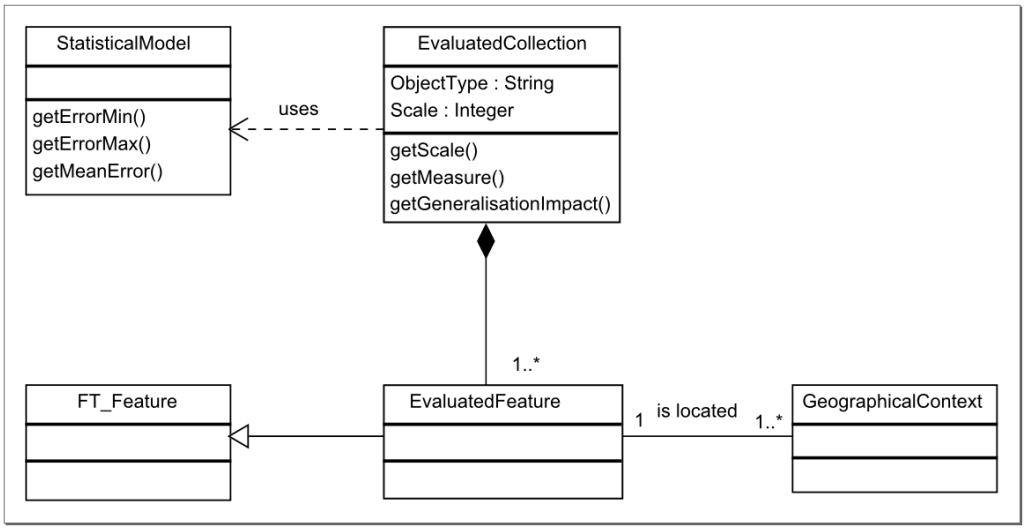


Figure 5.19 – Diagramme de classes du modèle d'estimation de l'impact de la généralisation cartographique

Le modèle général d'estimation de l'imprécision des mesures géométriques permet donc d'intégrer les différents modèles présentés précédemment dans un modèle global. Une architecture modulaire a donc été envisagée afin de pouvoir enrichir ce modèle général d'autres impacts de règles de représentation ou de processus de production qui n'ont pas été abordés dans le cadre de ce travail de thèse.

La modélisation des différentes composantes du modèle d'estimation étant à présent réalisée, nous allons présenter dans la section suivante comment est modélisée la combinaison des modèles pour l'estimation de l'imprécision globale de mesure.

5.2.2.3 Combinaison des impacts

Nous avons présenté jusqu'à présent la modélisation de chacune des composantes permettant d'estimer les impacts des règles de représentation et des processus de production. Ceci se justifie car nous souhaitons dans un premier temps pouvoir les estimer de manière indépendante. Cependant, comme nous l'avons vu, nous souhaitons également pouvoir combiner ces impacts afin de produire une estimation globale de l'imprécision de mesure.

Comme nous l'avons proposé dans la section 5.1, des méthodes de combinaison simples des impacts (par agrégation) ont été proposées pour produire une estimation globale. Ce sont ces méthodes qui ont été retenues pour le développement du modèle d'estimation. Cependant, d'autres méthodes de combinaison ont également été abordées en perspectives, notamment l'approche par composition de modèles. Bien que cette dernière approche n'ait pas été retenue dans le périmètre de ce travail de thèse, nous exposerons malgré tout son principe de modélisation.

a. Combinaison des impacts

Le principe de modélisation de la combinaison des impacts est très simple à implémenter, puisqu'il se base sur l'agrégation des impacts modélisés de manière indépendante.

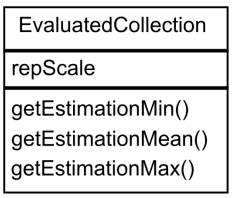


Figure 5.20 – Les méthodes utilisées pour combiner les impacts sélectionnés

Estimation globale de l'imprécision Ainsi, pour estimer l'imprécision globale des mesures géométriques, la classe *EvaluatedCollection* est utilisée.
Nous avons vu que l'estimation moyenne, ainsi que les bornes inférieures et supérieures de l'intervalle de confiance, ne sont pas identiques en fonction de l'échelle de représentation des données, et en particulier de son niveau de généralisation. L'attribut *repScale()* permet donc de caractériser l'échelle de représentation de la collection d'objets évalués.

Trois méthodes sont proposées pour estimer l'imprécision globale d'une mesure géométrique, en fonction de l'échelle de représentation des données :
- *getEstimationMean()* : calcule la mesure géométrique estimée (par agrégation quadratique des impacts pour les données peu généralisées, et par agrégation simple des impacts pour les données fortement généralisées),
- *getEstimationMin()* : calcule la mesure géométrique minimale estimée (correspondant à la mesure initiale dans les deux cas de figure),
- *getEstimationMax()* : calcule la mesure géométrique maximale estimée (correspondant à l'agrégation simple des impacts pour les données peu généralisées, et à l'impact maximal de la généralisation pour les données fortement généralisées).

En termes de modélisation, ce principe d'estimation globale offre l'avantage notable de pouvoir enrichir facilement le modèle général d'impacts supplémentaires (comme pourrait l'être un modèle d'impact de l'imprécision des acquisitions GNSS par exemple).

b. Composition de modèles

Le principe de modélisation de la combinaison des impacts par composition de modèles est quant à lui plus délicat à implémenter et surtout à enrichir de nouveaux impacts. Nous avons d'ailleurs fait remarquer que l'impact de la généralisation cartographique n'était pas pris en compte dans cette approche.

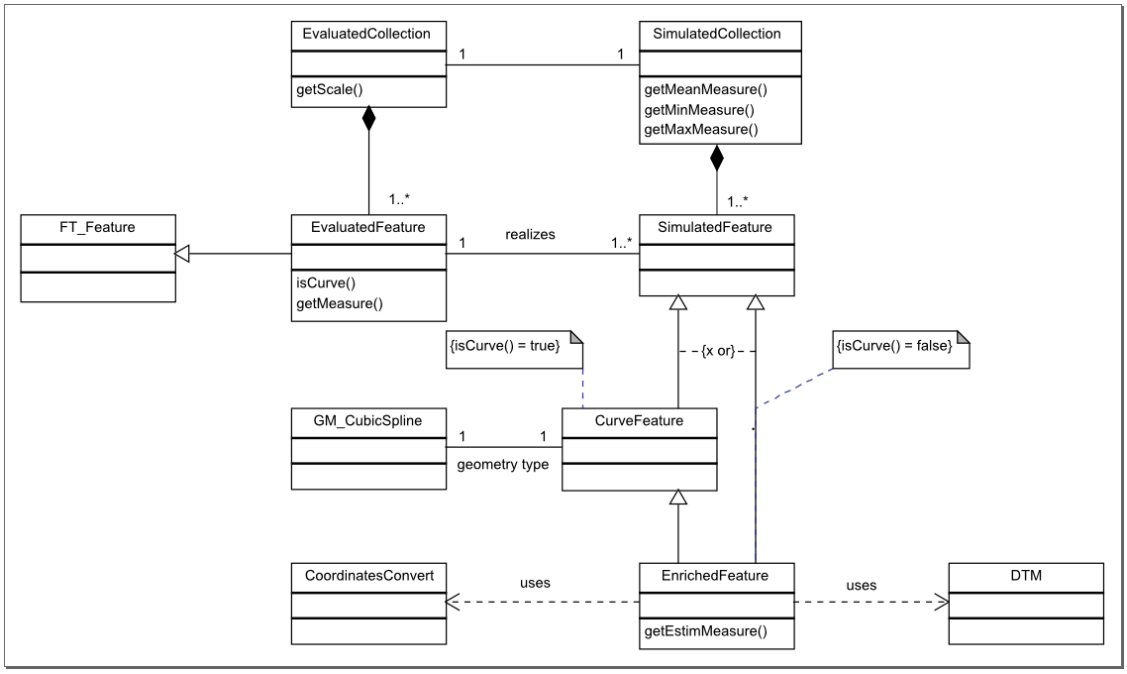


Figure 5.21 – Diagramme de classes du modèle d'estimation globale par composition de modèles

Composition des modèles Pour estimer l'imprécision globale des mesures géométriques avec l'approche par composition de modèles, une simulation de l'erreur de pointé doit tout d'abord être réalisée sur chaque objet (classe *EvaluatedFeature*) de la collection d'objet évalués (classe *EvaluatedCollection*).

Pour chaque objet simulé (classe *SimulatedFeature*), un objet enrichi est ensuite crée (classe *EnrichedFeature*) par affectation d'une information altimétrique à l'aide d'un modèle numérique de terrain (classe *DTM*) pour prendre en compte l'impact de la non-prise en compte du terrain.
Si l'objet enrichi représente une entité courbe du monde réel (méthode *isCurve()=true*), un objet enrichi courbe est crée (classe *CurveFeature*).
Enfin, pour prendre en compte l'impact de la projection cartographique, une classe de conversion des coordonnées projetées en coordonnées cartésiennes géocentriques est utilisée (classe *CoordinatesConvert*).

Après enrichissement géométrique, la méthode *getEstimMeasure* permet de calculer une mesure géométrique estimée sur chaque objet enrichi, prenant en compte les différents impacts sélectionnés.

Estimation globale de l'imprécision Pour estimer l'imprécision globale des mesures géométriques en utilisant l'approche par composition de modèles, la collection d'objets simulés (classe *SimulatedCollection*) est exploitée.
Trois méthodes sont utilisées pour exprimer l'imprécision globale de la mesure géométrique :
- *getMeanMeasure()* : calcule la mesure géométrique moyenne des objets simulés enrichis,
- *getMinMeasure()* : calcule la mesure géométrique minimale des objets simulés enrichis,
- *getMaxMeasure()* : calcule la mesure géométrique maximale des objets simulés enrichis.

Après avoir décrit les principes de modélisation du modèle général d'estimation de l'imprécision des mesures géométriques, la section suivante sera consacrée à son développement, à travers le prototype EstIM (Estimation de l'Imprécision des Mesures).

Synthèse Section Nous avons présenté dans cette section les principes de modélisation retenus dans le cadre de l'élaboration du modèle général d'estimation de l'imprécision des mesures géométriques. La principe général de modélisation des géométries a été présenté dans un premier temps. Dans un second temps, l'ensemble des classes nécessaires à l'estimation et la combinaison des impacts des règles de représentation et des processus de production ont été décrites. Le modèle général ainsi proposé présente une architecture modulaire, aisée à enrichir de modèles d'impact supplémentaires.

5.3 Développement du prototype EstIM

Cette section a pour objectif de présenter les choix relatifs au développement du modèle d'estimation de l'imprécision des mesures géométriques, en focalisant davantage sur ses aspects fonctionnels. Le modèle proposé est implémenté au sein du prototype EstIM (Estimation de l'Imprécision des Mesures). Ainsi, nous présenterons tout d'abord le fonctionnement envisagé du modèle EstIM, les interfaces homme-machine développées, et terminerons par un exemple de mise en œuvre.

5.3.1 Fonctionnement général du modèle

Le modèle général d'évaluation a été développé afin qu'un utilisateur de données géographiques puisse aisément opérer à une estimation de l'imprécision des mesures calculées sur une collection d'objets. Ainsi, le processus d'estimation peut être décomposé en trois activités principales (cf. figure 5.22) :
- la sélection des objets à évaluer,
- la sélection et l'estimation des impacts à prendre en compte,
- la combinaison des impacts.

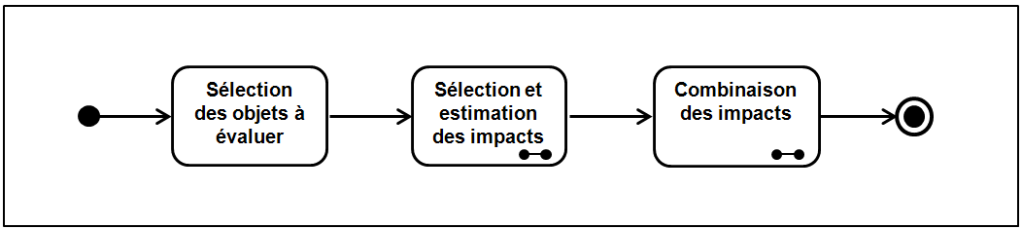


Figure 5.22 – Diagramme d'activité de haut niveau du modèle général d'estimation

Nous allons décrire plus précisément les choix retenus dans ces trois activités principales, afin d'assurer le bon fonctionnement du processus d'estimation.

5.3.1.1 Sélection des objets à évaluer

L'utilisation du modèle général d'estimation se veut être la plus flexible possible. Différentes exigences ont pour cela été formulées.

Tout d'abord, le modèle d'estimation doit pouvoir être exploité directement sur les objets d'une base de données, ou bien à la suite d'un traitement, comme un calcul d'itinéraire par exemple. Egalement, le processus d'estimation doit pouvoir être envisagé sur tous les objets de la base de données, ou bien uniquement sur une sélection (manuelle ou automatique) d'objets.

Afin de prendre en considération ces différents cas d'utilisation, nous proposons de ne procéder à l'estimation de l'imprécision des mesures que sur une **collection d'objets sélectionnés**. Cette option évite de devoir charger toute la base de données pour procéder au processus d'évaluation. Nous n'envisageons pour le moment que des **sélections manuelles** d'objets, pouvant par contre être le résultat de traitements préalables.

Ce choix dans le fonctionnement général du modèle laisse donc une part de choix à l'utilisateur. Dans le même temps, ceci garantit une flexibilité dans l'utilisation du modèle qui satisfait nos exigences.

5.3.1.2 Sélection et paramétrage des impacts

Une fois les objets sélectionnés par l'utilisateur, nous souhaitons le laisser libre de sélectionner les impacts qu'il désire estimer. Ainsi, un à cinq impacts peuvent être sélectionnés simultanément (cf. figure 5.23).

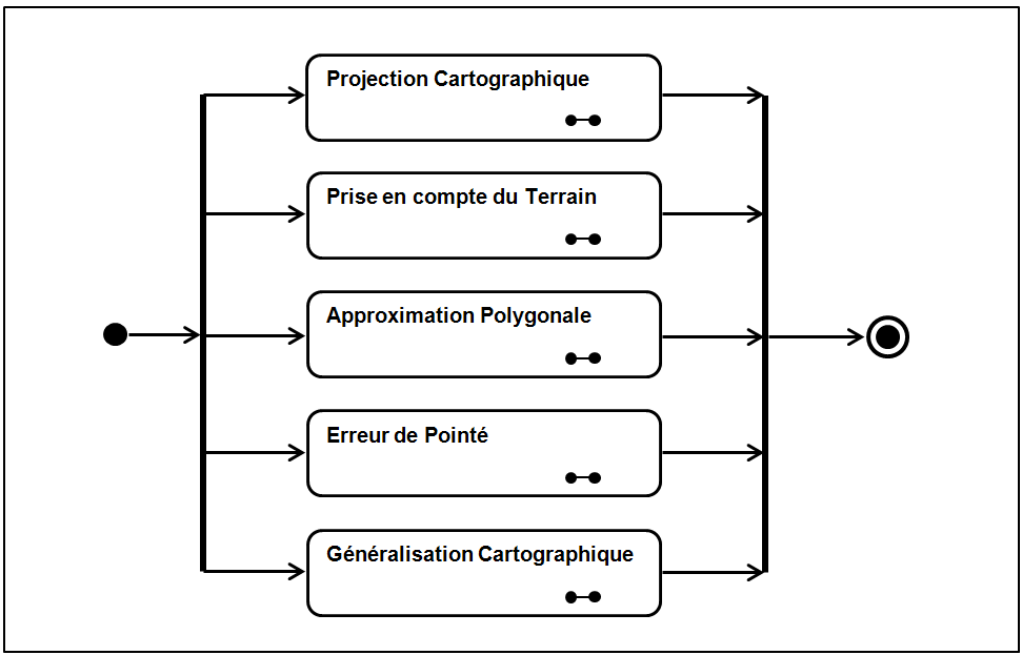


Figure 5.23 – Sous-diagramme d'activité correspondant à la sélection et au paramétrage des impacts à évaluer

Chaque impact sélectionné implique le paramétrage adapté du modèle sous-jacent, afin d'assurer le bon fonctionnement du processus d'estimation. De ce fait, la sélection et le paramétrage des impacts à estimer constituent la clé de voute du fonctionnement général du modèle. Nous allons décrire dans cette section chacune des étapes permettant de sélectionner et paramétrer correctement les modèles d'impact des règles de représentation et des processus de production pris en charge par le modèle.

Règles de représentation

Projection cartographique L'estimation de l'impact de la projection cartographique sur les mesures géométriques implique de disposer d'une grille d'altération linéaire (cf. figure 5.24). Ainsi, si l'utilisateur souhaite estimer cet impact, il doit préalablement charger cette grille. S'il n'en dispose pas, le processus d'estimation n'est pas activé et l'impact est considéré comme "null".

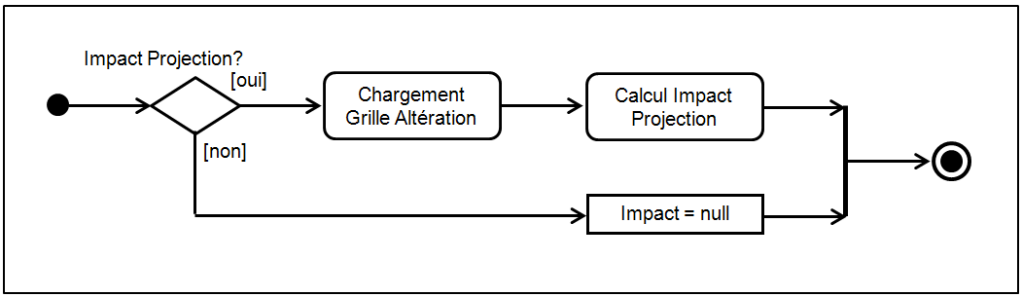


Figure 5.24 – Sous-diagramme d'activité relatif au paramétrage de l'impact de la projection cartographique

Prise en compte du terrain L'estimation de l'impact de la non-prise en compte de terrain implique de disposer d'un modèle numérique de terrain (cf. figure 5.25). Ainsi, l'estimation est conditionnée par le chargement de ce MNT. Si ce n'est pas le cas, le processus d'estimation n'est pas activé et l'impact est considéré comme "null".

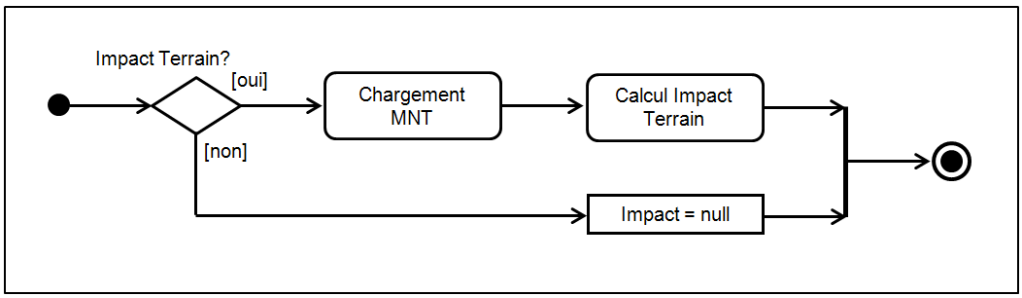


Figure 5.25 – Sous-diagramme d'activité relatif au paramétrage de l'impact de la non-prise en compte du terrain

Approximation polygonale des courbes Pour estimer l'impact de l'approximation polygonale des courbes, deux cas de figure sont envisagés. Soit l'estimation est réalisée sur l'ensemble de la collection d'objets sélectionnée, supposant donc que l'ensemble des objets représentent des entités courbes du monde réel. Soit une sélection des objets sinueux est proposée (cf. section 3.3.2.3), cette option n'étant applicable qu'à l'étude du réseau routier.

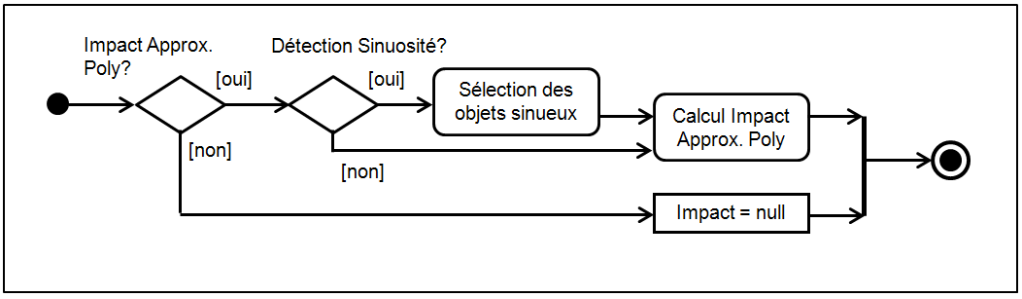


Figure 5.26 – Sous-diagramme d'activité relatif au paramétrage de l'impact de l'approximation polygonale des courbes

Processus de production

Erreur de pointé L'estimation de l'impact de l'erreur de pointé implique la connaissance de l'échelle moyenne de saisie des objets de la base de données évaluée. Ainsi, deux cas de figure ont été envisagés : soit l'utilisateur connait cette échelle et peut donc la paramétrer, soit il ne la connait pas, et dans ce cas la méthode d'estimation de l'échelle moyenne de saisie (proposées en section 4.1.2.2) est utilisée afin de l'informer.

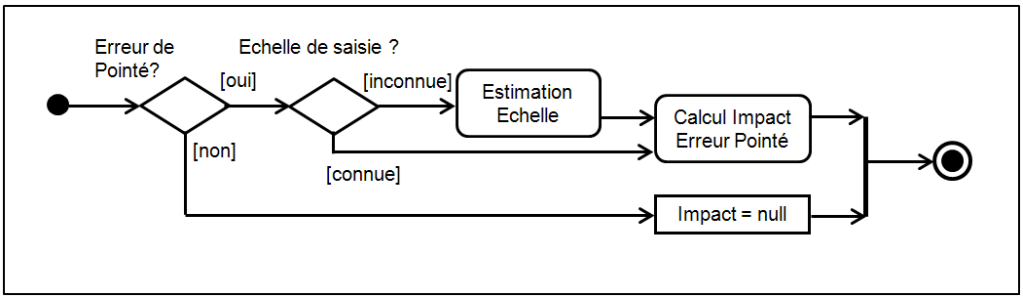


Figure 5.27 – Sous-diagramme d'activité relatif au paramétrage de l'impact de l'erreur de pointé

Généralisation cartographique L'estimation de l'impact de la généralisation cartographique suppose le paramétrage le plus conséquent de l'ensemble des impacts pris en charge par le modèle d'évaluation. Deux informations principales sont nécessaires au processus d'acquisition des connaissances :
– l'échelle de représentation des données sources,
– la délimitation du contexte géographique.
Plusieurs cas de figure ont été envisagés pour paramétrer le modèle (cf. figure 5.28).

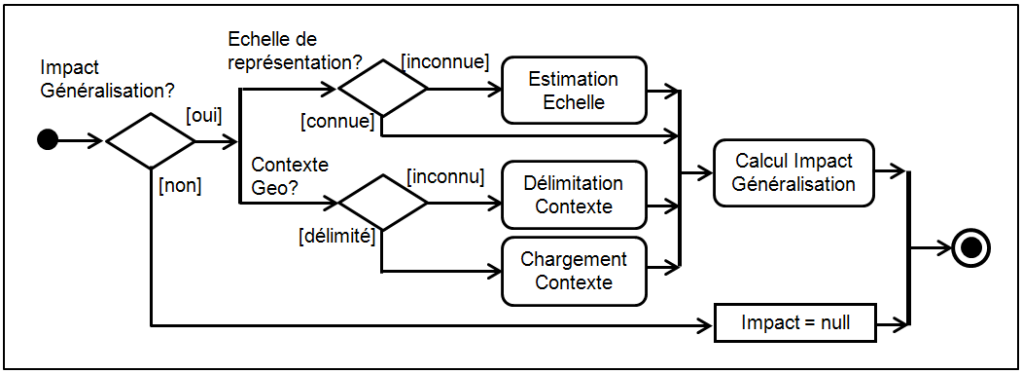


Figure 5.28 – Sous-diagramme d'activité relatif au paramétrage de l'impact de la généralisation cartographique

Concernant l'échelle de représentation des données sources, soit l'utilisateur la connait et est donc capable de paramétrer le modèle, soit il ne la connait pas, et dans ce cas les méthodes d'estimation de l'échelle de représentation (cf. section 4.2.2.1) sont activées. Ces méthodes ne sont proposées que sur l'étude du réseau routier, et impliquent le paramétrage d'une symbologie type.

Pour délimiter le contexte géographique, il est nécessaire de charger un fichier de délimitation des zones urbaines ou de montagne. Ainsi, l'utilisateur a la possibilité de charger un fichier de délimitation existant ou d'en créer un lui-même, à l'aide d'outils classiques de saisie de données. S'il ne peut satisfaire ces deux premiers cas de figure, nous proposons d'exploiter les méthodes de délimitation du contexte géographique exposées en section 4.2.2.2, afin de produire les fichiers de délimitation nécessaires à l'estimation de l'impact de la généralisation cartographique.

Une fois les différents impacts paramétrés et sélectionnés, une estimation de chacun d'entre eux est réalisée. La dernière étape du processus d'estimation repose sur la combinaison des impacts sélectionnés.

5.3.1.3 Combinaison des impacts

Dans le cadre du fonctionnement général du modèle d'estimation, nous envisageons le processus de combinaison des impacts de manière optionnelle. Ainsi, à la suite de l'estimation des impacts calculés de manière indépendante, la possibilité d'opérer à une agrégation des impacts est proposée à l'utilisateur (cf. figure 5.29).

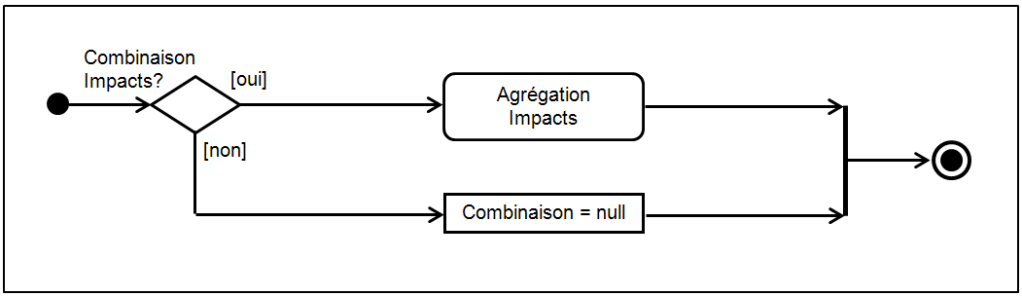


Figure 5.29 – Sous-diagramme d'activité relatif à la combinaison des impacts

A ce stade d'avancement, seule la stratégie par agrégation des impacts a été mise en œuvre dans le cadre de l'élaboration du modèle. La possibilité de choisir la méthode de combinaison souhaitée a malgré tout été envisagée dans le cadre de l'enrichissement du modèle d'évaluation.

Ainsi, après avoir présenté le fonctionnement général attendu du modèle EstIM, nous présenterons dans la section suivante les interfaces homme-machine développées afin de faciliter son utilisation.

5.3.2 Interface homme-machine

Cette section a pour objectif de présenter les interfaces homme-machine développées afin de permettre une utilisation conviviale et intuitive du modèle d'estimation de l'imprécision des mesures géométriques. Les interfaces présentées dans cette section ne constituent pour l'instant qu'un prototype du modèle EstIM.
Il avait déjà été exigé précédemment que le prototype développé dans le cadre de ce travail de thèse puisse s'intégrer sous forme de *plugin* au sein d'un environnement SIG existant. Le logiciel SIG OpenJUMP a pour cela été utilisé. Nous présenterons donc dans un premier temps l'environnement SIG utilisé pour développer le prototype, puis l'ensemble des interfaces développées pour permettre la mise en œuvre d'un processus d'estimation.

5.3.2.1 Le logiciel SIG OpenJUMP

Présentation du logiciel OpenJUMP Le logiciel OpenJump (initialement JUMP, pour "JAVA Unified Mapping Platform") est une application SIG modulaire, dont le développement a débuté à partir de 2002 par la société Vivid Solution, suite à un appel d'offre du ministère des ressources naturelles de la province de Colombie-Britannique (Canada). En 2007, suite à l'arrêt des financements, la licence est déposée en *OpenSource*, permettant la poursuite du développement du logiciel grâce aux efforts d'une communauté internationale (OpenJUMP, 2008).

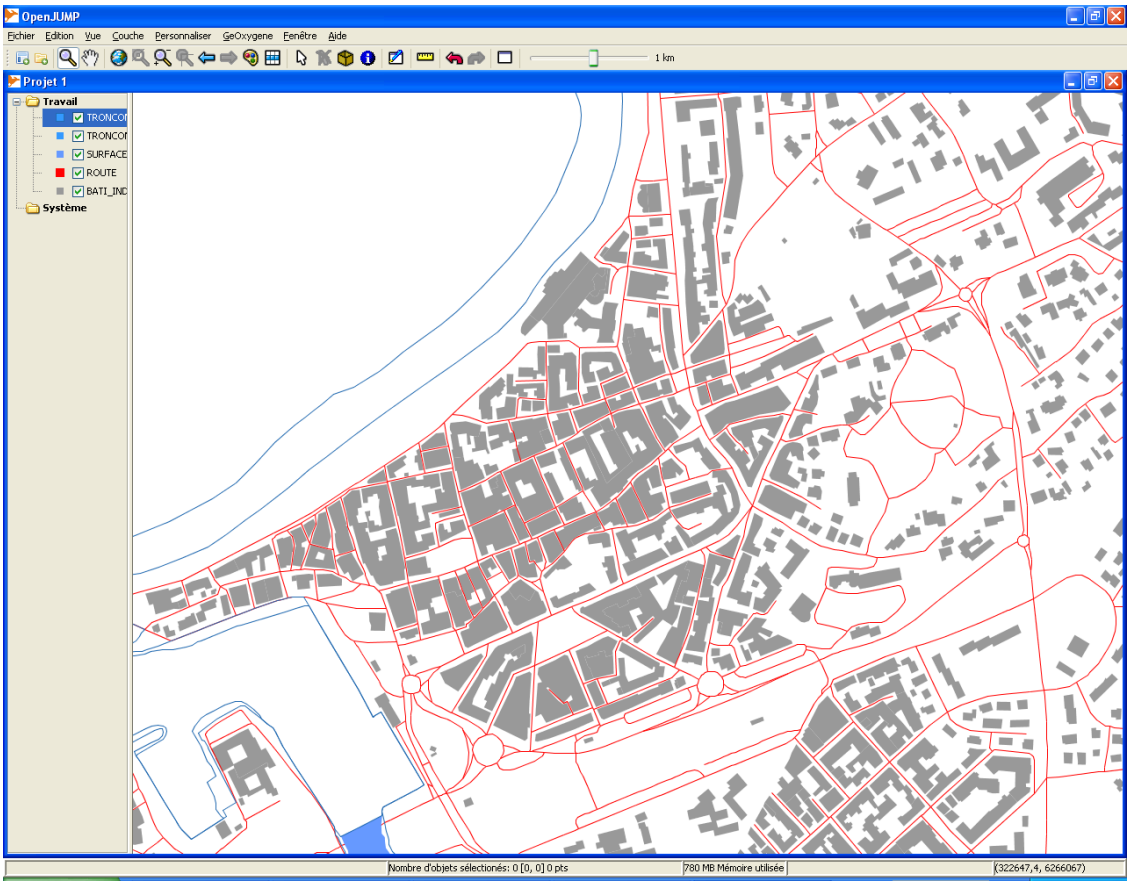


Figure 5.30 – Interface cartographique d'OpenJUMP

Intérêt du logiciel OpenJUMP Le logiciel OpenJUMP offre tout d'abord l'intérêt d'être un logiciel OpenSource, dont le code est ouvert et maintenu à jour par une communauté internationale. De plus, le logiciel OpenJUMP est développé en environnement JAVA, ce qui rend son utilisation facilement compatible avec la bibliothèque GeOxygene, exploitée dans le cadre de ce travail de thèse. Enfin, le logiciel OpenJUMP permet d'enrichir aisément son interface de *plugins* additionnels, permettant ainsi d'y intégrer le prototype du modèle d'estimation de l'imprécision des mesures géométriques.

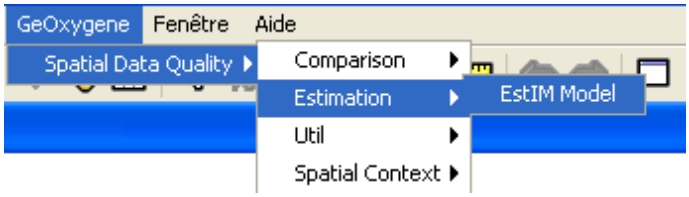


Figure 5.31 – Menu permettant d'accéder au modèle EstIM

5.3.2.2 Les interfaces du modèle EstIM

L'interfaçage du prototype EstIM s'organise autour d'une interface générale, permettant de sélectionner et de paramétrer les différents impacts sélectionnés. Nous présenterons tout d'abord cette interface générale, puis les interfaces d'aide au paramétrage du modèle et celles relatives à la visualisation des résultats.

Interface générale d'évaluation

L'interface générale permet la sélection et le paramétrage des impacts des règles de représentation et des processus de production proposés dans le cadre de ce travail de thèse. Cette interface est divisée en plusieurs sections (cf. figure 5.32) :
- une section spécifiant la base de données utilisée et la mesure géométrique étudiée (longueur ou surface),
- cinq sections correspondant à la sélection et au paramétrage de chaque impact étudié,
- une section relative à l'activation, ou non, de la combinaison des impacts et au lancement du processus d'estimation.

La sélection et le paramétrage de chacun des impacts sont donc réalisés de manière indépendante, l'utilisateur n'ayant qu'à sélectionner l'impact qu'il souhaite estimer, et le cas échéant, à le paramétrer. L'interface générale permet de saisir directement les informations nécessaires à chaque impact, c'est à dire :
- Projection cartographique : chemin vers la grille d'altération (au format *shapefile*),
- Prise en compte du terrain : chemin vers le modèle numérique de terrain (au format *ASCII Grid*),
- Approximation polygonale : détection, ou non, des objets sinueux,
- Erreur de pointé : échelle moyenne de saisie,
- Généralisation cartographique : échelle de représentation et chemins vers les fichiers de délimitation du contexte géographique (au format *shapefile*).

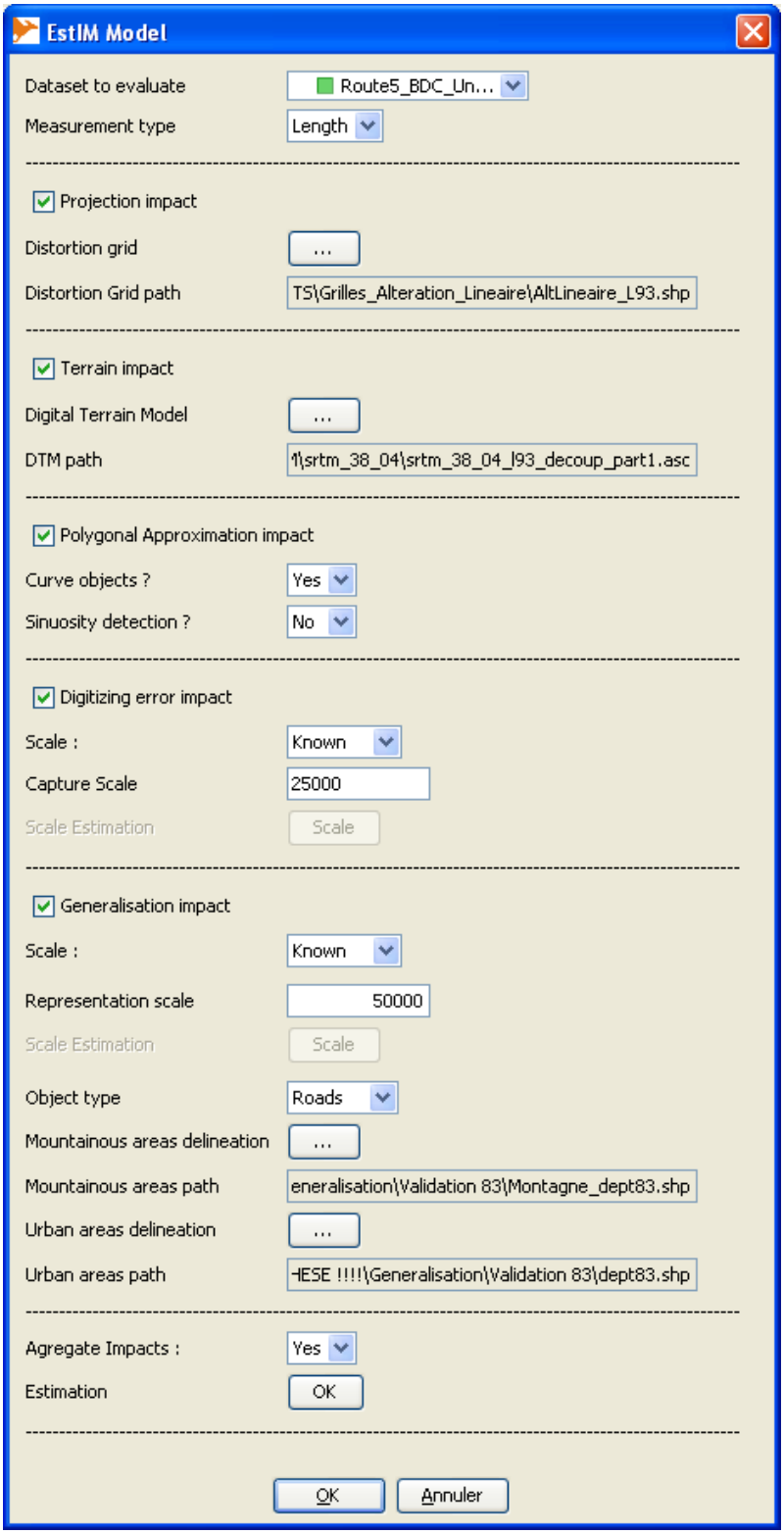


Figure 5.32 – Interface générale du modèle EstIM

Pour ces deux derniers impacts, des interfaces d'aide au paramétrage du modèle ont été prévues, afin d'accompagner l'utilisateur dans l'estimation des échelles de saisie et de représentation, ou dans la délimitation du contexte géographique des objets.

Enfin l'interface générale permet le lancement du processus d'estimation. Celui-ci est lancé à condition qu'une collection d'objets soit préalablement sélectionnée, via un traitement antérieur ou une sélection manuelle par l'utilisateur.

Interfaces d'aide au paramétrage du modèle

Pour faciliter l'estimation des impacts des processus de production, des interfaces d'aide au paramétrage du modèle sont accessibles depuis l'interface générale.

Erreur de pointé Seule la méthode d'estimation de l'échelle moyenne de saisie (proposée en section 4.1.2.2) est accessible depuis la section relative à l'impact de l'erreur de pointé. Si l'utilisateur sélectionne l'option "inconnue" dans le menu déroulant correspondant, un bouton est activé, permettant de l'informer de l'échelle de saisie estimée (cf. figure 5.33). L'utilisateur demeure libre par la suite de saisir l'échelle proposée ou de l'approximer.

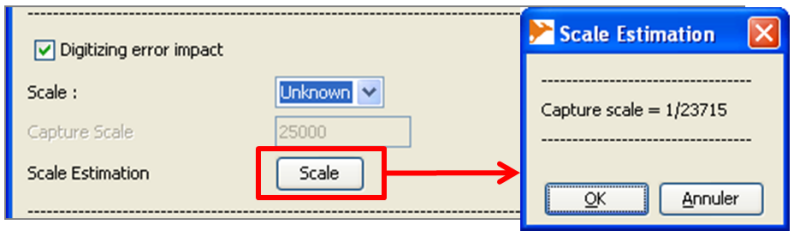


Figure 5.33 – Interface d'estimation de l'échelle moyenne de saisie

Généralisation cartographique Plusieurs méthodes d'aide au paramétrage sont accessibles depuis la section relative à l'impact de la généralisation cartographique, concernant l'estimation de l'échelle de représentation des données sources et la délimitation du contexte géographique.

La méthode d'estimation de l'échelle de représentation n'est proposée ici que sur les données routières (cf. section 4.2.2.1). Celle-ci est accessible si l'utilisateur sélectionne l'option "inconnue" dans le menu déroulant correspondant (cf. figure 5.34). L'interface proposée permet de choisir la méthode d'estimation (par superposition des routes voisines ou par empâtements de virages). De la même manière l'utilisateur demeure libre par la suite de saisir l'échelle proposée ou de l'approximer.

A ce stade de développement, la symbologie utilisée pour représenter chaque catégorie de tronçons de route est codée en dur dans l'application. Cependant, nous envisageons à terme d'exploiter des fichiers externes de symbologie, en utilisant les formats de description standards, comme le schéma SLD (*Style Layer Descriptor*).

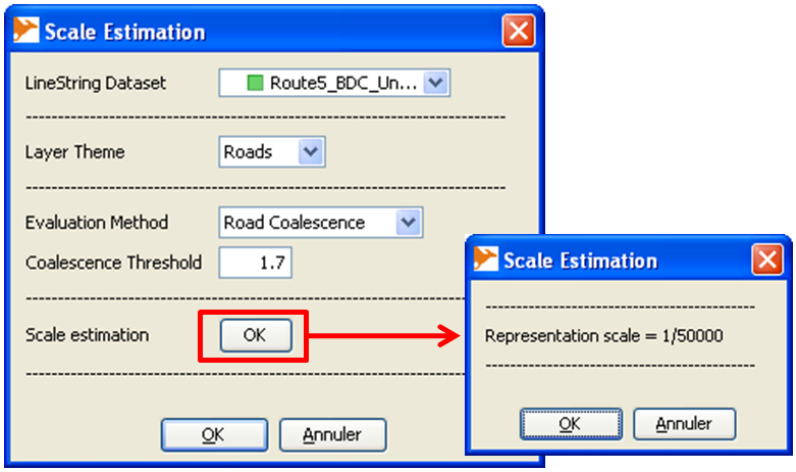


Figure 5.34 – Interface d'estimation de l'échelle de représentation

La section relative à l'impact de la généralisation cartographique permet également d'accéder aux interfaces permettant d'opérer à la délimitation du contexte géographique des objets, lorsque l'utilisateur n'en dispose pas (cf. figure 5.35).

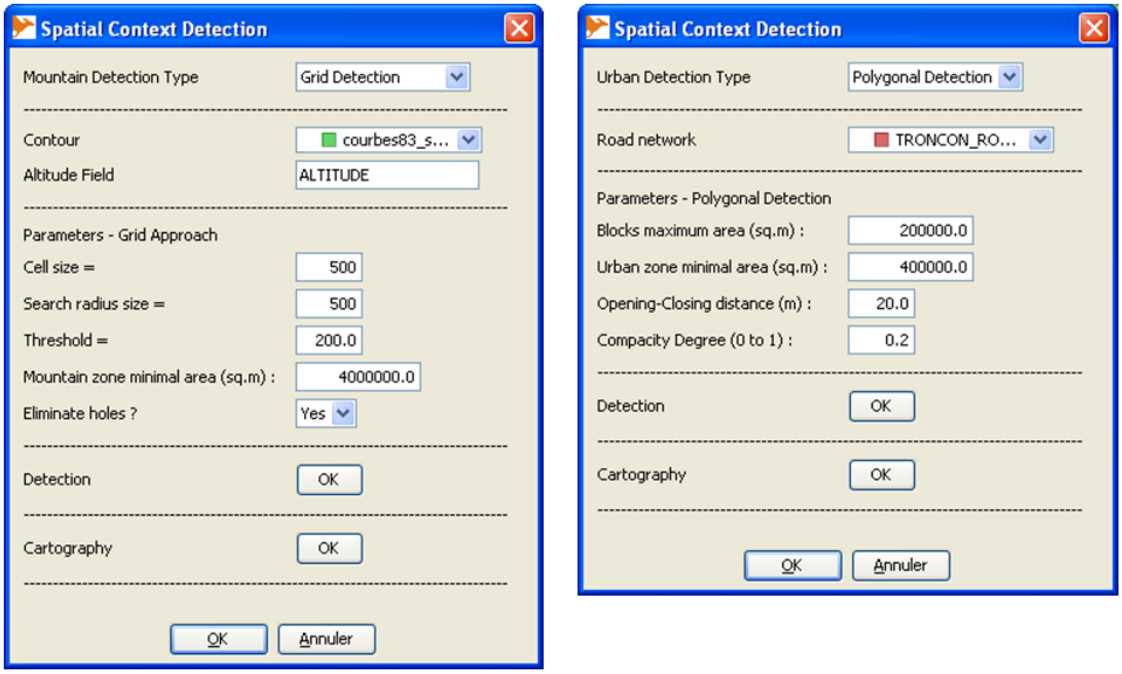


Figure 5.35 – Interfaces de délimitation des zones de montagne (à gauche) et des zones urbaines (à droite)

Ainsi, les méthodes proposées en section 4.2.2.2 ont été intégrées au sein du prototype, afin de délimiter les zones de montagnes (à partir d'un fichier de courbes de niveau) et les zones urbaines (à partir d'un réseau routier). Les résultats peuvent être sauvegardés (au format *shapefile*) pour ensuite être chargés dans le cadre du processus d'estimation.

Interfaces de communication des estimations

Enfin, des interfaces ont été prévues afin de communiquer les résultats d'un processus d'estimation à l'utilisateur. Deux interfaces de communication des estimations ont été prévues dans le cadre du développement du prototype :
– l'estimation des résultats par impact,
– l'estimation globale de l'imprécision de mesure.

L'interface de communication des estimations par impact permet de résumer, pour chaque impact sélectionné, l'estimation calculée par le modèle sous-jacent (cf. figure 5.36). L'information est uniquement formulée de manière textuelle pour l'instant. Les impacts qui n'ont pas été pris en compte sont mentionnés "non-sélectionnés".

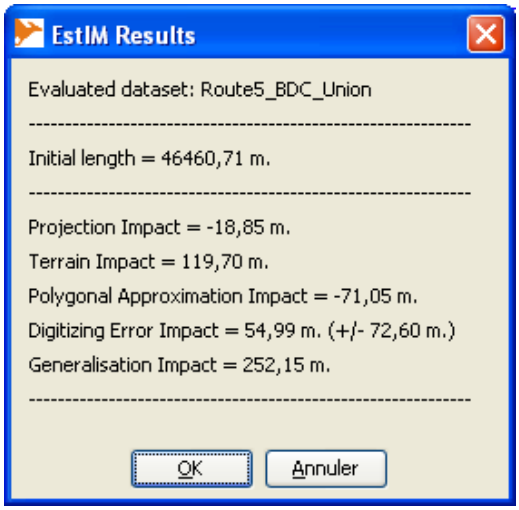


Figure 5.36 – Interface de communication des résultats par impact

Une seconde interface, proposant de communiquer une estimation agrégée des impacts, peut également être affichée à l'utilisateur, à condition que celui-ci ait sélectionné l'option "Agrégation des résultats" avant le lancement du processus d'estimation.

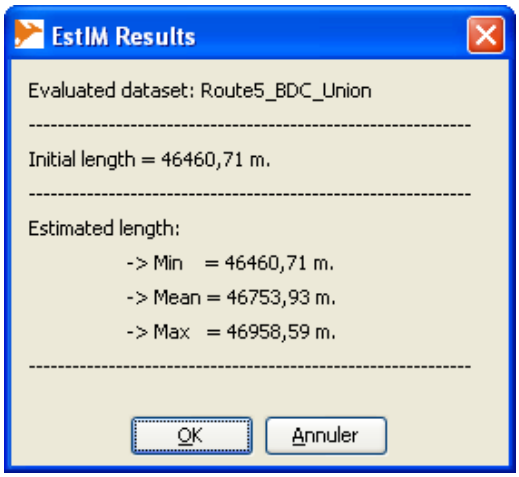


Figure 5.37 – Interface de communication de l'estimation globale

Tout comme l'interface de communication des estimations par impact, l'estimation globale est ici formulée de manière textuelle. D'autres formes de communication des résultats pourront être envisagées par la suite, sous forme de fonction de probabilité, ou de diagramme en radar par exemple.

Nous venons de présenter les interfaces homme-machine développées dans la version prototype du modèle EstIM. Bien évidemment, ces interfaces ne constituent pas un logiciel finalisé. Elle doivent être améliorées sur de nombreux points de vues (ergonomie, interactivité avec d'autres applications...).
Egalement, des tests d'utilisation de ces interfaces auraient pu être réalisées dans le cadre de ce travail de thèse. Par manque de temps, ceci ne pu être réalisé.
Malgré tout, cette phase de développement permet de matérialiser sous forme d'un outil concret et intégré dans un logiciel SIG, l'ensemble des principes exposés au cours de ce travail de thèse.

5.3.3 Mise en œuvre du modèle EstIM

Cette dernière section a pour objectif de mettre en œuvre le modèle EstIM en nous appuyant sur un cas d'utilisation courant. Pour cela, nous proposons de l'appliquer sur l'exemple d'un calcul d'itinéraire routier, qui constitue aujourd'hui l'une des applications les plus couramment employées. Nous présenterons tout d'abord l'expérimentation réalisée, et ensuite les résultats de l'estimation de l'imprécision de mesure.

5.3.3.1 Présentation de l'expérimentation

L'expérimentation proposée pour mettre en œuvre le modèle EstIM repose sur un calcul d'itinéraire calculé à partir de l'application *GoogleMaps*. L'itinéraire concerné est effectué entre les adresses suivantes :
– Départ : 40, place du Général De Gaulle, Saint-Lô
– Arrivée : 13, avenue de la Porte d'Italie, Paris

La tracé de l'itinéraire calculé par *GoogleMaps* est présenté en figure 5.38. La longueur totale communiquée suite au calcul d'itinéraire est de 312 kilomètres.

FIGURE 5.38 – Résultat du calcul d'itinéraire réalisé avec *GoogleMap*

Pré-traitements Afin d'estimer l'imprécision de la longueur calculée, l'itinéraire doit subir une série de pré-traitements. L'itinéraire est tout d'abord exporté au format KML, puis convertit au format Shapefile. Etant donné que l'itinéraire est initialement produit en coordonnées géographiques WGS84, il doit être ensuite re-projeté en projection Lambert93, afin de d'être évalué à l'aide du modèle EstIM.

La figure 5.39 présente ces différentes étapes de pré-traitements. Après projection, la mesure de longueur calculée totale de l'itinéraire est de 312,691 kilomètres.

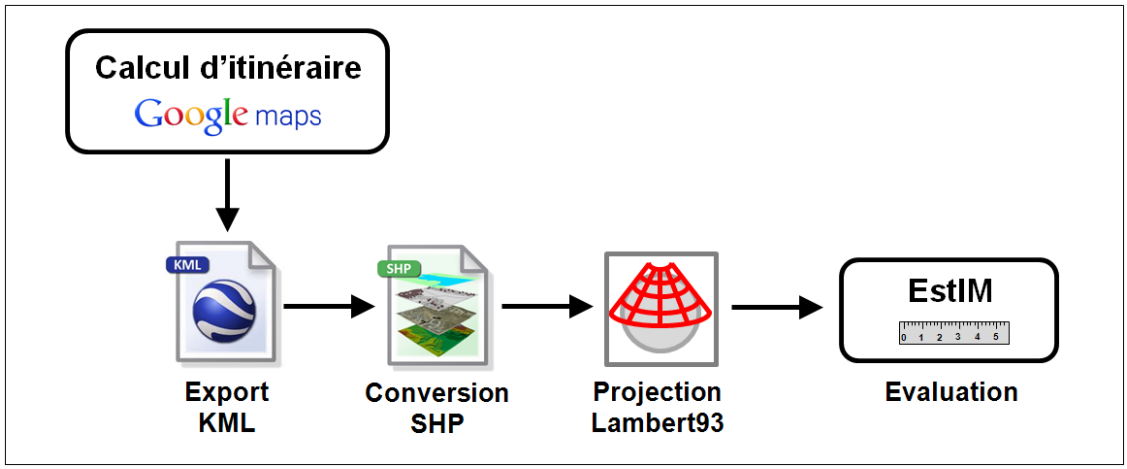


Figure 5.39 – Procédures d'export et de conversion de l'itinéraire calculé pour son évaluation avec le modèle EstIM

La section suivante va nous permettre de présenter les résultats de l'estimation de l'imprécision de la mesure de longueur calculée sur l'itinéraire expérimenté, par application du modèle EstIM.

5.3.3.2 Estimation de l'imprécision des mesures

Pour estimer l'imprécision de la mesure de longueur calculée sur l'itinéraire expérimenté, l'impact de la généralisation cartographique n'a pas été pris en compte. En effet, les données utilisées pour les calculs d'itinéraires au sein de l'application GoogleMaps, principalement les données réseau Multinet (TeleAtlas, 2009), produites par la société TeleAtlas, ne sont pas généralisées cartographiquement.

Méthodologie Seul l'impact de l'erreur de pointé sera donc pris en compte dans le cadre de l'évaluation de l'impact des processus de production. Suite à l'exploration des distances inter-points, l'imprécision de pointé a été fixée à 5,9 mètres, ce qui est relativement cohérent avec les spécifications du produit Multinet (TeleAtlas, 2009) et rejoint les observations de Helbich *et al.* (2011).

Concernant les règles de représentation, les impacts de l'approximation polygonale des courbes, de la projection cartographique, ainsi que de la non-prise en compte du terrain ont été estimés. Pour ce dernier impact, les altitudes extraites du MNE SRTM (Farr *et al.*, 2007) ont été utilisées.

Estimation des impacts La figure 5.40 présente les estimations calculées par le modèle EstIM pour les quatre impacts sélectionnés dans le cadre de l'évaluation. Nous observons que l'impact de l'approximation polygonale des courbes constitue la contribution la plus significative dans le cadre de cette évaluation. L'impact de l'erreur de pointé est relativement important, en particulier du fait du biais engendré par le modèle de simulation, tandis que l'impact du terrain est relativement secondaire et celui de la projection cartographique peut être considéré comme négligeable.

Estimation globale Etant donné que les données utilisées pour le calcul d'itinéraire sont faiblement généralisées, les résultats des agrégations simple (pour la longueur estimée maximale) et quadratique (pour la longueur estimée moyenne) sont exploités afin de produire une estimation globale de l'imprécision des mesures. Les résultats montrent que la longueur estimée moyenne de l'itinéraire est de 313,37 km. (soit une erreur de 0,22%) tandis que la longueur estimée maximale est de 313,77 km. (soit une erreur de 0,35%).

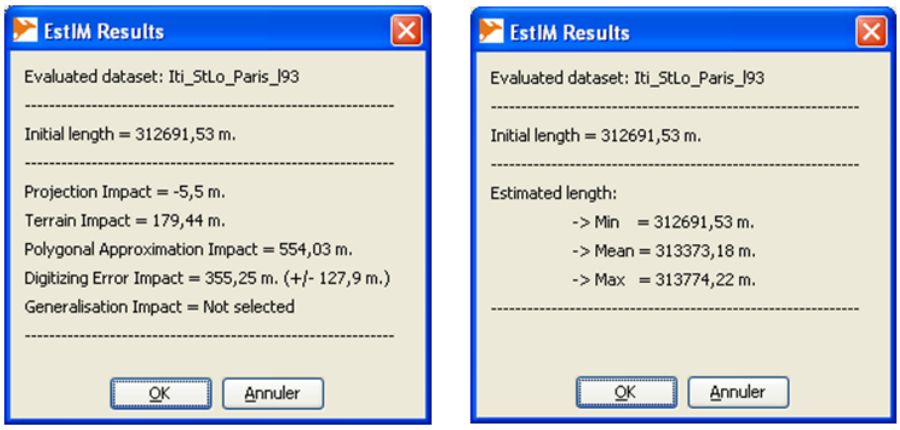


Figure 5.40 – Résultats de l'évaluation de l'itinéraire par le modèle EstIM

Cette estimation demeure insuffisante pour encadrer la distance réelle, puisqu'en réalisant le même itinéraire en voiture, la distance indiquée au compteur est de 316,3 kilomètres. Ces résultats démontrent cependant l'intérêt d'une telle approche, étant donné que l'estimation produite permet malgré tout de réduire l'erreur de longueur.

Nous venons de présenter un cas de mise en œuvre du prototype du modèle EstIM en aval d'un calcul d'itinéraire réalisé sur *GoogleMaps*. Ce type d'application constitue aujourd'hui un cas d'utilisation de l'information géographique des plus courants. L'exploitation du modèle d'estimation permet d'enrichir l'information proposée, en communiquant une imprécision de mesure, qui peut s'avérer potentiellement pertinente pour l'utilisateur. Nous voyons bien qu'un tel modèle d'estimation peut être intégré directement en aval de nombreuses applications nécessitant des mesures géométriques. Une identification des cas d'utilisations pouvant potentiellement solliciter une estimation de l'imprécision des mesures devra donc être entreprise par la suite, afin que le modèle proposé puisse s'adapter d'avantage aux besoins des utilisateurs.

Synthèse Section Cette section a permis de présenter la phase de développement du prototype EstIM. Les principes de fonctionnement du prototype ont été exposés dans un premier temps. Ceux-ci doivent satisfaire des exigences en termes de facilité d'utilisation. Ainsi, les options de sélection et de paramétrage des impacts ont été fortement simplifiées, et des méthodes d'accompagnement de l'utilisateur ont été prévues pour faciliter ce paramétrage. Dans un second temps, les interfaces homme-machine du prototype ont été présentées. Celles-ci ont également été développées afin d'être les plus intuitives possible pour l'utilisateur, en réponse aux exigences de fonctionnement du modèle. Enfin, une mise en œuvre du prototype EstIM a été proposée, en aval d'un calcul d'itinéraire.

Conclusion du chapitre

Ce dernier chapitre avait pour objectif de présenter la conception du modèle général d'estimation de l'imprécision des mesures géométriques, baptisé modèle EstIM.
L'objectif de ce modèle repose sur une estimation de chacun des impacts des règles de représentation et des processus de production, tel que cela a été proposé dans les chapitres 3 et 4, mais également sur une estimation globale de l'imprécision des mesures géométriques calculées.

La production d'une estimation globale implique une réflexion sur la combinaison des impacts, modélisés jusqu'à présent de manière indépendante. Or, nous savons que dans la réalité, ces impacts s'accumulent et inter-agissent de manière beaucoup plus complexe. Bien qu'imparfaites, deux méthodes de combinaison des impacts ont été expérimentées : l'agrégation des impacts, et la sélection de l'impact le plus significatif.
Les résultats des expérimentations ont montré que pour les données faiblement généralisées, la méthode par agrégation quadratique était la plus performante, alors que pour les données très généralisées, il s'agissait de la méthode par agrégation simple. Cependant nous avons pu voir que ces deux méthodes pouvaient s'avérer complémentaires, afin de produire un intervalle de confiance sur la mesure estimée.
Egalement, d'autres méthodes de combinaison des impacts ont été évoquées en perspectives, basées sur des approches par apprentissage automatique supervisé, ou par composition de modèles. Cependant ces méthodes n'ont pas été intégrées au sein du modèle général d'estimation.

Les principes de modélisation du modèle EstIM ont ensuite été présentés. Le modèle est bâti sur une architecture modulaire, permettant d'enrichir le modèle d'impacts supplémentaires par la suite.
Le fonctionnement général du modèle EstIM et le premier prototype développé ont finalement été présentés, en terminant par une mise en application dans le cadre d'un cas d'utilisation classique, celui d'un calcul d'itinéraire routier.

Conclusion Générale

Rappel des objectifs

Ce travail de thèse nous a permis de proposer un modèle général d'estimation de l'imprécision des mesures de longueur et de surface calculées à partir de la géométrie des objets d'une base de données géographiques vectorielles. L'objectif initial de cette thèse était de permettre de répondre à la question d'un utilisateur de données géographiques : "Si la mesure calculée avec mon logiciel SIG est de x mètres, de combien est l'erreur probable de mesure ?" Pour répondre à cette question, nous avons proposé un modèle permettant d'estimer les impacts de règles de représentation et de processus de production qui affectent la géométrie des objets géographiques vectoriels.

Résumé du mémoire

Cette thèse a tout d'abord permis de replacer la problématique de l'estimation de l'imprécision des mesures géométriques au sein d'un cadre thématique plus général, celui de l'évaluation de la qualité des données géographiques.

Ainsi, dans le chapitre 1, nous avons rappelé les causes d'imperfection des données géographiques. Nous avons montré que ces imperfections étaient principalement dues à deux processus dont les impacts s'accumulent : le processus d'abstraction du monde réel, et le processus de production des objets géographiques. Ce chapitre a également permis de réaliser un bilan des moyens actuels d'estimation de l'imprécision des mesures géométriques, montrant que les mesures et modèles développés ces dernières années s'attachent davantage à quantifier l'erreur de positionnement des objets, plutôt que l'erreur des mesures portées par la géométrie des objets.

Dans ce contexte, le chapitre 2 a permis de fixer les objectifs de ce travail de thèse. Suite à une analyse des besoins actuels en termes d'information sur la qualité géométrique des données géographiques, nous avons montré que les moyens actuels ne permettent pas d'estimer et de communiquer sur les erreurs de mesure géométrique. Ainsi un modèle d'estimation, dédié aux utilisateurs de données géographiques, a été proposé. Le modèle proposé doit permettre, sans données de référence, d'estimer l'imprécision des mesures géométriques de longueur et de surface, par connaissance des impacts de règles de représentation et de processus de production des données évaluées.

Dans le chapitre 3, nous avons proposé des méthodes d'estimation des impacts de règles de représentation. Trois règles de représentation étaient ici étudiées : la projection cartographique, la non-prise en compte du terrain, et l'approximation polygonale des courbes.

Pour estimer l'impact de la projection cartographique sur les mesures géométriques, des mesures sur ellipsoïde sont proposées. Ces mesures sont effectuées par enrichissement de la géométrie des objets d'une information d'altération linéaire. Pour évaluer l'impact de la non-prise en compte du terrain, un enrichissement de la géométrie par une information altimétrique est proposée, en exploitant des modèles numériques de terrain existants. Enfin, pour estimer l'impact de l'approximation polygonale des courbes, la reconstruction de géométries courbes par le biais de méthodes d'interpolation polynomiale est proposée.

Le chapitre 4 avait pour objectif de proposer des méthodes d'estimation des impacts de processus de production. Deux impacts ont été étudiés : l'erreur de pointé et la généralisation cartographique. Pour estimer l'impact de l'erreur de pointé, une approche par simulations est proposée. Cette approche repose sur la génération d'objets bruités, permettant de reproduire l'erreur de pointé et son impact sur les mesures géométriques. Des méthodes de paramétrage du bruit (par estimation de l'échelle moyenne de saisie) et de préservation du réalisme des objets bruités ont également été présentées. Pour estimer l'impact de la généralisation cartographique, une approche par tests est proposée. Cette approche repose sur des comparaisons entre objets homologues extraits de bases de données généralisés et d'une base de référence, permettant, par apprentissage, d'estimer un modèle statistique des erreurs de mesure. Pour cela, des méthodes de qualification des objets généralisées ont été développées, permettant de déterminer leur échelle de représentation ou leur contexte géographique.

L'objectif du chapitre 5 résidait dans la conception du modèle d'estimation de l'imprécision des mesures géométriques. Afin de produire une estimation globale de l'imprécision des mesures géométriques calculées, des méthodes de combinaison des impacts des règles de représentation et des processus de production ont été proposées. Nous avons montré que l'utilisation de méthodes simples permet déjà de réduire l'erreur résiduelle de mesure, ce qui constitue une information intéressante pour un utilisateur. Cependant, du fait de la complexité du processus d'accumulation des erreurs, d'autres méthodes de combinaison ont été évoquées en perspectives. Les principes de modélisation et de fonctionnement général du modèle ont ensuite été présentés et un prototype du modèle EstIM (Estimation de l'Imprécision des Mesures) a été développé.

Contributions

Un cadre conceptuel pour les erreurs de mesure

Le premier apport de ce travail réside dans la conceptualisation des composantes de l'erreur de mesure géométrique. Nous avons cherché à identifier les différentes sources d'erreurs impactant les mesures géométriques et les avons divisées en deux catégories : celles relatives aux règles de représentation des objets géographiques, et celles relatives aux processus de production de la géométrie des objets. Cette différenciation avait pour principal objectif d'éclaircir le champ des composantes de l'erreur finale de mesure, et en particulier la relation géométrie/représentation, qui, pour chaque combinaison proposée, implique une erreur de mesure différente.

Une première réflexion sur la combinaison des impacts

Contrairement aux modèles d'estimation proposés dans la littérature, qui cherchent davantage à estimer globalement les erreurs de mesure sans vraiment s'intéresser aux causes de ces erreurs, nous avons cherché dans ce travail à identifier les causes d'erreurs et à modéliser leur impact sur les mesures géométriques, la principale difficulté résidant dans leur combinaison pour produire une estimation globale. Bien que la modélisation de la combinaison des impacts demeure inaboutie, ce travail a malgré tout permis d'aborder la question complexe de l'accumulation des erreurs sur les mesures géométriques. Les résultats obtenus montrent notamment qu'en fonction des caractéristiques des objets évalués, la méthode de combinaison des impacts la plus performante n'est pas la même.

Un outil d'estimation de l'imprécision des mesures géométriques

Dans le cadre conceptuel ainsi défini, nous avons proposé un outil logiciel, permettant à un utilisateur de données géographiques d'estimer l'imprécision des mesures de longueur, par connaissance des règles de représentation et des processus de production impactant ces données. Le modèle proposé a été implémenté et mis en œuvre au sein du prototype EstIM, intégré au logiciel SIG Libre OpenJUMP.

Des méthodes d'enrichissement géométrique des objets

Le modèle développé a tout d'abord permis de mettre au point un certain nombre de méthodes d'enrichissement de la géométrie des objets vectoriels, afin d'estimer les impacts de règles de représentation. L'utilisation de données externes (MNT, grilles d'altération) pour enrichir les données évaluées a été privilégiée pour estimer les impacts de la projection cartographique et de la non-prise en compte du terrain. Des méthodes d'interpolation polynomiale ont également été mobilisées pour estimer l'impact de l'approximation polygonale des courbes, étant donné que dans la majorité des logiciels SIG aujourd'hui, les objets courbes demeurent représentés sous forme de lignes polygonales.

Un modèle de simulation de l'erreur de pointé

Un modèle de simulation a été implémenté pour estimer l'erreur de pointé. Outre le choix de la méthode de paramétrage du bruit gaussien, basée sur l'estimation de l'échelle moyenne de saisie, notre modèle de simulation permet de préserver le réalisme des objets simulés, en intégrant des contraintes de respect de forme et de corrélation des erreurs entre points successifs.

Un modèle statistique d'impact de la généralisation cartographique sur les objets routiers

L'estimation de l'impact de la généralisation cartographique s'est basée sur une approche par tests, permettant, par le biais de comparaisons entre bases de données généralisées et une base de données de référence, d'extraire par apprentissage un modèle d'estimation des erreurs de mesure. Le modèle statistique ainsi développé n'a pu être validé que sur les objets routiers. Cette application a malgré tout permis de quantifier les impacts de l'échelle de représentation et du contexte géographique des objets routiers sur les mesures de longueur.

Des méthodes de qualification des données

Pour permettre l'estimation de chacun des impacts sélectionnés, des méthodes de qualification des données géographiques ont été développées et implémentées au sein du modèle d'estimation. Ces méthodes peuvent être classées en deux catégories : les méthodes basées sur des données externes, et les méthodes basées sur une analyse des données. Dans la première catégorie, nous pouvons citer les méthodes de délimitation du contexte géographique des objets nécessaires à l'estimation de l'impact de la généralisation cartographique. Dans la seconde catégorie, nous retiendrons notamment les méthodes d'estimation des échelles de saisie ou de représentation des données évaluées.

Enseignements

Dans le cadre de cette thèse, nous avons supposé que la connaissance des règles de représentation et des processus de production permettait d'estimer l'imprécision des mesures géométriques. Nous avons montré que même en utilisant des méthodes de combinaison simples des impacts, il était possible de réduire l'erreur résiduelle, ce qui confirme notre hypothèse de départ. Ainsi, l'élaboration d'un tel modèle d'estimation peut donc s'avérer utile pour un utilisateur, ce qui constituait notre objectif initial

Au cours de ce travail, nous avons également pu voir à quel point le processus d'évaluation requiert un processus d'acquisition de connaissances spécifiques. En ce sens, la qualification des objets évalués (par la connaissance de l'échelle de saisie ou de représentation, du contexte géographique) permet d'apporter des informations indispensables à l'estimation de l'imprécision des mesures portées par la géométrie des objets. Nous avons vu que cette information pouvait être déduite par le biais de méthodes automatiques guidées par des connaissances d'experts.

Bien que les métadonnées permettent principalement de communiquer des informations d'imprécision géométrique portant sur l'erreur de positionnement des objets, nous avons vu par le biais d'expérimentations que la connaissance de l'erreur moyenne quadratique peut être utile au paramétrage du modèle de simulation de l'erreur de pointé (dans le cas des données n'ayant pas subi l'impact de la généralisation cartographique). Ceci démontre que l'élaboration d'un tel modèle d'estimation peut s'avérer complémentaire des outils de communication classiques comme les métadonnées, tant en sortie (pour le paramétrage du modèle), qu'en entrée (pour remplir des champs relatifs à l'imprécision des mesures géométriques), comme illustré en figure 6.1.

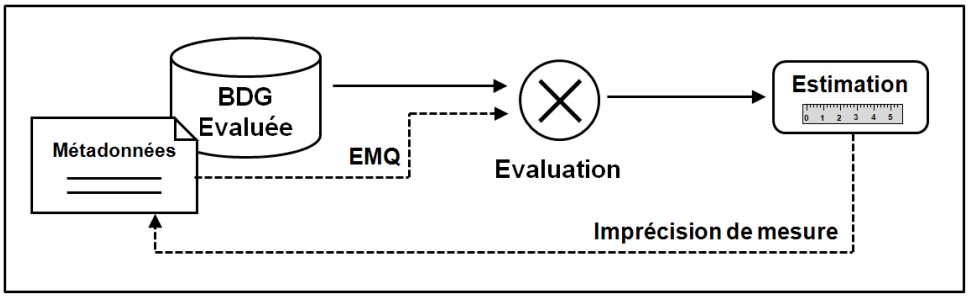


FIGURE 6.1 – Relations possibles entre le modèle d'estimation et les métadonnées

Perspectives

La mise en perspective de ce travail méthodologique ouvre de nombreuses pistes de recherche auxquelles le périmètre de cette thèse n'aurait jamais pu prétendre répondre. Nous essaierons d'ordonner ces perspectives par ordre de priorité.

Intégration de processus supplémentaires

L'architecture modulaire pensée dans le cadre de l'élaboration du modèle d'estimation suppose la possibilité de l'enrichir de modèles permettant d'estimer les impacts de règles de représentation ou de processus de production supplémentaires. La priorité réside dans l'intégration d'un modèle d'impact de l'imprécision des acquisitions GNSS, étant donné la popularité actuelle de ce moyen d'acquisition. Cependant la modélisation de son impact sur les mesures géométriques demeure complexe, du fait des nombreux processus générateurs d'erreurs impactant les signaux reçus (satellite, propagation atmosphérique, environnement ou encore qualité du récepteur).

Croisement de processus de production hétérogènes

Si nous avons souligné la complexité de croiser les impacts affectant les mesures géométriques, il est nécessaire de garder à l'esprit que des processus de production hétérogènes cohabitent bien souvent au sein d'une même base de données. Pourtant nous avons cherché pour l'instant à les croiser de manière indépendante, considérant que tous les objets sont affectés des mêmes processus de production. Pourtant nous savons bien que ce n'est pas le cas dans la réalité. Par exemple, les objets de la BDTOPO peuvent être produits par le biais de restitution photogrammétrique et de relevés GNSS de précision. Parfois même les deux modes de production cohabitent au sein d'un même objet (saisi par relevé GNSS, puis modifié par digitalisation manuelle). Les questions relatives à la combinaison de ces processus doivent donc être entreprises, et prennent même une résonance particulièrement forte dans le cas des données géographiques collaboratives (VGI) où des processus et des sources de données hétérogènes cohabitent bien souvent au sein d'une même classe d'objets (Girres et Touya, 2010).

Enrichissement du modèle d'impact de la généralisation cartographique

Le modèle statistique d'estimation de l'impact de la généralisation cartographique n'a pour l'instant été développé et validé que sur l'étude du réseau routier. Concernant ce modèle à proprement parler, de nombreuses améliorations peuvent être apportées. Tout d'abord, l'estimation de l'imprécision des mesures de longueur en fonction de la longueur étudiée n'a pu être formalisée. Bien que nous sachions que l'imprécision relative de mesure réduit lorsque la taille du tronçon augmente, sa quantification demeure nécessaire. Egalement, le modèle proposé a été réalisé sans prendre en compte la classification des routes étudiées. Une étude doit donc être envisagée en ce sens, afin de voir si l'erreur de longueur et son imprécision varient en fonction de la classification de la route étudiée.

Ce modèle statistique doit cependant être enrichi à d'autres classes d'objets (réseau hydrographique, voies ferrées) et aux mesures de surface, qui comme nous l'avons vu, ne sont pas affectées de la même manière en fonction des caractéristiques des objets (s'il forment ou non une partition de l'espace) et de leur taille.

Egalement, le modèle d'estimation de l'impact de la généralisation cartographique pourrait être porté à l'étude des cartes anciennes, qui ont elles aussi subi des processus de généralisation cartographique, se manifestant par des modifications de position et de forme des objets différentes à celles étudiées dans le cadre de travail.
Enfin, ces travaux montrent aussi l'intérêt, dans le cadre de processus automatique de généralisation cartographique, qui sont aujourd'hui majoritairement employés en production, de conserver les liens avec les objets de référence, afin notamment de conserver la mesure initiale portée par les objets, et ainsi quantifier véritablement l'impact de chaque processus de généralisation cartographique mobilisé.

Amélioration de la qualification des données

Pour permettre une estimation de l'imprécision des mesures, la qualification des données évaluées, par le biais de méthodes d'acquisition des connaissances spécifiques, se révèle être un préalable indispensable. Les méthodes développées dans le cadre de ce travail de thèse ont été envisagées afin de répondre à des hypothèses de départ (comme l'influence du contexte géographique sur les erreurs de longueur des routes). Cependant, nous considérons que le panel d'indicateurs de qualification des données doit être enrichi afin d'affiner le processus d'évaluation. Par exemple, l'évaluation de l'impact de la généralisation cartographique pourrait être enrichi d'indicateurs supplémentaires, comme la sinuosité des routes, ou encore leur dénivelé, afin d'estimer plus précisément l'imprécision des mesures en fonction de la configuration spatiale des objets. Nous pourrions ainsi envisager à terme un processus d'évaluation entièrement guidé par les connaissances associées aux objets, permettant d'estimer l'imprécision des mesures géométriques en tenant compte de leurs caractéristiques, comme proposé en section 5.1.3.

Intégration de mesures supplémentaires

Avec l'avènement des données représentées en trois dimensions, et la naissance de logiciels SIG 3D, la quantification de l'imprécision des mesures géométriques à trois dimensions va également se poser. Des modèles spécifiques doivent donc être envisagés afin d'estimer ces imprécisions, d'autant plus que les objets concernés sont généralement acquis avec des processus de production qui n'ont pas été abordés dans le cadre de ce travail de thèse. Cependant, les ordres de grandeur des erreurs impliquées par les objets concernés ne sont pas les mêmes, puisqu'ils présentent généralement de petites dimensions, comme les bâtiments par exemple.

Ces perspectives montrent l'importance de prolonger les études et recherches sur l'évaluation de l'impact de la qualité des données géographiques. Ceci concerne à la fois l'estimation des erreurs de mesure géométrique, mais également la propagation de ces erreurs sur des applications concrètes sollicitant l'information géographique (calculs d'itinéraires, d'écoulement...). Ces conclusions soulignent également la nécessité de disposer de données de référence, nécessaires à la calibration des différents modèles, et de corpus de données indispensables à la qualification des données évaluées.

Annexes

Annexe A. Descriptif des mesures d'exactitude géométrique de la spécification technique ISO 19138

Annexe B. Descriptif des mesures proposées dans l'Arrêté du 16 septembre 2003

Annexe C. Principes généraux de modélisation orientée-objet

Annexe D. Présentation et résultats du test de digitalisation manuelle

Annexe E. Résultats des comparaisons de longueurs sur les routes généralisées

Annexe A. Descriptif des mesures d'exactitude géométrique de la spécification technique ISO 19138

La spécification technique ISO 19138 (ISOTS19138, 2006) propose un ensemble de mesures permettant d'évaluer la qualité des bases de données géographiques. Ces mesures sont subdivisées par éléments, puis par sous-éléments. Nous présentons ici les mesures liées à l'élément "Exactitude de Position" (*Positionnal Accuracy*). Les mesures d'exactitude de position sont subdivisées en mesures d'exactitude absolue (ou externe) et en mesures d'exactitude relative (ou interne).

A.1 Exactitude absolue

Les mesures d'exactitude absolue (ou externe) sont divisées en trois catégories :
– les mesures générales d'incertitude de positionnement,
– les indicateurs d'incertitude de positionnement vertical,
– les indicateurs d'incertitude de positionnement horizontal.

Mesures générales Les mesures générales proposent notamment le calcul de la valeur moyenne d'incertitude de positionnement (*Mean value of positional uncertainties*) entre le jeu de données évalué et le jeu de données de référence. Cette valeur moyenne est basée sur le calcul de la distance euclidienne entre les points homologues des deux jeux de données comparés. Cette mesure est également calculée sur des comparaisons ne prenant pas en compte les valeurs aberrantes, c'est à dire les erreurs de position trop importantes entre points homologues (*Mean value of positional uncertainties excluding outliers*). Une valeur seuil est déterminée à l'avance afin d'éliminer ces erreurs de position considérées aberrantes. A partir de cette valeur seuil, sont également calculés le nombre et le taux d'erreurs de positions dépassant une erreur tolérée. Enfin, une matrice de Variance-Covariance des coordonnées est également proposée dans cette catégorie des mesures générales. La matrice de Variance-Covariance est en effet calculée afin de compenser les erreurs accumulées lors de relevés topographiques.

Incertitude de positionnement vertical Les mesures d'incertitude de positionnement vertical sont en grande partie issues de l'accord de normalisation (STANAG, 2002) de l'agence de normalisation de l'OTAN. Ces mesures sont principalement basées sur le calcul de l'écart-type (*Standard linear error*) des erreurs de hauteur (entre le jeu de données à évaluer et le jeu de données de référence). Cet indicateur signifie qu'il existe une probabilité de 68.3% que les erreurs de positionnement vertical soient comprises dans un intervalle centré sur la moitié de la valeur calculée. Les autres mesures d'incertitude de positionnement vertical reprennent ce principe avec des pourcentages de probabilité différents :
– 50% (*Linear error probable*),
– 90% (*Linear map accuracy at 90% significance level*),
– 95% (*Linear map accuracy at 95% significance level*),
– 99% (*Linear map accuracy at 99% significance level*),
– 99,7% (*Near certainty linear error*).
Il est à noter que dans cette catégorie des mesures d'incertitude de positionnement vertical est également proposé le calcul de d'erreur moyenne quadratique en positionnement vertical (*Root Mean Square Error*).

Incertitude de positionnement horizontal Comme pour les mesures d'incertitude de positionnement vertical, les mesures d'incertitude de positionnement horizontal sont en grande partie issues de l'accord de normalisation (STANAG, 2002) de l'agence de normalisation de l'OTAN. Les mesures proposées sont également basées sur le calcul de l'écart-type des erreurs de position des coordonnées en x et en y (entre le jeu de données à évaluer et le jeu de données de référence). Chacun des indicateurs calculés définit un rayon au sein duquel la probabilité de présence des erreurs de position est de :
– 39.4% (*Circular standard deviation*),
– 50% (*Circular error probable*),
– 90% (*Circular map accuracy standard*),
– 95% (*Circular error at 95% significance level*),
– 99.8% (*Circular near certainty error*).
Cette catégorie propose également le calcul de l'erreur moyenne quadratique en positionnement horizontal (*RRoot Mean Square Error in planimetry*), ainsi que le calcul d'ellipses d'incertitude et de confiance, permettant de qualifier les éventuelles corrélations d'erreurs entre coordonnées en x et en y.

A.2 Exactitude relative

Uniquement deux mesures d'exactitude relative (ou interne) sont proposées dans le cadre de la spécification technique ISO 19138, à savoir les mesures d'erreur relative horizontale et verticale. Que ce soit en planimétrie ou en altimétrie, ces mesures sont basées sur la détermination préalable de combinaisons de paires de points homologues entre le jeu de données à évaluer et le jeu de données de référence. La différence des erreurs (en planimétrie ou altimétrie) entre chaque paire de points est ensuite calculée, afin d'en déterminer son écart-type. Le résultat de la mesure est livré en transformant l'écart-type en probabilité à 90%.

Annexe B. Descriptif des mesures proposées dans l'Arrêté du 16 septembre 2003

Contexte La France s'est dotée en 2003 d'un nouveau moyen de validation des travaux topographiques réalisés par l'Etat, les collectivités locales et leurs établissement publics. Cet arrêté (qui remplace l'ancien arrêté de 1980) concerne les travaux relatifs aux points du canevas géodésique, aux objets géographiques relatifs à ce canevas, ainsi qu'aux images rectifiées.

La notion de classes de précision L'arrêté de 2003 intègre la notion de **classes de précision**, notée $[xx]$ et exprimée en centimètres. La classe de précision permet de qualifier la précision géométrique des objets produits et sert de terme commun entre un donneur d'ordre (qui peut exiger une classe de précision de 5 cm par exemple) et un exécutant qui devra mettre en œuvre les méthodes afin de la respecter. La notion de classes de précision $[xx]$ implique l'emploi d'un "modèle standard".

Principe d'évaluation Afin de contrôler l'appartenance d'une base de données géographique à une classe de précision $[xx]$, un échantillon d'objets est extrait de la base de données. Les coordonnées des objets composant l'échantillon sont comparées à celles des données de contrôle par le biais de comparaisons ponctuelles, basées sur la distance euclidienne. Pour respecter la classe de précision testée, l'échantillon de points doit respecter plusieurs conditions basées sur des seuils établis selon des lois statistiques. Nous décrivons comment s'appliquent ces conditions dans le cas du modèle standard.

Dans le cas d'un **modèle standard**, l'échantillon de points testés doit respecter les trois conditions suivantes :

- Critère a : Ce critère se base sur l'écart moyen en position E_{moyPos}, qui correspond à la moyenne arithmétique des distances euclidiennes entre les coordonnées des points de l'échantillon et leurs homologues dans les données de référence. L'écart moyen en position E_{moyPos} doit respecter le seuil suivant :

$$E_{moyPos} < [xx] * (1 + \frac{1}{2 * C^2}) \tag{7.9}$$

où C est un coefficient de sécurité et $C \geq 2$

- Critère b : Pour ce critère, le nombre d'objets dépassant le seuil T_1 doit être inférieur à la valeur N'. Avec :

$$T_1 = k * [xx] * (1 + \frac{1}{2 * C^2}) \tag{7.10}$$

$$N' = 0.01 * N + 0.232 * \sqrt{N} \tag{7.11}$$

où N est le nombre d'objets de l'échantillon et k prenant la valeur suivante en fonction du nombre de dimensions pour les coordonnées : 3.23 pour une dimension, 2.42 pour deux dimensions, 2.11 pour trois dimensions.

- Critère c : Aucun objet ne doit dépasser le seuil T_2 avec :

$$T_2 = 1.5 * T_1 \tag{7.12}$$

Si l'échantillon testé respecte simultanément les trois critères, appelé également **gabarit d'erreurs**, on considère que la classe de précision $[xx]$ est respectée. Il est tout à fait possible dans ce cas de tester une autre classe de précision, afin de mieux qualifier la géométrie des objets de la base de données. Enfin, il est tout à fait possible de créer des gabarits d'erreurs personnalisés en définissant pour chaque seuil un taux de rejet.

Annexe C. Principes généraux de modélisation orientée-objet

Nous présenterons ici quelques principes généraux de modélisation objet, en utilisant le langage UML (*Unified Modeling Language*) pour sa formalisation.

C.1 Définitions

Le concept d'objet L'approche orientée-objet est une méthode de modélisation basée sur une représentation abstraite des entités du monde réel sous forme d'**objets**. Muller (1997) définit le concept d'objet par une identité, un état et un comportement.

- L'identité caractérise l'existence propre de l'objet au sein d'un modèle, indépendamment de son état. Le concept d'identité ne se représente pas en modélisation.
- L'état regroupe les valeurs instantanées de tous les attributs d'un objet dans le temps. L'état peut évoluer dans le temps.
- Le comportement regroupe toutes les compétences d'un objet et regroupe donc les actions et réactions que celui-ci est susceptible d'avoir. Chaque élément du comportement est appelé méthode.

Par exemple, avec l'approche orientée-objet, un objet "Bâtiment" peut être caractérisé par les attributs "Hauteur" et "Adresse" et par les méthodes "Construire" et "Détruire".

Classes et instances Le concept de **classe** permet de regrouper un ensemble d'objets possédant des états et des comportements communs. Chaque classe est donc définie par des champs (les attributs) et des opérations (les méthodes). La classe génère des objets, appelés **instances** de classe (par le mécanisme d'instanciation), qui possèdent la structure de la classe et sur lesquels les opérations définies peuvent s'appliquer. La figure 7.2 illustre une instance "Batiment1" de la classe "BATIMENT" avec les attributs la caractérisant.

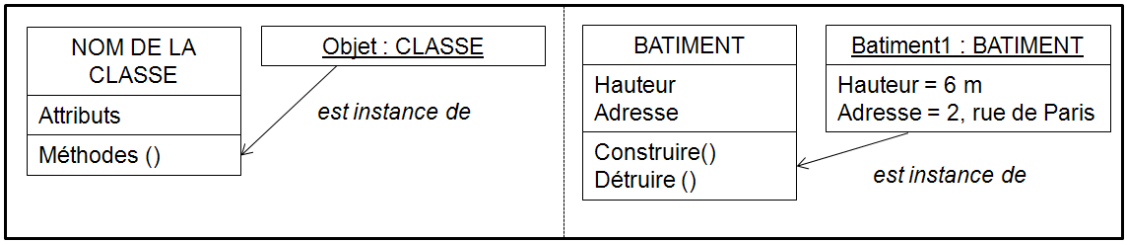


FIGURE 7.2 – Exemple d'une instance "Batiment1" de la classe "BATIMENT"

C.2 Les relations entre classes

Il existe plusieurs types de relations entres classes : les relations d'association, d'agrégation, et de spécialisation/généralisation.

Association L'association exprime une connexion sémantique durable entre deux classes, qui décrit un ensemble de liens entre objets. Une association est en général décrite par un nom, des cardinalités (qui décrivent le nombre d'objets minimum et maximum d'une classe cible participant à une relation avec un objet d'une classe source).

La figure 7.3 illustre une relation d'association entre les classes "BATIMENT" et "ACTIVITE".

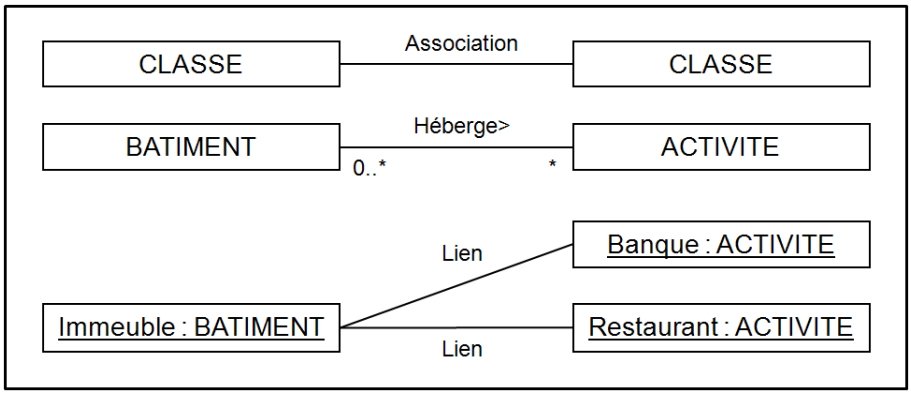


FIGURE 7.3 – Exemple de relation d'association entre la classe "BATIMENT" et la classe "ACTIVITE"

Agrégation L'agrégation est un cas particulier d'association dissymétrique exprimant une relation de contenance. L'inverse de l'agrégation est la décomposition.
La figure 7.4 illustre une relation d'agrégation entre la classe "BATIMENT" et les classes "TOIT" et "MUR".

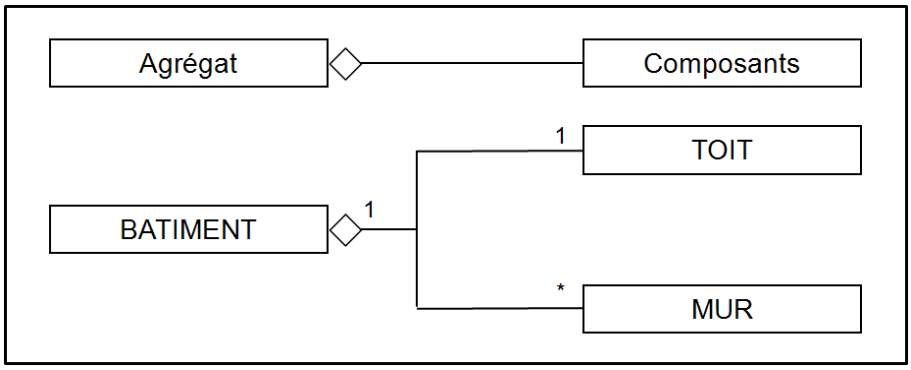


FIGURE 7.4 – Exemple de relation d'agrégation entre la classe "BATIMENT" et les classes "TOIT" et "MUR"

Spécialisation / Généralisation Ces deux types de relation permettent de traduire des relations entre classes par une notion d'héritage. La spécialisation consiste à affiner une classe existante en sous-classes plus spécifiques.

La généralisation traduit la relation inverse, et permet de regrouper les caractéristiques communes à plusieurs classes au sein d'une super-classe.
La figure 7.5 illustre une relation de généralisation par une super-classe "BATIMENT".

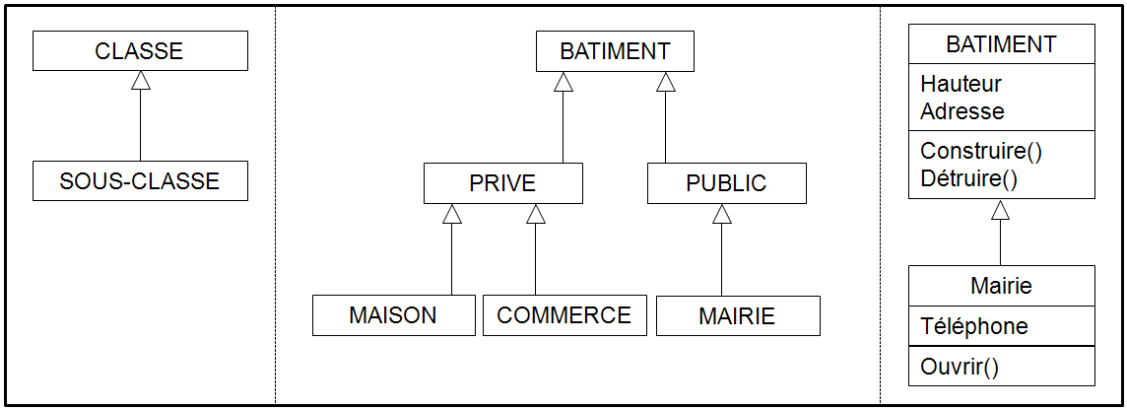


FIGURE 7.5 – Exemple de relation de généralisation par une super-classe "BATIMENT"

La modélisation par approche orientée-objet porte un intérêt important en termes de modélisation des données géographiques. Cette approche offre la possibilité de modéliser sous forme de classes d'objets des phénomènes complexes, et ainsi traduire des raisonnements plus difficilement modélisables qu'avec de simples primitives graphiques.

Annexe D. Présentation et résultats du test de digitalisation manuelle

D.1 Questionnaire utilisé

Dans le cadre de mon travail de thèse, je cherche à modéliser l'impact de l'erreur de pointé sur les mesures géométriques de longueur. Afin de valider le modèle, un test de digitalisation manuelle doit être réalisé. Ce test consiste à digitaliser une route à plusieurs échelles (1 :10.000 et 1 :25.000), à l'aide du logiciel libre *QuantumGIS*.

Préalablement à ce test, quelques questions pour établir votre profil :

Avez-vous déjà réalisé de la digitalisation manuelle ? Si c'est le cas, à quelle fréquence ?

1. Jamais
2. Occasionnellement
3. Fréquemment
4. Tous les jours

Si vous en avez déjà réalisé, sur quels types d'objets géographiques ? ...

Le test de digitalisation peut à présent commencer. Pour cela, vous trouverez dans le répertoire "TestDigit", situé sur le bureau de l'ordinateur, deux projets QGIS nommés :
– TestDigit10000.qgs
– TestDigit25000.qgs

Double-cliquez sur le projet "TestDigit10000.qgs". Le tracé d'une route (en mode raster) s'affiche à l'échelle 1 :10.000.

Afin de numériser cette route, faites un clic-droit sur la couche "Route10000.shp" et sélectionnez l'option "Basculer en mode édition". Vous pourrez digitaliser la route après avoir activé le mode capture, à l'aide du bouton Une fois la route saisie, la validation de la digitalisation s'effectue par un clic-droit.

Important : L'échelle de saisie est fixée par défaut au 1 :10.000, il ne faut pas la modifier (en particulier avec la molette de souris). L'ensemble de la digitalisation de la route doit être réalisée à échelle constante. Egalement, ne capturez la route qu'en une seule fois. Evitez donc de la numériser en plusieurs objets.

Une fois la saisie achevée, sauvegardez votre fichier dans le répertoire "Résultats" (situé dans le répertoire "TestDigit") en le nommant "TestDigit10000votrenom.shp". Pour cela, faites un clic-droit sur la couche "Route10000.shp" et sélectionnez l'option " Sauvegardez sous... ".

Fermez le projet. Vous pouvez à présent recommencer l'opération au 1 :25.000 en double-cliquant sur le projet " TestDigit25000.qgs ".

D.2 Résultats du test de digitalisation

Nous présenterons dans cette section les résultats du test de digitalisation manuelle à travers :
- le profil de l'échantillon,
- les résultats de la digitalisation au 1 :10.000,
- les résultats de la digitalisation au 1 :25.000.

ID	Fréquence de saisie	Types d'objets
1	Occasionnellement	Tous types d'objets géographiques
2	Jamais	—
3	Occasionnellement	Entités administratives, routes
4	Occasionnellement	Cours d'eau, forêts, villes, routes
5	Occasionnellement	Routes, bâtiments...
6	Jamais	—
7	Occasionnellement	Parcelles agricoles, bâtiments, trajectoires
8	Fréquemment	Routes, rivières, parcelles, points remarquables...
9	Occasionnellement	Cartes anciennes
10	Occasionnellement	Limites administratives, parcelles
11	Occasionnellement	Bâtiments et routes
12	Occasionnellement	Limites administratives, routes...
13	Jamais	—
14	Fréquemment	Données anciennes, plans Vasserot, données géo diverses
15	Occasionnellement	Objets ponctuels sur les cartes de Cassini
16	Occasionnellement	Routes, itinéraires, marqueurs, parcelles agricoles
17	Occasionnellement	Zones de végétation...
18	Occasionnellement	Routes sur OSM
19	Occasionnellement	Objets géographiques divers
20	Occasionnellement	Tous types d'objets géographiques

TABLE 7.8 – Profil général de l'échantillon

				Ecarts de Position		Dist. inter-points	
ID	**Long.(m)**	**Erreur(m)**	**NbPts**	**Moyen**	**Hausdorff**	**Min**	**Médiane**
1	4241,05	25,00	103	0,09	6,01	7,48	31,86
2	4230,80	35,24	89	0,95	10,21	8,37	24,96
3	4285,33	-19,29	86	0,36	6,56	5,92	34,40
4	4248,94	17,10	103	0,82	9,70	5,29	23,67
5	4251,39	14,66	97	0,29	7,91	10,91	33,15
6	4271,03	-4,99	121	0,49	10,91	5,29	26,46
7	4222,74	43,31	90	1,10	8,35	0,00	32,19
8	4263,50	2,54	95	0,14	6,27	0,00	26,19
9	4236,12	29,92	61	1,12	11,78	13,23	52,38
10	4272,32	-6,28	125	0,00	7,60	5,92	26,59
11	4223,48	42,56	98	0,30	9,22	7,48	29,22
12	4277,56	-11,51	134	0,09	9,04	5,29	21,17
13	4285,89	-19,84	57	0,60	12,20	0,00	68,99
14	4257,05	9,00	95	0,17	6,37	5,92	31,86
15	4336,37	-70,32	74	0,46	13,06	7,94	37,42
16	4271,80	-5,76	99	0,45	8,46	0,00	28,62
17	4255,81	10,24	103	0,48	8,33	0,00	30,17
18	4306,01	-39,97	75	0,14	7,26	0,00	35,99
19	4285,90	-19,85	71	0,42	11,04	17,75	57,79
20	4300,40	-34,36	110	0,07	6,20	7,94	28,13

TABLE 7.9 – Résultats du test de digitalisation au 1 :10.000

				Ecarts de Position		Dist. inter-points	
ID	**Long.(m)**	**Erreur(m)**	**NbPts**	**Moyen**	**Hausdorff**	**Min**	**Médiane**
1	12750,34	107,67	175	1,25	21,80	13,23	66,15
2	13032,46	-174,45	137	1,15	25,13	14,79	73,06
3	12914,35	-56,33	176	0,61	18,14	13,23	59,90
4	13263,07	-405,05	233	1,61	32,72	0,00	46,77
5	12735,26	122,75	154	1,74	26,61	14,79	72,76
6	12859,45	-1,44	216	0,54	26,00	13,23	51,66
7	12699,10	158,92	158	0,16	20,88	0,00	71,55
8	13038,57	-180,55	186	0,49	23,87	0,00	53,33
9	12749,38	108,63	136	1,65	32,15	9,35	79,65
10	12820,41	37,60	214	0,82	22,27	6,61	51,66
11	12749,28	108,73	159	1,57	18,63	13,23	73,06
12	12932,64	-74,63	319	0,77	32,71	0,00	33,07
13	12927,66	-69,65	134	0,96	25,84	6,61	79,65
14	12736,95	121,06	169	1,05	22,47	14,79	66,15
15	13117,16	-259,15	140	1,29	32,02	13,23	75,42
16	12780,41	77,61	195	0,70	28,80	14,79	59,53
17	12797,81	60,21	211	0,46	22,40	0,00	52,92
18	12976,42	-118,41	153	0,06	23,52	0,00	68,10
19	12956,27	-98,25	134	0,85	20,75	23,85	89,97
20	12929,26	-71,24	186	0,04	21,07	14,79	68,10

TABLE 7.10 – Résultats du test de digitalisation au 1 :25.000

Annexe E. Résultats des comparaisons de longueurs sur les routes généralisées

Nous présenterons dans cet annexe les distributions des erreurs relatives de longueur entre tronçons de route généralisés (extraits de la BDCARTO, des données TOP100, et des données de la Carte Régionale) et leurs homologues de référence (extraits de la BDTOPO). Les distributions des erreurs relatives de longueur sont classées en fonction du contexte géographique des objets et de la base de données généralisée testée.

Ces résultats ont été utilisés afin de mettre au point le modèle statistique d'estimation de l'imprécision de longueur sur les routes généralisées.

Par ailleurs, la figure 7.6 présente l'interface développée afin de réaliser l'appariement et les comparaisons de mesures géométriques entre objets homologues.

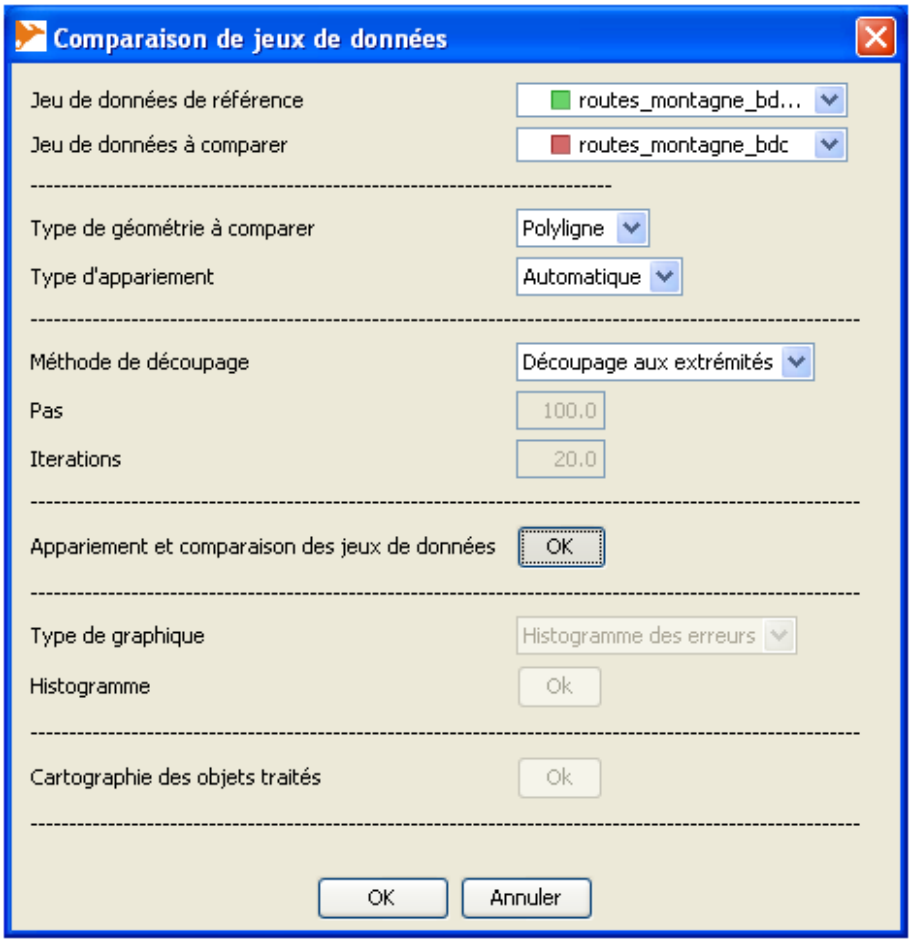


FIGURE 7.6 – Interface d'appariement et de comparaison des mesures géométriques entre objets homologues

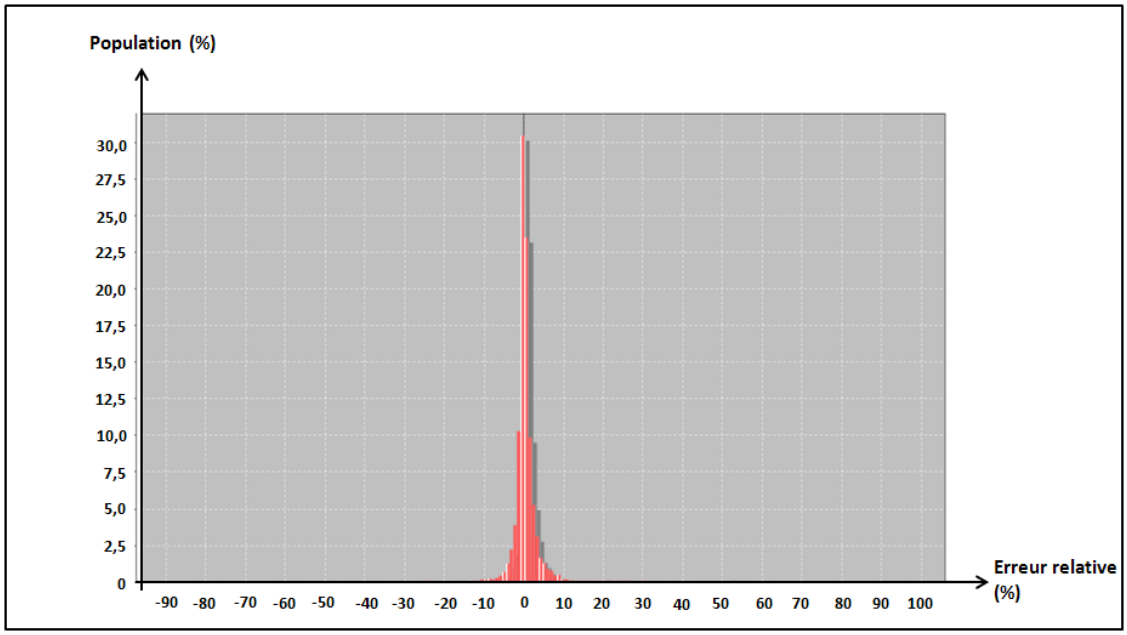


FIGURE 7.7 – Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone rurale

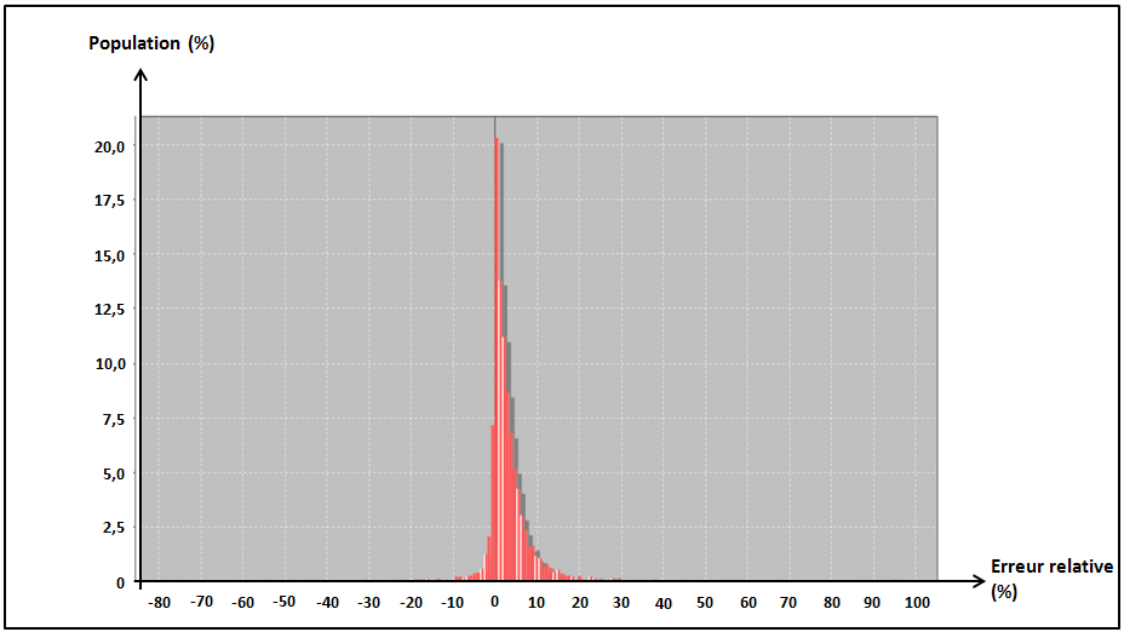


FIGURE 7.8 – Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone rurale

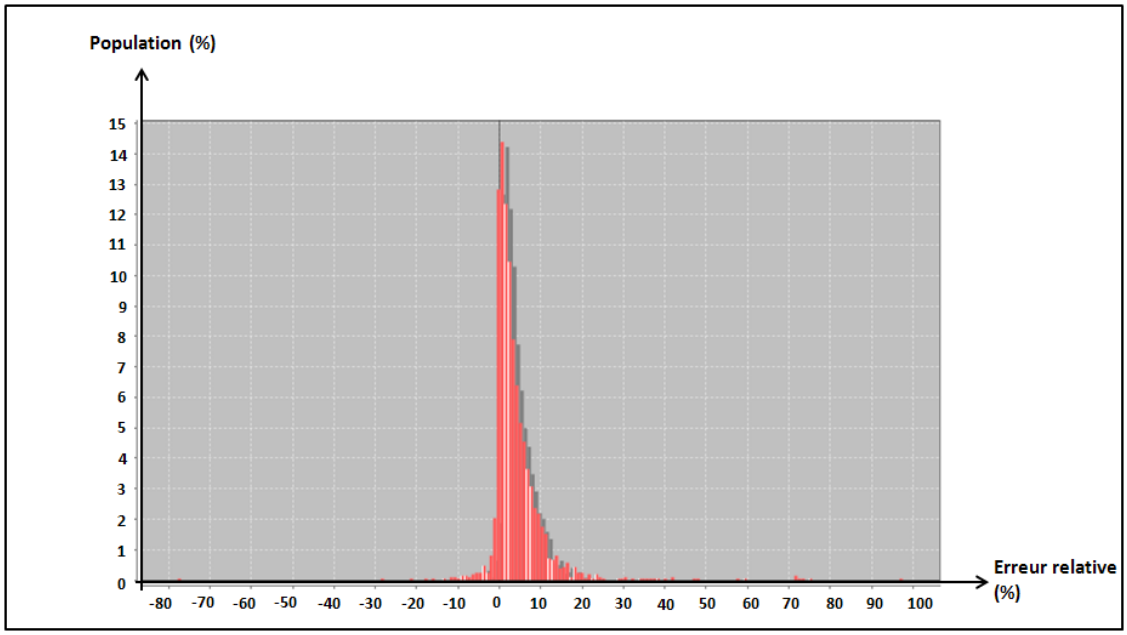


FIGURE 7.9 – Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone rurale

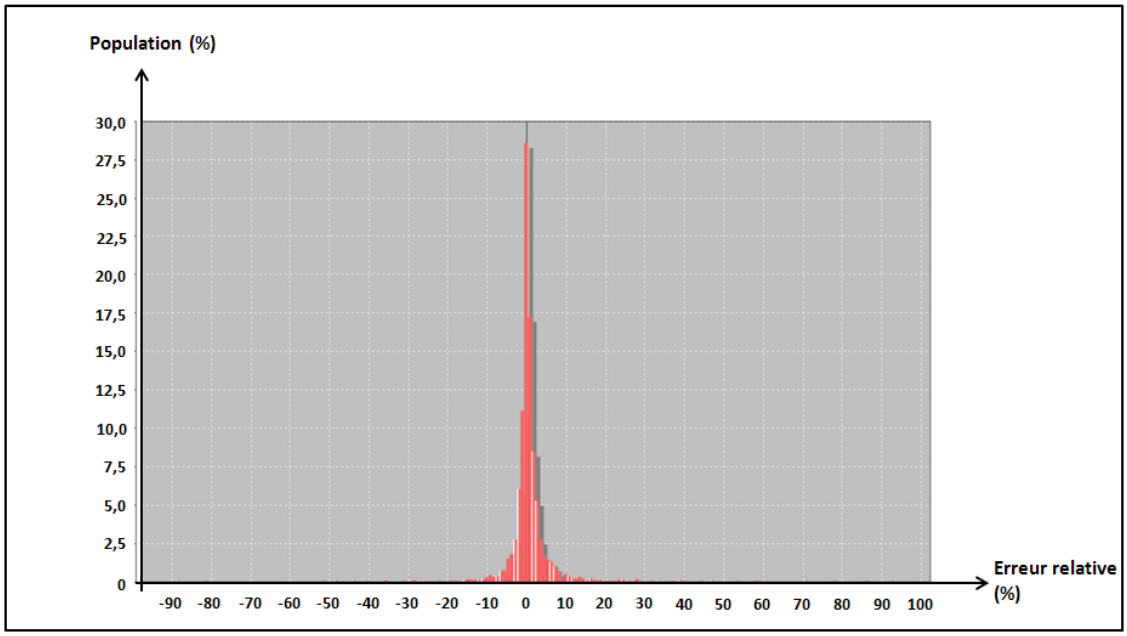


FIGURE 7.10 – Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone urbaine

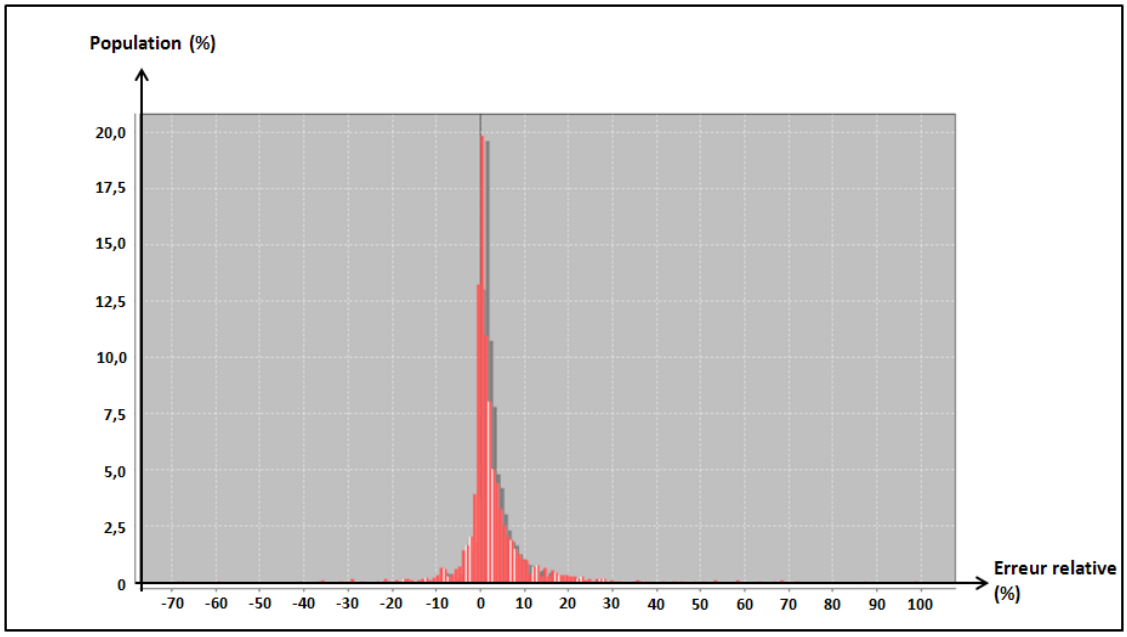


FIGURE 7.11 – Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone urbaine

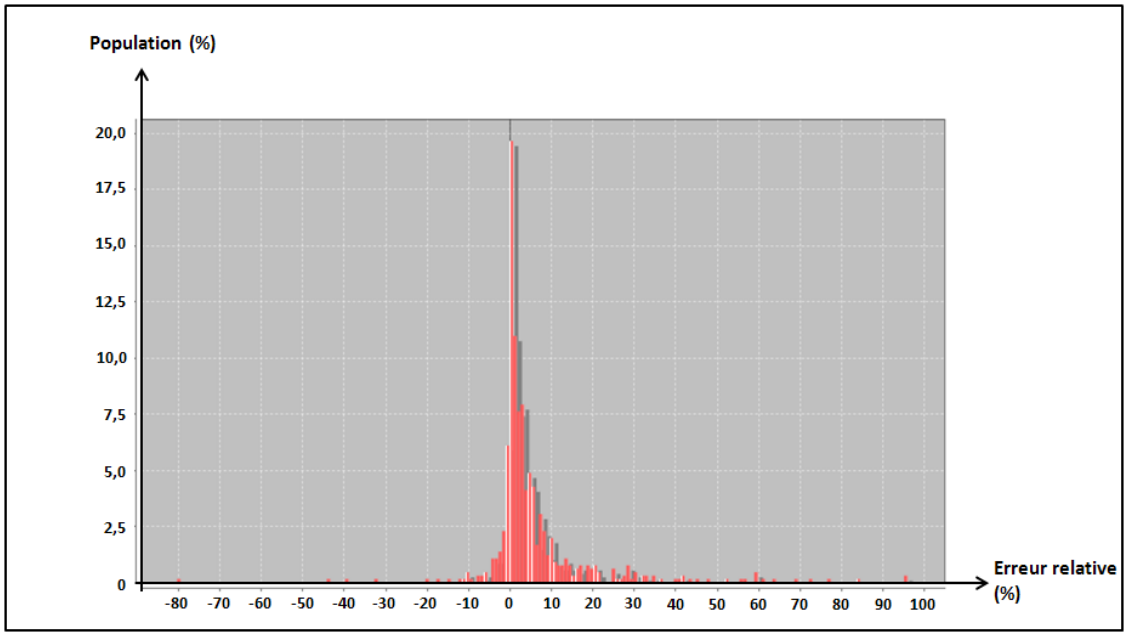


FIGURE 7.12 – Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone urbaine

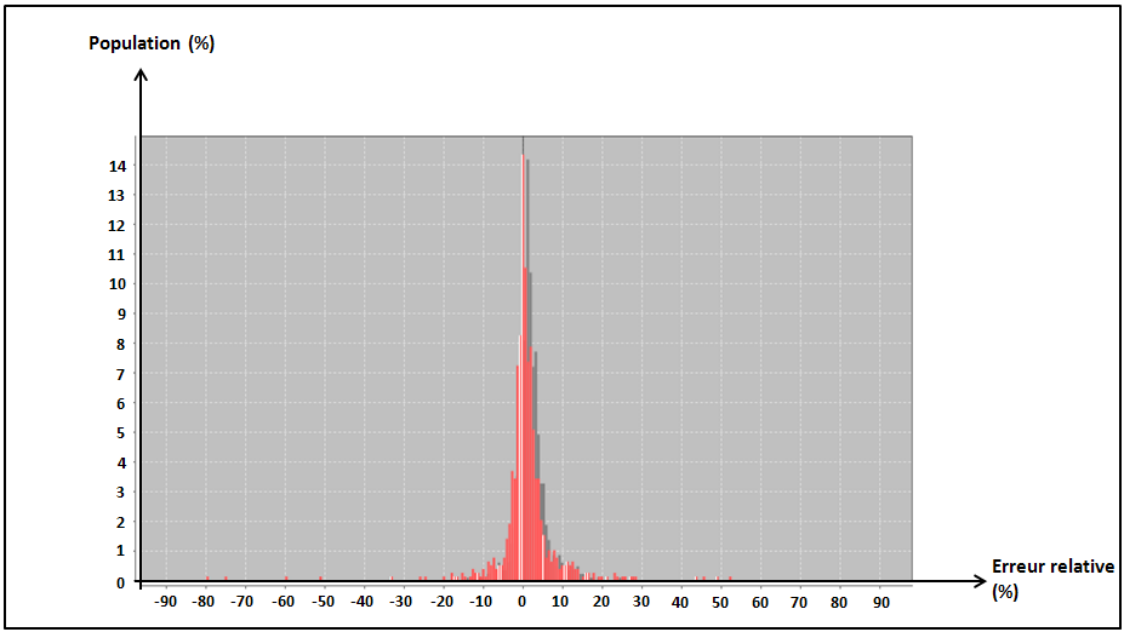


FIGURE 7.13 – Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone de montagne

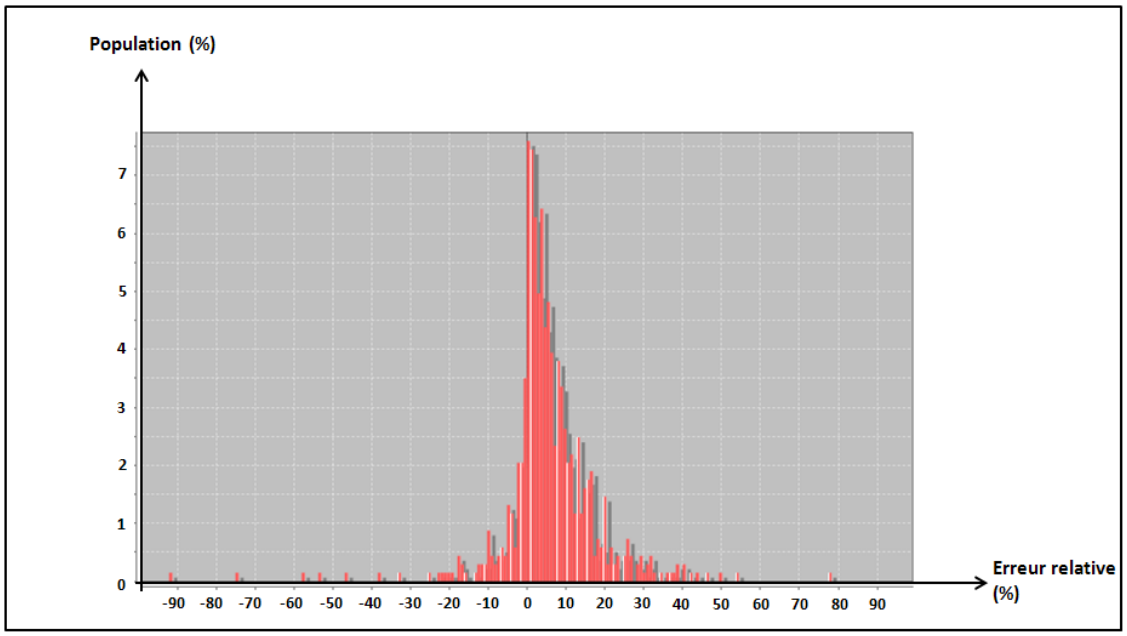


FIGURE 7.14 – Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone de montagne

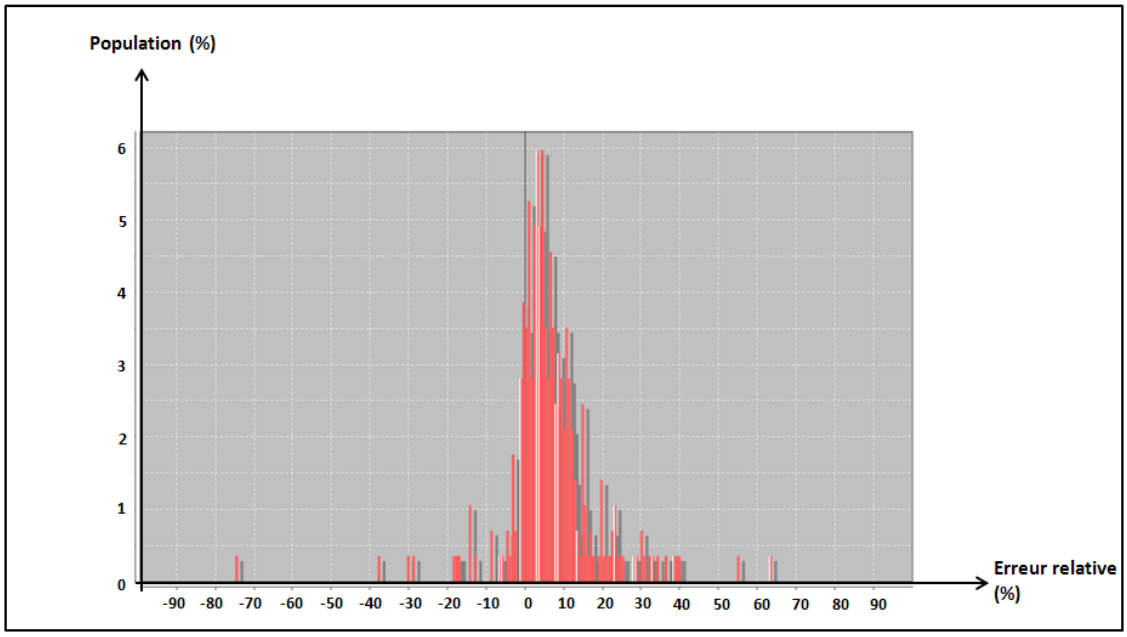


FIGURE 7.15 – Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone de montagne

Bibliographie

Abbas, I. (1994). *Base de données vectorielles et erreur cartographique : Problèmes posés par le contrôle ponctuel Une méthode alternative fondée sur la distance de Hausdorff : le contrôle linéaire*. Thèse de doctorat, Université Paris 7. 54

Affholder, J.-G. (2002). Exagération des formes basée sur une nouvelle modélisation du linéaire routier. *In Journées de la Recherche de l'IGN*, Saint-Mandé, France. 142

AFNOR (1955). Vocabulaire international de métrologie - concepts fondamentaux et généraux et termes associés. afnor nf x07-001. Rapport technique, Association Française de Normalisation. 26

Agumya, A. et Hunter, G. (1996). Assessing fitness for use of spatial information : Information utilisation and decision uncertainty. *In Proceedings of the GIS/LIS '96 Conference*, pages 359–370, Denver, Colorado. 44

Alt, H. et Godau, M. (1995). Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry and Applications*, 5:75–91. 55

Arkin, E. M., Chew, L. P., Huttenlocher, D. P., K., K. et Mitchel, J. S. B. (1991). An efficient computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:121–147. 58

Aronoff, S. (1989). *Geographic information systems : a management perspective*. Ottawa : WDL Publications. 47

Badard, T. et Lemarié, C. (2002). Associer les données : l'appariement. *In* Ruas, A., éditeur : *Généralisation et représentation multiple*, pages 163–183. Hermes et Lavoisier. 55, 359

Bard, S. (2004). *Méthode d'évaluation de la qualité des données géographiques généralisées*. Thèse de doctorat, Université Paris 6. 38, 202

Bédard, Y. (1987). Uncertainties in land information systems databases. *In Proceedings of the 8th International Symposium on Computer-Assisted Cartography*, pages 175–184, Baltimore, Maryland. American Society for Photogrammetry and Remote Sensing. 44

Beard, K. et Buttenfield, B. (1991). Visualization of spatial data quality. Rapport technique, National Center for Geographic Information and Analysis, Castine, Maine. 50

Beard, K. M. (1989). Use Error : The Neglected Error Component. *In Proceedings of the AUTOCARTO 9 Conference*, pages 808–817. 13, 29, 44, 88

Bel Hadj Ali, A. (2001). *Qualité géométrique des entités géographiques surfaciques. Application à l'appariement et définition d'une typologie des écarts géométriques.* Thèse de doctorat, Université Marne-La-Vallée. 58, 229, 248, 365

Blakemore, M. (1984). Generalisation and Error in Spatial Data Bases. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 21(2):131–139. 38, 202

Blaschke, T., Lang, S. et Hay, G., éditeurs (2008). *Object-based image analysis.* Springer. 36

Boffet, A. (2000). Creating urban information for cartographic generalisation. *In Proceedings of the 9th International Symposium on Spatial Data Handling*, Beijing, China. 223, 224, 364

Boin, A. T. et Hunter, G. J. (2008). What communicates quality to the spatial data consumer ? *In* Stein, A., Bijker, W. et Shi, W., éditeurs : *Quality Aspects in Spatial Data Mining.* 50, 75

Bolstad, P. V., Gessler, P. et Lillesand, T. M. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, 4(4):399–412. 62, 79, 162

Bonin, O. (2002). *Modèles d'erreurs dans une base de données géographiques et grandes déviations pour des sommes pondérées ; application à l'estimation d'erreurs sur un temps de parcours.* Thèse de doctorat, Université Paris 6. 66, 138, 163, 165, 166, 179, 361, 362

Booch, G., Rumbaugh, J. et Jacobson, I. (2000). *Le guide de l'utilisateur UML.* Eyrolles. 97

Bosser, P. (2011). *GNSS : Systèmes Globaux de Positionnement par Satellite.* Ecole Nationale des Sciences Géographiques - Département Positionnement Terrestre et Spatial. 32, 33, 359, 369

Botton, S. (2010). *Systèmes de référence et de coordonnées : Représentations planes de l'ellipsoïde.* Ecole Nationale des Sciences Géographiques - Département Positionnement Terrestre et Spatial. 109

Bouchon-Meunier, B. (1995). *La logique floue et ses applications.* Addison-Wesley France. 41, 42, 43, 45, 52

Bouteloup, D. (2006). *Topométrie : Mesure des distances.* Ecole Nationale des Sciences Géographiques. 109

Brasebin, M. (2009). Geoxygene : An open 3d framework for the development of geographic applications. *In Proceedings of the 12th International Conference on Geographic Information Science (AGILE'09)*, Hanovre, Germany. 132, 292

Brassel, K. E. et Weibel, R. (1988). A review and conceptual framework of automated map generalization. *International journal of geographical information systems*, 2(3):229–244. 38

Brown, J. et Heuvelink, G. (2007). The data uncertainty engine (due) : A software tool for assessing and simulating uncertain environmental variables. *Computers & Geosciences*, 33:172–190. 50, 66

Bucher, B., Brasebin, M., Buard, E., Grosso, E. et S., M. (2009). Geoxygene : built on top of the expertness of the french nma to host and share advanced gi science research results. *In Proceedings of the International Opensource Geospatial Research Symposium 2009 (OGRS'09)*, Nantes, France. 292

Burrough, P. A. (1986). *Principles of geographical information systems for land resources assessment.* Clarendon Press, Oxford Oxfordshire and New York. 28

Caspary, W. et Scheuring, R. (1992). Error-band as Measures of Geographic Accuracy. *In Proceedings of EGIS'92*, pages 226–233. 62

Chandelier, L. (2011). *La prise de vues photogrammétriques.* École Nationale des Sciences Géographiques - Département Imagerie Aérienne et Spatiale. 168, 169, 184

Chaudhry, O. Z. et Mackaness, W. A. (2008). Creating Mountains out of Mole Hills : Automatic Identification of Hills and Ranges Using Morphometric Analysis. *Transactions in GIS*, 12(5):567–589. 219, 220

Chaudry, O. et Mackaness, W. A. (2008). Automatic identification of urban settlement boundaries for multiple representation databases. *Computers, Environment and Urban Systems*, 32(2):95–109. 223

Cheung, C. K. et Shi, W. (2004). Estimation of the positional uncertainty in line simplification in gis. *Cartographic Journal*, 41:37–45. 67

Cheung, C. K., Shi, W. et Zhou, X. (2004). A Probability-based Uncertainty Model for Point-in-Polygon Analysis in GIS. *GeoInformatica*, 8(1):71–98. 67

Chrisman, N. R. (1982). A Theory of Cartographic Error and its Measurement in Digital Data Bases. *In Proceedings the AUTOCARTO 5 Conference*, pages 159–168. 62, 63

Chrisman, N. R. (1984). The Role Of Quality Information In The Long-Term Functioning Of A Geographic Information System. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 21(2):79–88. 46

Chrisman, N. R. (1989). Errors In Categorical Maps : Testing versus Simulation. *In Proceedings the AUTOCARTO 9 Conference*, pages 521–529. 61

Chrisman, N. R. (1991). A diagnostic test for error in categorical maps. *In Proceedings the AUTOCARTO 10 Conference*, pages 330–348. 13, 61

Chrisman, N. R. (1995). Living with error in geographic data : Truth and responsibility. *In Proceedings of the GIS '95 Conference*, pages 12–17. 29

Chrisman, N. R. et Yandell, B. S. (1988). Effects of Point Error on Area Calculations : A Statistical Model. *Surveying and Mapping*, 48:241–246. 68, 79

Cohen, S. D. et Guibas, L. J. (1997). Partial matching of planar polylines under similarity transformations. *In Proceedings of the 8th Annual ACMSIAM Symposium on Discrete Algorithms*. 57

Collins, F. et Smith, J. (1994). Taxonomy for error in gis. *In Proceedings of the International Symposium on the spatial accuracy of Natural Resource Data Bases, ASPRS*, pages 1–7. 29

Comber, A., Fisher, P. et Wadsworth, R. (2005). What is land cover ? *Environment and Planning B : Planning and Design*, 32:199–209. 87, 88, 360

Cuenin, R. (1972). *Cartographie Générale. Notions générales et principes d'élaboration.* Eyrolles, collection scientifique de l'institut géographique national édition. 199, 208

Dana, P. H. (1997). *The Geographer's Craft Project.* Department of Geography, The University of Colorado, Boulder. 81, 360

David, B. et Fasquel, P. (1997). Qualité d'une base de données géographique : concepts et terminologie. *Bulletin d'Information de l'IGN*, 67. 19, 47, 48, 53

De Bruin, S. (2008). Modelling Positional Uncertainty of Line Features by Accounting for Stochastic Deviations from Straight Line Segments. *Transactions in GIS*, 12(2):165–177. 177

De Bruin, S., Heuvelink, G. B. M. et Brown, J. D. (2008). Propagation of positional measurement errors to agricultural field boundaries and associated costs. *Computers and Electronics in Agriculture*, 63(2):245–256. 66, 68, 79, 162

Devillers, R. (2004). *Conception d'un système multidimensionnel d'information sur la qualité des données géospatiales.* Thèse de doctorat, Université Laval - Université Marne-La-Vallée. 13, 44, 50

Devillers, R. (2005). Qualité de l'information géographique : concepts généraux. *In* Devillers, R. et Jeansoulin, R., éditeurs : *Qualité de l'information géographique*, pages 37–48. Hermès, Lavoisier. 29, 46, 359

Devillers, R., Stein, A., Bédard, Y., Chrisman, N. R., Fisher, P. et Shi, W. (2010). Thirty Years of Research on Spatial Data Quality : Achievements, Failures, and Opportunities. *Transactions in GIS*, 14(4):387–400. 41

Devogèle, T. (1997). *Processus d'intégration et d'appariement de bases de données géographiques, Application à une base de données routières multi-échelles.* Thèse de doctorat, Université de Versailles. 39, 55, 56, 229

Douglas, D. et Peucker, T. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122. 139, 150, 186, 204, 224

Duchêne, C. (2002). *Généralisation par agents communicants : Le modèle CARTACOM. Application aux données topographiques en zone rurale.* Thèse de doctorat, Université Paris 6. 38, 205

Duckham, M. et McCreadie, J. (2002). Error aware gis development. *In* Shi, W., Fisher, P. et Goodchild, M., éditeurs : *Spatial Data Quality*, pages 63–75. Taylor & Francis. 50

Dunham, J. G. (1986). Optimum uniform piecewise linear approximation of planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):67–75. 137

Dutton, G. (1992). Handling positionnal uncertainty in spatial databases. *In Proceedings of the 5th International Symposium on Spatial Data Handling*, Charleston, South Carolina, USA. 63, 360

EEA (2007). Corine land cover 2006 technical guidelines. Rapport technique, European Environment Agency. 219, 233

Epstein, E. F., Hunter, G. J. et Agumya, A. (1998). Liability insurance and the use of geographical information. *International Journal of Geographical Information Science*, 12:203–214. 44

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D. et Alsdorf, D. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45:33. 131, 132, 135, 232, 273, 318

Faïz, S. O. (1999). *Systèmes d'Information Géographiques : Information Qualité et Data Mining*. Editions C.L.E., Tunis, Tunisie. 50

Fisher, P. (2003). Models of uncertainty in spatial data. *Geographical Information System*, 1:191–203. 41, 42, 43, 359

Fisher, P., Comber, A. et Wadsworth, R. (2005). Nature de l'incertitude pour les données spatiales. *In* Devillers, R. et Jeansoulin, R., éditeurs : *Qualité de l'Information Géographique*, pages 49–64. Hermès et Lavoisier. 41, 42, 43, 359

Fisher, P. et Tate, N. (2006). Causes and consequences of error in digital elevation models. *Progress in Physical Geography*, 30:467–489. 130

Fouqué, L. (1999). Simulation d'erreurs dans une base de données géographique. Mémoire de stage de dess de mathématiques appliquées, Université Rennes 1. 165, 166, 179, 362

Franck, A. U. et Navratil, G. (2011). Ignoring correlation leads to bone shaped confidence regions and other counter-intuitive aspects of spatial data quality. *In* Fonte, C., Gonçalves, L. et Conçalves, G., éditeurs : *Proceedings of the 7th Internattional Symposium on Spatial Data Quality*, pages 95–102, Coimbra, Portugal. 64, 360

Frank, A. U. (2007). Data quality ontology : an ontology for imperfect knowledge. *In Proceedings of the 8th international conference on Spatial information theory (COSIT'07)*, pages 406–420, Berlin, Heidelberg. Springer-Verlag. 41

Franklin, C. (1992). An introduction to geographic information systems : Linking maps to databases. *Database*, 15:13–21. 18

Gaffuri, J. (2008). *Généralisation automatique pour la prise en compte de thèmes champ : le modèle GAEL*. Thèse de doctorat, Université Paris-Est. 37, 199, 200, 201, 359, 363

Garnesson, P. et Giraudon, G. (1992). L'Approximation polygonale : bilans et perspectives. Research Report RR-1621, INRIA. 137

Gervais, M. (2004). *Pertinence d'un manuel d'instructions au sein d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques.* Thèse de doctorat, Université Laval, Québec. 75

Gesbert, N., Libourel, T. et Mustière, S. (2004). Apport des spécifications pour les modèles de bases de données géographiques. *Revue Internationale de Géomatique*, 14(2):239–257. 20

Girres, J.-F. (2011a). An evaluation of the impact of cartographic generalisation on length measurement computed from linear vector databases. *In Proceedings of the 25th International Cartographic Conference (ICC'11)*, Paris, France. 223, 228, 233, 234, 237, 365, 369

Girres, J.-F. (2011b). A model to estimate length measurements uncertainty in vector databases. *In Proceedings of the 7th International Symposium on Spatial Data Quality*, pages 83–88, Coimbra, Portugal. 98

Girres, J.-F. et Julien, P. (2010). Estimation of digitizing and polygonal approximation errors in the computation of length in vector databases. *In Proceedings of the 9th International Symposium on Spatial Accuracy Assessement in Natural Resources & Environmental Sciences (ACCURACY'10)*, Leicester, UK. 69

Girres, J.-F. et Ruas, A. (2010). Estimation of imprecision in length and area computation in vector databases including production processes description. *In Proceedings of the 14th International Symposium on Spatial Data Handling (SDH'10), Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, Hong-Kong. 89

Girres, J.-F. et Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459. 325

Goodchild, M., Shortridge, A. et P, F. (1999). Encapsulating simulation models with geospatial data sets. *In* Lowell, K. et Jaton, A., éditeurs : *Spatial Accuracy Assessment : Land Information Uncertainty in Natural Resources*, pages 123–129. Ann Arbor Press. 68, 162

Goodchild, M. F. (1995). Sharing imperfect data. *In* Onsrud, H. J. et Rushton, G., éditeurs : *Sharing Geographic Information*, pages 413–425. Rutgers University Press. 41, 43, 44

Goodchild, M. F. (2004). A general framework for error analysis in measurement-based GIS. *Journal of Geographical Systems*, 6(4):323–324. 68, 69, 79

Goodchild, M. F. (2007). Citizens as voluntary sensors : spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2:24–32. 13

Goodchild, M. F. et Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, pages 299–306. 62, 360

GRIFFITH, D. A. (1989). Distance calculations and errors in geographic databases. *In* GOODCHILD, M. F. et GOPAL, S., éditeurs : *The Accuracy Of Spatial Databases*, pages 81–90. Taylor & Francis. 68

GROSSO, E., PERRET, J. et BRASEBIN, M. (2012). Geoxygene : an interoperable platform for geographical application development. *In Innovative Software Development in Gis*, chapitre 3, pages 67–90. John Wiley & Sons. 132, 292

GUPTILL, S. C. et MORRISSON, J. L., éditeurs (1995). *Elements of Spatial Data Quality*. Pergamon Press Inc, 1st édition. 47

HAKLAY, M. (2010). How good is volunteered geographical information ? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B : Planning and Design*, 37(4):682–703. 62

HARRIE, L. et SARJAKOSKI, T. (2002). Simultaneous graphic generalization of vector data sets. *Geoinformatica*, 6(3):233–261. 67

HE, Y. (2007). *Spatial Data Quality Management*. Thèse de doctorat, The University of New South Wales. 50

HELBICH, M., AMELUNXEN, C. et NEIS, P. (2011). Comparative spatial analysis of positional accuracy of openstreetmap and proprietary geodata. *In Proceedings of the 14th International Conference on Geographic Information Science (AGILE'11)*, Utrecht, Netherlands. 318

HENRYE, J.-B. (2005). *Cours de Topographie et Topométrie Générale*. Université Louis Pasteur de Strasbourg - Ecole et Observatoire des Sciences de la Terre. 31

HEO, J., KIM, J. H. et KIM, J. W. (2009). A new methodology for measuring coastline recession using buffering and non-linear least squares estimation. *International Journal of Geographical Information Science*, 23(9):1165–1177. 62

HEUVELINK, G. B. M. (1998). *Error Propagation in Environmental Modelling with GIS*. CRC Press. 26

HEUVELINK, G. B. M., BROWN, J. D. et van LOON, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Systems*, 21:497–513. 66, 67, 177, 360

HIRT, C., FILMER, M. et FEATHERSTONE, W. (2010). Comparison and validation of recent freely-available aster-gdem ver1, srtm ver4.1 and geodata dem-9s ver3 digital elevation models over australia. *Australian Journal of Earth Sciences*, 57:337–347. 131

HOPE, S. et HUNTER, G. J. (2007). Testing the effects of positional uncertainty on spatial decision-making. *International Journal of Geographical Information Science*, 21(6):645–665. 50

HUNTER, G. J. (1998). Managing uncertainty in gis. Rapport technique, NCGIA Core Curriculum in GIScience. 41, 44

HUNTER, G. J. et BEARD, K. (1992). Understanding error in spatial databases. *Australian Surveyor*, 37:108–119. 29, 359

HUNTER, G. J. et GOODCHILD, M. F. (1996). A New Model For Handling Vector Data Uncertainty in Geographic Information Systems. *URISA Journal*, 8(1):51–57. 66, 162, 166, 190, 360

IGN (2011a). Bdalti version 1 - descriptif de contenu. Rapport technique, Institut Géographique National. 132

IGN (2011b). Bdcarto version 3.1 - descriptif de contenu. Rapport technique, Institut Géographique National. 23, 132, 170, 258

IGN (2011c). Bdtopo version 2.1 - descriptif de contenu. Rapport technique, Institut Géographique National. 20, 21, 23, 122, 132, 170, 359

IGN (2012a). Descriptif technique - carte routière france. Rapport technique, Institut Géographique National. 202

IGN (2012b). Fiche produit - carte régionale routière et touristique. Rapport technique, Institut Géographique National. 201, 202

IGN (2012c). Fiche produit - carte top100 tourisme et découverte. Rapport technique, Institut Géographique National. 201, 202

IGN (2012d). Scan25 version 2 - descriptif de contenu. Rapport technique, Institut Géographique National. 201

IGN-SDC (2004). Spécifications du fond cartographique de l'atlas routier a4 au 1/250.000. Rapport technique, Institut Géographique National - Service De la Cartographie. 209

IGN-SDC (2007). Spécifications du fonds cartographique top100. Rapport technique, Institut Géographique National - Service De la Cartographie. 209

IGN-SGN (2010). Circé france version 4. Rapport technique, Institut Géographique National - Service de Géodésie et Nivellement. 114

IGN-SHOM (1992). Trait de côte histolitt verison 2.0 - descriptif technique. Rapport technique, Institut Géographique National - Service Hydrographique et Océanographique de la Marine. 168

ISO19107 (2003). Information géographique – schéma spatial. Rapport technique, International Organization for Standardization. 293

ISO19113 (2002). Information géographique - principes qualité. Rapport technique, International Organization for Standardization. 48, 49, 50

ISO19114 (2003). Information géographique - procédures d'évaluation de la qualité. Rapport technique, International Organization for Standardization. 48, 49

ISO19115 (2003). Information géographique - métadonnées. Rapport technique, International Organization for Standardization. 48, 49

ISO8402 (1994). Norme internationale - management de la qualité et assurance de la qualité - vocabulaire. Rapport technique, International Organization for Standardization. 19, 46

ISODIS19157 (2012). Information géographique - qualité de données. Rapport technique, International Organization for Standardization. 49, 58

ISOTS19138 (2006). Information géographique - mesures de la qualité des données. Rapport technique, International Organization for Standardization. 48, 49, 50, 58, 328

Jaara, K. et Lecordix, F. (2011). Extraction of cartographic contour lines using digital terrain model (dtm). *The Cartographic Journal*, 48:131–137. 221

João, E. (1998). *Causes And Consequences Of Map Generalization*. CRC Press. 38, 202, 203, 204, 205, 233, 235, 363

Julien, P. (2008). Note sur les erreurs dans le calcul des longueurs et des aires dans une base de données en mode vectoriel. Rapport technique, Institut Géographique National - Laboratoire MATIS. 69

Keefer, B., Smith, J. et Gregoire, T. (1988). Simulating manual digitizing error with statistical models. *In Proceedings of GIS/LIS'88 Conference*, pages 475–483. 162

Kiiveri, H. T. (1997). Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, pages 33–52. 68, 69, 79

Kochanek, D. et Bartels, R. (1984). Interpolating splines with local tension, continuity and bias control. *ACM SIGGRAPH*, 18(3):33–41. 144

Kresse, W. et Fadaie, K. (2004). *ISO standards for geographic information*. Berlin : Springer. 47

Kunt, M. (2000). *Traitement de l'information : reconnaissance des formes et analyse de scènes*. Presse Polytechniques et Universitaires Romandes. 146

Leung, Y., Ma, J.-H. et Goodchild, M. F. (2004). A general framework for error analysis in measurement-based GIS Part 2 : The algebra-based probability model for point-in-polygon analysis. *Journal of Geographical Systems*, 6(4):355–379. 68

Leung, Y. et Yan, J. (1998). A location error model for spatial features. *International Journal of Geographical Information System*, 12:607–620. 65, 67

Liu, Z. J., Wang, J. et Liu, W. (2005). Building extraction form high resolution imagery based on multi-scale object oriented classification and probabilistic hough transform. *In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'05)*, Seoul, South Korea. 36, 359

Lohani, B. (2010). *Airborne Altimetric LiDAR : Principle, Data Collection, Processing and Applications*. Indian Institut of Technology Kanpur - Department of Civil Engineering. 34

Longley, P. A., Goodchild, M. F., Maguire, D. J. et Rhind, D. W., éditeurs (2005). *Geographic Information Systems and Science*. John Wiley & Sons. 44

McMaster, R. (1986). A statistical analysis of mathematical measures fior linear simplification. *The American Cartographer*, 23:103–116. 56

Muller, P.-A. (1997). *Modélisation objet avec UML*. Eyrolles. 332

Mustière, S. (2001). *Apprentissage supervisé pour la généralisation cartographique*. Thèse de doctorat, Université Paris 6. 38, 79, 139, 148, 149, 151, 152, 153, 200, 201, 203, 205, 212, 213, 362, 363, 364

Mustière, S. et Devogèle, T. (2008). Matching Networks with Different Levels of Detail. *GeoInformatica*, 12(4):435–453. 53, 55, 229, 235, 240, 270

Oksanen, J. (2006). *Digital Elevation Model Error in Terrain Analysis*. Thèse de doctorat, University of Helsinki. 130

Olteanu, A.-M. (2008). *Fusion de connaissances pour l'appariement de données géographiques. Proposition d'une approche s'appuyant sur la théorie des fonctions de croyance*. Thèse de doctorat, Université Paris-Est. 41, 43, 53, 55, 56, 58, 229, 359, 360

OpenJUMP (2008). Openjump gis - the free and open sourcejava-based desktop gis. 310

Perkal, J. (1954). On epsilon length. *Bulletin de l'académie polonaise des sciences*, 4:399–403. 61

Plazanet, C. (1996). *Enrichissement des bases de données géographiques : analyse de la géométrie des objets linéaires pour la généralisation cartographique*. Thèse de doctorat, Université Marne-La-Vallée. 56, 148, 149, 151, 362

Prisley, S., Gregoire, T. et Smith, J. (1989). The mean and variance of area estimates computed in an arc-node geographical information system. *Photogrammetric Engineering and Remote Sensing*, 55:1601–1612. 68

Qiu, J. et Hunter, G. J. (2002). A gis with the capacity for managing data quality information. *In* Shi, W., Goodchild, M. et Fisher, P., éditeurs : *Spatial Data Quality*, pages 230–250. Taylor & Francis. 50

Quodverte, P. (1994). *Cartographie numérique et information géographique*. Thèse de doctorat, Université d'Orléans. 18

Ray, C. (2009). Dispositifs mobiles, ubiquité : Mobilité, contexte et localisation. *In Ecole d'été MAGIS*, Aussois, France. 80, 360

Ray, C., Comblet, F., Jean-Marie, B. et Le Roux, Y.-M. (2010). Wireless and information technologies supporting intelligent location-based services. *In* Zhou, M.-T., Zhang, Y. et Yang, L., éditeurs : *Wireless Technologies in Intelligent Transportation Systems*, pages 225–265. Nova Science Publishers. 80

Rüeger, J. (1996). *Electronic Distance Measurement - An Introduction*. Springer-Verlag. 109

Ruas, A. (1999). *Modèle de généralisation des données géographiques à base de contraintes et d'autonomie*. Thèse de doctorat, Université Marne la Vallée. 38, 201, 205

Ruas, A. (2002). Echelle et niveau de détail. *In* Ruas, A., éditeur : *Généralisation et représentation multiple*, chapitre Echelle et niveau de détail, pages 26–44. Hermes, Lavoisier, Paris. 22, 23, 37, 139, 169, 207, 359

Ruas, A. et Mustière, S. (2005). Bases de données géographiques et cartographiques à différents niveaux de détail. *Bulletin du Comité Français de Cartographie*, 185:5–14. 21, 169

Schneider, M. (1999). Uncertainty management for spatial data in databases : Fuzzy spatial data types. *In Proceedings of the 6th International Symposium On Advances in Spatial Databases*, Hong Kong, China. 41

Serra, J., éditeur (1983). *Image Analysis and Mathematical Morphology*. Academic Press Inc. 223, 224, 226

Servigne, S., Lesage, N. et Libourel, T. (2005). Composantes qualité et métadonnées (standardisation et mise en oeuvre standardisée). *In* Devillers, R. et Jeansoulin, R., éditeurs : *Qualité de l'information géographique*, pages 205–237. Hermès, Lavoisier. 49, 75

Sheeren, D. (2005). *Méthodologie d'évaluation de la cohérence inter-représentations pour l'intégration de bases de données spatiales*. Thèse de doctorat, Université Paris 6. 39

Shi, W., Ehlers, M. et Tempfli, K. (1999). Analytical modelling of positional and thematic uncertainties in the integration of remote sensing and geographical information systems. *Transactions in GIS*, 3:119–136. 63, 64, 360

Shi, W. et Liu, W. (2000). A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science*, pages 51–66. 63

Shortridge, A. (2001). Characterizing uncertainty in digital elevation models. *In* Hunsaker, C., Goodchild, M., Friedl, M. et Case, T., éditeurs : *Spatial Uncertainty in Ecology : Implications for Remote Sensing and GIS Applications*, pages 238–257. Springer : New York, NY. 130

Sillard, P. (2000). *Les projections et référentiels cartographiques*. Ecole Nationale des Sciences Géographiques. 107, 361

Smithson, M. (1989). *Ignorance and Uncertainty : Emerging Paradigms*. Springer Verlag. 42

Snyder, J. (1987). *Map projections - a working manual*. United States Geological Survey, usgs professional paper 1395 édition. 106

STANAG (2002). Evaluation of land maps, aeronautical charts and digital topographic data. Rapport technique, OTAN. 59, 328, 329

TeleAtlas (2009). Multinet teleatlas. Rapport technique, TeleAtlas. 318

Tong, X. et Shi, W. (2010). Measuring positional error of circular curve features in Geographic information systems (GIS). *Computers & Geosciences*, 36:861–870. 140

Tong, X., Shi, W. et Liu, D. (2003). An error model of circular curve features in GIS. *In Proceedings of the 11th ACM international symposium on Advances in geographic information systems*, GIS'03, pages 141–146, New York, NY, USA. ACM. 140

Touya, G. (2011). *Le modèle CollaGen. Collaboration de processus automatiques pour la généralisation cartographique des paysages hétérogènes.* Thèse de doctorat, Université Paris-Est. 201, 205, 219, 220, 221, 224, 248, 364

Touya, G., Grosso, E., Mustière, S. et Duchêne, C. (2006). Automatisation de l'appariement de la bdcarto avec la bdtopo pays et de la généralisation de la bdcarto à partir de la bdtopo pays, en vue de l'intégration de ces bases. la bdcarto fait peau neuve. Rapport technique, Institut Géographique National. 39

Unwin, D. (1995). Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, 19:549–558. 50

Van Oort, P. (2005). *Spatial Data Quality : From Description To Application.* Thèse de doctorat, Wageningen University. 47, 48, 76, 369

Vauglin, F. (1997). *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires.* Thèse de doctorat, Université Marne-La-Vallée. 19, 46, 47, 48, 57, 58, 61, 64, 65, 162, 163, 360

Vauglin, F. (2002). A Practical Study on Precision and Resolution in Vector Geographical Databases. *In* Shi, W., Fisher, P. et Goodchild, M. F., éditeurs : *Spatial Data Quality.* Taylor & Francis. 139

Veregin, H. (2000). Quantifying positional error induced by line simplification. *International Journal of Geographical Information Science*, 14(2):113–130. 67

Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review*, 22:88–93. 114

Visvalingam, M. et Whyatt, J. D. (1991). The douglas-peucker algorithm for line simplification : re-evaluation through visualization. *Computer Graphics Forum*, 9(3):213–228. 139

Wall, K. et Danielsson, P. (1984). A fast sequential method for polygonal approximation of digitized curves. *Computer Vision, Graphics and Image Processing*, 28:220–227. 137, 138, 361

Walter, V. (2008). Automatic interpretation of vector databases with a raster-based algorithm. *In Proceedings of the 21st ISPRS Congress - Commission II*, Beijing, China. 220, 223, 224

Williams, M., Cornford, D., Bastin, L., Ingram et R., B. (2008). Uncertml : an xml schema for exchanging uncertainty. *In Proceedings of the 16th GIS Research UK Annual Conference (GISRUK 2008)*, Manchester Metropolitan University. 50

Wolf, P. R. et Ghilani, C. D. (1997). *Adjustment Computations : Statistics and Least Squares in Surveying and GIS (Wiley Series in Surveying and Boundary Control).* Wiley-Interscience. 62

Worboys, M. F. (1998a). Computation with imprecise geospatial data. *Computers, Environment and Urban Systems*, 22:85–106. 41

Worboys, M. F. (1998b). Imprecision in finite resolution spatial data. *Geoinformatica*, 2:257–280. 41

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353. 42, 43

Zhang, J. et Goodchild, M. F. (2002). *Uncertainty in Geographical Information*. CRC Press, 1 édition. 42

Table des figures

1.1 Une base de données géographiques vue à travers le filtre des spécifications . 20
1.2 Exemple de fiche de spécifications extraite de la BDTOPO (IGN, 2011c) . . 20
1.3 Extrait du schéma de données de la BDCARTO pour le thème routier . . . 21
1.4 La résolution spatiale de données maillées : représentation d'une route et d'un bâtiment, d'après Ruas (2002) 22
1.5 Une même courbe décrite par deux successions de points de granularités différentes, d'après Ruas (2002) 23
1.6 Un même réseau routier saisi selon les spécifications de la BDCARTO (en bleu) et de la BDTOPO (en rouge) 24
1.7 Les notions de précision et de biais représentées par une distribution statistique 27
1.8 Les notions d'exactitude, de précision et de biais représentées par une série de mesures sur une cible 28
1.9 Impact de la notion d'échelle sur l'exactitude d'une série de mesures 28
1.10 Classification des erreurs proposée par Hunter et Beard (1992), traduite par Devillers (2005) 29
1.11 Le concept de processus de production d'une base de données géographiques 30
1.12 Calcul des coordonnées d'un point P_3, à partir d'un réseau de référence de points P_1, et d'une densification locale de points P_2. La taille du cercle pointillé autour de chaque point illustre l'imprécision de localisation. 31
1.13 Calcul des coordonnées d'un récepteur par trilatération entre trois satellites, d'après Bosser (2011). 32
1.14 Exemple d'un nuage de points acquis par relevé Lidar sur les Champs Elysées (source IGN) 34
1.15 Classes d'objets de bâtiments et de routes de la BDTOPO issues de relevés photogrammétriques 35
1.16 Extraction de la géométrie des bâtiments (en rouge) à partir de la classification orientée-objet d'une image satellite Quickbird, d'après Liu *et al.* (2005) 36
1.17 Impact de la généralisation de bases de données, d'après Gaffuri (2008) . . . 37
1.18 Les deux natures d'imperfections affectant les bases de données géographiques 40
1.19 Taxonomie de l'incertitude de Fisher (2003) et Fisher *et al.* (2005) et des modélisations associées 42
1.20 Les grands types d'imperfections, d'après Olteanu (2008) 43
1.21 Les deux notions de la qualité des données géographiques 46
1.22 Distance euclidienne entre deux points 54
1.23 Distance de Hausdorff entre deux polylignes 54
1.24 Distances de Hausdorff et de Fréchet, d'après Badard et Lemarié (2002) . . 55

1.25 Principe de calcul de la distance moyenne . . . 56
1.26 Principe de calcul de la distance surfacique . . . 57
1.27 Calcul de la distance radiale entre deux objets surfaciques . . . 57
1.28 Calcul de la fonction angulaire de deux objets surfaciques, d'après Olteanu (2008) . . . 58
1.29 Le modèle "Epsilon-Band" . . . 61
1.30 Le modèle proposé par Goodchild et Hunter (1997) . . . 62
1.31 Le modèle d'imprécision en forme d'ellipse avec un échantillon de points . . . 62
1.32 Le modèle "Error Band" . . . 63
1.33 Le modèle proposé par Dutton (1992) . . . 63
1.34 Le modèle proposé par Shi *et al.* (1999) . . . 64
1.35 Les trois types de modélisation des zones de tolérance, d'après Franck et Navratil (2011) . . . 64
1.36 Le modèle GES proposé par Vauglin (1997) . . . 65
1.37 Le modèle proposé par Hunter et Goodchild (1996) . . . 66
1.38 Objets rigides et déformables, d'après Heuvelink *et al.* (2007) . . . 67

2.1 Imprécision géométrique et besoins utilisateurs . . . 76
2.2 Un objet mal positionné portant une erreur de mesure faible, et un objet bien positionné portant une erreur de mesure importante . . . 77
2.3 Un calcul d'itinéraire routier réalisé avec Google Map : Quelle est l'imprécision de la mesure ? . . . 78
2.4 Inadéquation d'une approche par simulations pour modéliser l'impact de la généralisation cartographique . . . 80
2.5 Comparaison des traces de trois récepteurs GPS (GlobalSat BT-338 en rouge, Emtac BTGPS S3 en bleu, et Royaltek RBT-2110 en vert), d'après Ray (2009) . . . 80
2.6 Impact de trois projections cartographiques sur les Etats-Unis, d'après Dana (1997) . . . 81
2.7 Représentation 2D et 2D5 d'une limite communale . . . 82
2.8 Impact de l'approximation polygonale sur la géométrie d'une route . . . 82
2.9 Principe général du processus d'évaluation . . . 86
2.10 Un processus d'évaluation sans données de référence . . . 87
2.11 La définition d'une forêt selon plusieurs pays, d'après Comber *et al.* (2005) . . . 88
2.12 Processus impliqués dans les erreurs de mesures géométriques . . . 89
2.13 Un processus d'évaluation détaillant chaque contribution à l'imprécision de la mesure . . . 91
2.14 Un processus d'évaluation produisant une estimation agrégée de l'imprécision de la mesure . . . 91
2.15 Sélection des impacts à modéliser par l'utilisateur . . . 93
2.16 Données externes et connaissances nécessaires au paramétrage du modèle . . . 94
2.17 Un modèle d'évaluation intégré dans une interface SIG . . . 95
2.18 Diagramme de classes de haut niveau du modèle d'estimation de l'imprécision des mesures géométriques . . . 97

3.1 Représentation exagérée du géoïde terrestre . . . 103
3.2 Un ellipsoïde défini par son demi-grand axe et son demi-petit axe . . . 104
3.3 Dimensions des ellipsoïdes Clarke 1880, Hayford 1909, et IAG-GRS80 . . . 104

3.4 Calcul de coordonnées cartésiennes dans un système de référence terrestre . 105
3.5 Principe de calcul de coordonnées géographiques 106
3.6 Projection cylindrique conforme de Marcator et indicateur de Tissot, d'après Sillard (2000) 107
3.7 Projections Lambert 93 et Coniques Conformes 9 zones 108
3.8 Principe de calcul de distances sur ellipsoïde entre deux points A et B . . . 109
3.9 Altération linéaire des projections Lambert I, II étendu, III, IV, et Lambert93 selon la latitude (source : IGN) 110
3.10 Principe d'estimation de l'impact de la projection cartographique 114
3.11 Projection du corroyage régulier en Lambert93 115
3.12 L'altération linéaire de la projection Lambert93 (positive en rouge et négative en bleu) 116
3.13 Suréchantillonnage de la limite départementale des Pyrénées-Atlantiques à partir de la grille d'altération 117
3.14 Recherche des quatre points les plus proches dans la grille d'altération . . . 118
3.15 Réseau routier primaire et limite administrative du département des Pyrénées-Atlantiques 119
3.16 Renseignement de l'altitude des tronçons de route de la BDTOPO en coordonnées tridimentionnelles 121
3.17 Renseignement de l'altitude des tronçons de route de la BDTOPO version 1, au format *shapefile* 122
3.18 Vue tridimensionnelle de deux polygones de la classe "Cimetières" de la BDTOPO (sur la commune d'Hendaye) 123
3.19 Enrichissement de la géométrie des objets évalués pour le calcul de mesures géométriques en 2D5 124
3.20 Etapes de transformation d'un modèle numérique de terrain d'un maillage carré vers un réseau de triangles réguliers 125
3.21 Principe de suréchantillonnage d'une géométrie linéaire par intersection avec un réseau de triangles réguliers 126
3.22 Principe de suréchantillonnage d'une géométrie surfacique par intersection avec un réseau de triangles réguliers 128
3.23 Représentation exagérée en trois dimensions de la géométrie enrichie d'une limite communale 128
3.24 Triangles entièrement inclus (en bleu) et portions de triangles incluses (en vert) dans le contour d'un objet surfacique 129
3.25 Différence entre un MNT (en noir) et un MNE (en rouge) 131
3.26 Les MNT et MNE utilisés pour l'expérimentation 132
3.27 Limite communale et réseau routier de la commune d'Herbeys 133
3.28 Erreurs de longueur du réseau routier de la commune d'Herbeys entre mesures en 2D et en 2D5 134
3.29 Erreurs de surface de la limite administrative de la commune d'Herbeys entre mesures en 2D et en 2D5 134
3.30 Illustration de l'algorithme de Wall et Danielsson (1984) 138
3.31 Une route courbe du monde réel représentée par une ligne polygonale dans une base de données, d'après Bonin (2002) 138
3.32 Impact de la granularité sur la représentation d'un virage par la BDTOPO (en rouge) et la BDCARTO (en bleu) 140

3.33 Une entité courbe du monde réel (en gris), représentée sous forme de ligne polygonale (en noir), sur laquelle est reconstruite une géométrie courbe réaliste (en rouge) . . . 140
3.34 Principe d'estimation de l'impact de l'approximation polygonale des courbes 141
3.35 Une courbe de Bézier cubique définie par les points de contrôle P_0, P_1, P_2 et P_3 . . . 142
3.36 Construction d'une courbe de Bézier cubique définie par les points de contrôle P_0, P_1, P_2 et P_3 selon la méthode de Casteljau . . . 143
3.37 Construction d'une interpolation par splines cubiques d'Hermite passant par les points de contrôle P_0, P_1, P_2 et P_3 selon les tangentes T_0, T_1, T_2 et T_3 . 143
3.38 Impact du paramètre de tension sur la reconstruction d'une courbe . . . 144
3.39 Hétérogénéité des distances inter-points d'un réseau routier . . . 145
3.40 Exemple de reconstruction de courbe irréaliste . . . 146
3.41 Reconstruction d'une courbe réaliste suite au suréchantillonnage de la polyligne . . . 146
3.42 Distribution des longueurs de segments des tronçons de route de la BDCARTO dans le département des Pyrénées-Atlantiques . . . 147
3.43 Classification de la sinuosité de portions de routes, d'après Plazanet (1996) . 148
3.44 Effet du lissage gaussien sur une polyligne pour différentes valeurs de σ . . . 149
3.45 Tableaux de séquences et d'identification des points d'inflexion d'une polyligne 150
3.46 Résultat de la détection des points d'inflexion sur une polyligne de la BDTOPO, suite à un lissage gaussien de paramètres $\sigma = 85$ m. . . . 151
3.47 Rapport base sur longueur . . . 151
3.48 Classification de la sinuosité de tronçons de route, selon la méthode de Mustière (2001) en *sinuosité nulle* (en bleu), *virages peu sinueux* (en vert), *sinuosité hétérogène* (en orange), et *épingles à cheveux* (en rouge) . . . 152
3.49 Extraction du réseau routier de la BDCARTO sur les communes de Aldudes, Banca, et Urupel (Pyrénées-Atlantiques) . . . 154
3.50 Distribution des distances inter-points du réseau routier expérimenté . . . 155
3.51 Différence de longueur en fonction de la tension . . . 156
3.52 Résultats des reconstructions de courbes avec un paramètre de tension de 0,1 à 0,9 . . . 156
3.53 Résultat de reconstruction de courbe avec un paramètre de tension de 0,7 . 157
3.54 Modélisation de l'impact des règles de représentation au sein du processus global de génération d'erreur de mesure . . . 159

4.1 Imprécision de localisation d'un point selon une probabilité suivant une loi normale en x et en y . . . 163
4.2 Principe de déplacement d'un point selon deux méthodes . . . 163
4.3 Une polyligne représentant un tronçon de route dans la BDCARTO, et 10 réalisations de son bruitage . . . 164
4.4 Effet du paramétrage de la fonction gaussienne sur le bruitage des sommets d'une polyligne . . . 164
4.5 Incohérences topologiques impliquées par un bruitage trop important . . . 165
4.6 Bruitage d'une polyligne avec corrélations (en tiretés), suivant un variogramme (à droite) déterminant la valeur de σ', d'après Bonin (2002) et Fouqué (1999) . . . 166
4.7 Une polyligne au caractère rectiligne (en rouge) et une simulation (en bleu) 166

4.8 Gammes d'échelles de saisie et de représentation de la BDTOPO et de la BDCARTO . . . 170
4.9 Distribution des distances inter-points à partir d'une extraction des tronçons de route de la BDTOPO . . . 171
4.10 Exemple de fonctions permettant de déterminer la pondération q du bruit gaussien en fonction de l'angularité α du point de la géométrie . . . 175
4.11 Bruitage d'une polyligne en prenant en compte l'angularité . . . 175
4.12 Imprécision de localisation relative au degré des nœuds . . . 176
4.13 Bruitage d'une polyligne en prenant en compte l'angularité et le degré des nœuds . . . 177
4.14 Effet du lissage gaussien sur des tronçons de route bruités de la BDCARTO 178
4.15 Principe de translation d'un point B en fonction des erreurs aux extrémités A et C d'une polyligne . . . 179
4.16 Exemple d'objets surfaciques représentant une partition de l'espace (les limites communales, en vert) et n'en représentant pas une (les surfaces d'eau, en bleu) . . . 180
4.17 Tronçons de route utilisés pour le test de saisie, représentés au 1 :10.000 et au 1 :25.000 . . . 182
4.18 Localisation des 10 points sélectionnés pour étudier l'imprécision de pointé des tronçons saisis . . . 183
4.19 Résultats de l'étude de l'imprécision de pointé (en mètres) à partir des 10 points sélectionnés, au 1 :10.000 (en bleu) et au 1 :25.000 (en rouge) . . . 184
4.20 Localisation du tronçon de route (en rouge) utilisé pour étudier l'imprécision des mesures de longueur . . . 185
4.21 Distribution des distances inter-points des routes saisies au 1 :10.000 . . . 187
4.22 Distribution des distances inter-points des routes saisies au 1 :25.000 . . . 187
4.23 Extraction du réseau routier de la BDCARTO sur les cantons d'Hendaye et de Saint-Jean-de-Luz . . . 191
4.24 Distribution des distances inter-points sur le réseau routier de la BDCARTO 192
4.25 Distribution des erreurs de longueur entre routes initiales et simulées . . . 193
4.26 Limites communales de la BDCARTO sur les cantons d'Hendaye et de Saint-Jean-de-Luz . . . 194
4.27 Distribution des distances inter-points sur les limites communales de la BDCARTO . . . 194
4.28 Extraction des nœuds des limites communales . . . 195
4.29 Distribution des erreurs de surface entre limites communales initiales et simulées . . . 196
4.30 Sur-estimation systématique du périmètre d'un objet surfacique simulé par rapport à sa surface. . . . 197
4.31 Les opérations de généralisation, d'après Mustière (2001) et Gaffuri (2008) . 200
4.32 Impact de la généralisation cartographique sur la représentation du Havre de Regnéville à trois échelles de représentation différentes . . . 201
4.33 La rocade de Rennes numérisée à partir de trois sources cartographiques . . 202
4.34 Impact du changement d'échelle sur la longueur d'objets généralisés, d'après João (1998) . . . 203
4.35 Impact du changement d'échelle sur la longueur d'objets routiers généralisés, d'après João (1998) . . . 203

4.36 Principe d'élaboration d'un modèle statistique (1) et application de ce modèle sur une base de données à évaluer (2) . . . 206
4.37 Deux routes voisines représentées sur la carte TOP100 (n°106) . . . 208
4.38 Algorithme d'estimation de l'échelle de représentation basé sur la superposition de symboles de routes voisines . . . 209
4.39 Surface intersectée entre symboles de routes voisines en fonction de l'échelle de représentation du réseau routier de données TOP100 . . . 211
4.40 Surface intersectée entre symboles de routes voisines en fonction de l'échelle de représentation du réseau routier de la carte Régionale . . . 211
4.41 Superposition de symboles routiers détectée sur les données de la Carte Régionale au 1 :150.000 . . . 212
4.42 Empâtement d'une ligne symbolisée, d'après Mustière (2001) . . . 212
4.43 Rapport entre la distance bord-axe de la route D et la largeur du demi-symbole L . . . 213
4.44 Nombre d'empâtements détectés en fonction de l'échelle de représentation sur les tronçons de route extraits des données TOP100 (en vert) et de la Carte Régionale (en bleu) . . . 214
4.45 Limite d'empâtement d'un virage extrait des données TOP100, en comparaison avec sa représentation cartographique . . . 215
4.46 Variation de la surface intersectée (en trait plein) et du nombre d'empâtements détectés (en pointillés) entre deux échelles successives pour les données TOP100 . . . 216
4.47 Détection conjointe de superpositions de symboles de routes et de zones d'empâtement à partir des données TOP100 . . . 216
4.48 Variation de la surface intersectée (en trait plein) et du nombre d'empâtements détectés (en pointillés) entre deux échelles successives pour les données de la Carte Régionale . . . 217
4.49 Détection conjointe de superpositions de symboles de routes et de zones d'empâtement sur les données de la Carte Régionale . . . 217
4.50 Principe de détéction des zones de montagne avec un critère de dénivelé, d'après Touya (2011) . . . 220
4.51 Conversion du MNE SRTM en courbes de niveau . . . 221
4.52 Zones de montagne détectées selon les deux formulations du critère de dénivelé 222
4.53 Résultat de la délimitation de la zone de montagne suite aux opérations de nettoyage et de lissage . . . 222
4.54 Résultat de la délimitation des zones urbaines et péri-urbaines par la méthode de Boffet (2000), d'après Touya (2011) . . . 224
4.55 Petites faces en milieu urbain et grandes faces en milieu rural . . . 225
4.56 Polygones à éliminer pour la détection des zones urbaines . . . 225
4.57 Résultats de la délimitation des zones urbaines sur le département des Pyrénées-Atlantique, à partir du réseau routier de la BDCARTO . . . 226
4.58 Résultats d'un appariement entre les tronçons de route de la BDTOPO (en trait plein) et de la BDCARTO (en tiretés) en trois classes : "Apparié" (en vert), "Incertain" (en orange), "Non apparié" (en rouge) . . . 229
4.59 Pré-traitements des polylignes pour une comparaison normalisée . . . 230
4.60 Comparaison de tronçons de routes avec une longueur de référence de 800 m, entre les données TOP100 (en bleu) et la BDTOPO (en rouge) . . . 231

4.61 Un même réseau routier représenté par la BDTOPO (en rouge), la BDCARTO (en bleu), les données TOP100 (en vert) et celles de la Carte Régionale (en orange) . . . 232
4.62 Délimitation du contexte géographique en zones de montagne (en bleu), urbaines (en rouge) et rurales (en vert) . . . 232
4.63 Erreur moyenne de longueur relative à la longueur du tronçon de référence, en zone de montagne (bleu), rurale (vert) et urbaine (rouge), pour les données BDCARTO (rond), TOP100 (triangle), et de la Carte Régionale (carré), d'après Girres (2011a) . . . 234
4.64 Comparaison de tronçons de route de longueurs hétérogènes, des données TOP100 (en bleu) et de la BDTOPO (en rouge) . . . 236
4.65 Distribution des erreurs de longueur sur le réseau routier de la BDCARTO localisé en zone de montagne . . . 237
4.66 Imprécision de l'erreur relative de longueur des tronçons de route de la BDCARTO en zones rurale (en vert) et de montagne (en bleu) . . . 238
4.67 Délimitation des zones de montagne (en bleu), urbaines (en rouge) et rurales (en vert) dans le département du Var . . . 240
4.68 Les dix itinéraires tirés aléatoirement sur le réseau routier des données TOP100 dans le département du Var . . . 241
4.69 L'itinéraire B extrait de la BDTOPO (en rouge) et des données TOP100 (en bleu) . . . 242
4.70 Les dix itinéraires tirés aléatoirement sur le réseau routier de la BDCARTO dans le département du Var . . . 244
4.71 Les opérations de simplification de bord, d'agrégation et de dilatation d'objets surfaciques . . . 247
4.72 Résultat d'un appariement de surfaces par un lien 1-n, avec l'algorithme de Bel Hadj Ali (2001) . . . 248
4.73 Fusion de polygones adjacents pour la comparaison de surfaces . . . 249
4.74 Localisation des 382 emprises communales utilisées pour le test de comparaisons de surfaces . . . 250
4.75 Relation entre les surfaces des communes homologues de la BDTOPO et des données TOP100 . . . 251
4.76 Distribution des erreurs de surface relatives des limites communales de la BDCARTO . . . 252
4.77 Relation entre l'erreur de surface relative et la surface évaluée sur les limites communales de la BDCARTO . . . 253
4.78 Distribution des erreurs de surface relatives des surfaces d'eau de la BDCARTO . . . 255
4.79 Distribution des erreurs de surface relatives des surfaces d'eau des données TOP100 . . . 256
4.80 Une surface d'eau représentée par la BDTOPO, la BDCARTO et les données TOP100 (du vert foncé au vert clair) . . . 256
4.81 Relation entre l'erreur de surface relative et la surface évaluée sur les surfaces d'eau de la BDCARTO . . . 257
4.82 Relation entre les surfaces d'eau de la BDTOPO et de la BDCARTO . . . 258
4.83 Modélisation de l'impact des processus de production au sein du processus global de génération d'erreur de mesure . . . 262

5.1 Principe général de l'approche par combinaison des estimations 266
5.2 Les 20 itinéraires au sein de leur contexte géographique (montagne en bleu, urbain en rouge, et rural en vert) 269
5.3 Exemple de l'itinéraire 13, représenté par la BDTOPO (en rouge) et les données TOP100 (en vert) 272
5.4 Erreurs relatives de référence et résiduelles estimées sur les itinéraires extraits de la BDCARTO 279
5.5 Erreurs relatives de référence et résiduelles estimées sur les itinéraires extraits des données TOP100 279
5.6 Exemple fictif d'application des arbres de décision à l'estimation de l'imprécision des mesures 285
5.7 Principe de fonctionnement d'un neurone formel 286
5.8 Principe général de l'approche par composition ordonnée de modèles 287
5.9 Calcul de la distance entre les points A et B en projection (à gauche), sur ellipsoïde (au centre) et en coordonnées cartésiennes (à droite) 288
5.10 Géométries simulées enrichies d'un tronçon de route de la BDCARTO . . . 289
5.11 Diagramme de classes du modèle d'estimation de l'imprécision des mesures géométriques 292
5.12 Diagramme de classes relatif à la modélisation des géométries dans GeOxygene 293
5.13 Diagramme de classes relatif à la modélisation de la géométrie des objets à évaluer 294
5.14 Diagramme de classes du modèle d'estimation de l'impact de la projection cartographique 296
5.15 Diagramme de classes du modèle d'estimation de l'impact de la non-prise en compte du terrain 297
5.16 Diagramme de classes du modèle d'estimation de l'impact de l'approximation polygonale des courbes 298
5.17 Diagramme de classes du modèle d'estimation de l'impact de l'erreur de pointé 299
5.18 Diagramme de classes du modèle de comparaison des mesures géométriques entre données généralisées et données de référence 300
5.19 Diagramme de classes du modèle d'estimation de l'impact de la généralisation cartographique 301
5.20 Les méthodes utilisées pour combiner les impacts sélectionnés 302
5.21 Diagramme de classes du modèle d'estimation globale par composition de modèles 303
5.22 Diagramme d'activité de haut niveau du modèle général d'estimation 305
5.23 Sous-diagramme d'activité correspondant à la sélection et au paramétrage des impacts à évaluer 306
5.24 Sous-diagramme d'activité relatif au paramétrage de l'impact de la projection cartographique 307
5.25 Sous-diagramme d'activité relatif au paramétrage de l'impact de la non-prise en compte du terrain 307
5.26 Sous-diagramme d'activité relatif au paramétrage de l'impact de l'approximation polygonale des courbes 307
5.27 Sous-diagramme d'activité relatif au paramétrage de l'impact de l'erreur de pointé 308
5.28 Sous-diagramme d'activité relatif au paramétrage de l'impact de la généralisation cartographique 308

5.29 Sous-diagramme d'activité relatif à la combinaison des impacts 309
5.30 Interface cartographique d'OpenJUMP 310
5.31 Menu permettant d'accéder au modèle EstIM 311
5.32 Interface générale du modèle EstIM 312
5.33 Interface d'estimation de l'échelle moyenne de saisie 313
5.34 Interface d'estimation de l'échelle de représentation 314
5.35 Interfaces de délimitation des zones de montagne (à gauche) et des zones urbaines (à droite) 314
5.36 Interface de communication des résultats par impact 315
5.37 Interface de communication de l'estimation globale 315
5.38 Résultat du calcul d'itinéraire réalisé avec *GoogleMap* 317
5.39 Procédures d'export et de conversion de l'itinéraire calculé pour son évaluation avec le modèle EstIM 318
5.40 Résultats de l'évaluation de l'itinéraire par le modèle EstIM 319
6.1 Relations possibles entre le modèle d'estimation et les métadonnées 324
7.2 Exemple d'une instance "Batiment1" de la classe "BATIMENT" 332
7.3 Exemple de relation d'association entre la classe "BATIMENT" et la classe "ACTIVITE" 333
7.4 Exemple de relation d'agrégation entre la classe "BATIMENT" et les classes "TOIT" et "MUR" 333
7.5 Exemple de relation de généralisation par une super-classe "BATIMENT" . 334
7.6 Interface d'appariement et de comparaison des mesures géométriques entre objets homologues 338
7.7 Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone rurale 339
7.8 Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone rurale 339
7.9 Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone rurale 340
7.10 Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone urbaine 340
7.11 Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone urbaine 341
7.12 Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone urbaine 341
7.13 Distribution des erreurs de longueur relatives des tronçons de route de la BDCARTO en zone de montagne 342
7.14 Distribution des erreurs de longueur relatives des tronçons de route des données TOP100 en zone de montagne 342
7.15 Distribution des erreurs de longueur relatives des tronçons de route des données de la Carte Régionale en zone de montagne 343

Liste des tableaux

1.1 Volumetrie du réseau routier du département des Pyrénées-Atlantiques représenté par la BDTOPO et la BDCARTO . . . 24
1.2 Impact des sources d'erreurs sur la précision d'un positionnement absolu et différentiel, d'après Bosser (2011) . . . 33
1.3 Eléments d'évaluation de la qualité des données géographiques proposés dans cinq sources, d'après Van Oort (2005) . . . 47
1.4 Eléments et sous-éléments d'évaluation de la qualité de la norme ISO 19113 48
1.5 Mesures d'exactitude absolue et relative proposées dans la spécification technique ISO 19138 pour l'élément "Exactitude de Position" . . . 59

3.1 Différences de longueurs entre la polyligne initiale et les courbes reconstruites avec ou sans suréchantillonnage . . . 155

4.1 Distances inter-points médianes de trois classes d'objets de la BDTOPO et la BDCARTO . . . 172
4.2 Imprécision de pointé (en mètres) à partir des 10 points sélectionnés . . . 184
4.3 Imprécision des mesures de longueur (en mètres) au 1 :10.000 et au 1 :25.000 186
4.4 Pourcentages d'estimations de longueurs correctes pour les routes saisies au 1 :10.000 . . . 189
4.5 Pourcentages d'estimations de longueurs correctes pour les routes saisies au 1 :25.000 . . . 189
4.6 Spécifications simplifiées des largeurs de symboles routiers pour la carte TOP100 et la Carte Régionale . . . 209
4.7 Erreur de longueur moyenne absolue et relative en fonction de la longueur de tronçon pour les données de la Carte Régionale (CR), TOP100 (T100) et de la BDCARTO (BDC), d'après Girres (2011a) . . . 233
4.8 Résultat des comparaisons de longueurs sur l'ensemble du réseau routier des Pyrénées-Atlantiques . . . 236
4.9 Pourcentage d'erreurs de longueur relatives comprises dans un intervalle de σ, 2σ, et 3σ autour de l'erreur moyenne relative . . . 238
4.10 Contexte géographique des itinéraires extraits des données TOP100 . . . 242
4.11 Erreurs de longueur des itinéraires extraits des données TOP100 . . . 242
4.12 Longueurs de référence estimées par application du modèle statistique sur les itinéraires extraits des données TOP100 . . . 243
4.13 Structure des itinéraires extraits de la BDCARTO selon le contexte géographique . . . 244
4.14 Erreurs de longueur entre les itinéraires extraits de la BDCARTO et leurs homologues de la BDTOPO . . . 245

4.15 Longueurs de référence estimées par application du modèle statistique sur les itinéraires extraits de la BDCARTO 245
4.16 Résultat des comparaisons de surfaces sur les limites communales 251
4.17 Pourcentage d'erreurs de surface relatives comprises dans un intervalle de σ, 2σ, et 3σ autour de l'erreur moyenne 252
4.18 Résultats des comparaisons de surfaces sur les surfaces d'eau 255

5.1 Structure des itinéraires extraits de la BDCARTO et des données TOP100 selon le contexte géographique 270
5.2 Erreurs de longueur des vingt itinéraires comparés 271
5.3 Estimations absolues et relatives des impacts sur les itinéraires extraits de la BDCARTO 274
5.4 Estimations absolues et relatives des impacts sur les itinéraires extraits des données TOP100 275
5.5 Combinaison des impacts sur les itinéraires de la BDCARTO 278
5.6 Combinaison des impacts sur les itinéraires extraits des données TOP100 . . 278
5.7 Erreurs moyennes quadratiques de référence et résiduelles estimées 280
7.8 Profil général de l'échantillon 336
7.9 Résultats du test de digitalisation au 1 :10.000 337
7.10 Résultats du test de digitalisation au 1 :25.000 337

Publications

GIRRES J.-F. et RUAS A. (2009). Conception of a model to evaluate geometrical imprecisions from under-qualified vector data for decision making. *In Proceedings of the 6th International Symposium on Spatial Data Quality (ISSDQ'09).* 5-8 Juillet, St John's (Canada)

TOUYA G. et GIRRES J.-F. (2009). Première évaluation de la qualité des données libres d'OpenStreetMap en France. *In Colloque International de Géomatique et d'Analyse Spatiale SAGEO'09.* 25-27 Novembre, Paris (France)

GIRRES J-F. (2010). Evaluation et communication de l'impact de la précision géométrique, Proposition d'une approche basée sur la connaissance des processus de production. *In Rencontres SIG-La Lettre.* 4-6 mai, ENSG, Marne la Vallée (France)

GIRRES J.-F. et RUAS A. (2010). Estimation of Imprecision in Length and Area Computation in Vector Databases Including Production Processes Description. *In Proceedings of the 14th International Symposium on Spatial Data Handling (SDH'10), Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science.* 26-28 Mai, Hong Kong (Chine)

GIRRES J.-F. et JULIEN P. (2010). Estimation of digitizing and polygonal approximation errors in the computation of length in vector databases. *In Proceedings of the 9th International Symposium on Spatial Accuracy Assessement in Natural Resources and Environmental Sciences (ACCURACY'10).* 20-23 Juillet, Leicester (UK)

GIRRES, J.-F. et TOUYA G. (2010). Quality assessment of the french OpenStreetMap dataset, *Transactions in GIS.* 14(4). 435-459

GIRRES J.-F. (2011). Conception d'un modèle d'évaluation des imprécisions géométriques dans les bases de données géographiques vectorielles. *In Journées de la Recherche de l'IGN.* 11 Mars, Saint-Mandé (France)

GIRRES J.-F. (2011). An evaluation of the impact of cartographic generalisation on length measurement computed from linear vector databases. *In Proceedings of the 25th International Cartographic Conference (ICC'11).* 3-8 Juillet, Paris (France)

GIRRES J.-F. (2011). A model to estimate length measurements uncertainty in vector databases. *In Proceedings of the 7th International Symposium on Spatial Data Quality (ISSDQ'11).* 12-14 Octobre, Coimbra (Portugal)