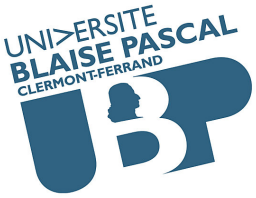
Initialize and Calibrate a Dynamic Stochastic Microsimulation Model : application to the SimVillages Model

Maxime Lenormand



N° d'ordre : D.U. 2315
EDSPIC : 597

UNIVERSITÉ BLAISE PASCAL - CLERMONT II

ÉCOLE DOCTORALE
SCIENCES POUR L'INGÉNIEUR DE CLERMONT-FERRAND

THÈSE

Présentée par

MAXIME LENORMAND

Master Statistiques et Traitement des Données

pour obtenir le grade de

DOCTEUR D'UNIVERSITÉ

SPÉCIALITÉ : Informatique

Initialize and Calibrate a Dynamic Stochastic Microsimulation Model: Application to the *SimVillages* Model

Soutenue publiquement le 12 décembre 2012 devant le jury composé de :

Président :	**Laurent SERLET** Professeur des universités, Université Blaise-Pascal, Clermont-Ferrand
Rapporteurs :	**Jean-Pierre NADAL** Directeur de recherche, CNRS, Paris
	Philippe TOINT Professeur des universités, Université de Namur (FUNDP), Namur
Examinateurs :	**Marc BARTHELEMY** Chercheur, Institut de Physique Theorique, Gif-sur-Yvette
	Jean-Michel MARIN Professeur des universités, Université Montpellier 2, Montpellier
Directeur de thèse :	**Guillaume DEFFUANT** Directeur de recherche, Irstea, Clermont-Ferrand
Invitée :	**Sylvie HUET** Ingénieur d'études, Irstea, Clermont-Ferrand

Acknowledgement

Je tiens tout d'abord à remercier Sylvie Huet et Guillaume Deffuant pour avoir su me guider tout au long de ces trois années. Sylvie, merci pour tes conseils, ta patience, ta disponibilité et aussi pour ton pif incroyable capable de repérer (ou de tomber malencontreusement sur) le moindre bug présent dans un code ou un article. Guillaume, merci de m'avoir accordé un peu de ton temps si précieux, merci pour l'aide que tu m'as apportée, la confiance que tu m'as donnée et pour les longues heures que tu as passées sur nos articles et sur mon manuscrit. Je vous remercie tous les deux d'avoir trouvé le juste équilibre entre encadrement et autonomie. Trouvez dans ces quelques lignes le témoignage de ma profonde gratitude. Merci pour tout.

Mes remerciements vont également à Floriana Gargiulo avec qui c'est un réel plaisir de travailler, j'espère que nous continuerons à collaborer ensemble dans les années futures.

Je tiens à remercier mon "esclave", comme il se qualifie lui même, l'informatismique Nicolas Dumoulin sans qui ma thèse aurait duré trente années au lieu de trois. Un grand merci à toi Nico !!

Merci également aux membres de mon comité de thèse Timoteo Carletti, Didier Blanchet, Hervé Monod, Fabien Campillo et plus particulièrement merci à Franck Jabot pour sa contribution à ma thèse.

Mes remerciements vont également à Jean-Pierre Nadal et Philippe Toint qui ont accepté d'être rapporteurs de cette thèse. Je remercie aussi les autres membres du jury Marc Barthélémy, Jean-Michel Marin et Laurent Serlet.

Merci également à tous les membres du Lisc. En particulier, merci à Clairus, Wei Wei, Bruni et JD.

Mes remerciements suivants vont à l'ensemble des membres du projet PRIMA. En particulier je tiens à remercier Olivier Aznar, Eliska Kozáková, Omar Baqueiro Espinosa, Olivier Barreteau, Mario Njavro et Marian Raley.

M M M

Résumé Etendu

L'objectif de ce travail est de développer des outils statistiques et des modèles informatiques pour initialiser et calibrer les modèles de microsimulation dynamique stochastique. Nous avons développé ces outils et modèles dans le cadre de l'élaboration du modèle *SimVillages*, certains développements sont très spécifiques à ce modèle, d'autres plus génériques.

L'hypothèse à la base de la microsimulation est que se placer au niveau de l'individu donne plus de chance de comprendre ce qui se passe à un niveau plus agrégé. C'est dans cette optique que l'on utilise la microsimulation dynamique. L'idée est de créer une société virtuelle où l'individu est l'entité à la base du système. Cette société virtuelle devra être statistiquement semblable à la société "réelle" sur les indicateurs qui nous intéressent, définis en fonction des objectifs du modèle. On fait ensuite évoluer cette société dans le temps en essayant de reproduire les faits passés pour comprendre comment ils se sont produit et ainsi tenter d'anticiper l'avenir. Mais pour créer un tel modèle informatique plusieurs questions se posent. Tout d'abord comment créer la population synthétique de base du modèle en absence de donnée détaillée sur la population à reproduire ? Comment extraire l'information des données pour construire la dynamique du modèle ? Si il existe des paramètres du modèle inconnus, comment estimer leurs valeurs ? Ce travail est consacré à ces questions.

Dans ce résumé étendu, je présente, dans un permier temps, le modèle de microsimulation dynamique *SimVillages* servant de cadre applicatif aux travaux de la thèse dans le but de donner aux lecteurs une idée plus précise du contexte dans lequel s'inscrivent les développements méthodologiques présentés. Dans un second temps, je présente un résumé des chapitres.

Le modèle de microsimulation *SimVillages*

Le modèle de microsimulation dynamique stochastique *SimVillages* a été développé durant le projet Européen PRIMA[1]. Son objectif est de permettre de mieux comprendre les différences d'évolution des municipalités rurales. Dans ce modèle, on fait l'hypothèse que l'évolution de ces communes dépend, d'une part, des interactions entre municipalités à travers le navettage et la consommation de service et d'autre part du nombre d'emplois dans les différents secteurs d'activité (fixé de manière exogène à l'aide de scenarios) et d'emplois de services de proximité (supposés dépendant des caractéristiques de la municipalité).

Le modèle *SimVillages* appartient à la famille des modèles de microsimulation. Les origines de l'approche par microsimulation remontent à la fin des années cinquante (Orcutt, 1957). Elle est la première approche à avoir pris en compte le niveau individuel dans la modélisation des systèmes complexes mais elle fait maintenant partie d'une

[1] PRototypical policy Impacts on Multifunctional Activities in rural municipalities - EU 7th Framework Research Programme ; 2008-2011 ; https://prima.cemagref.fr/the-project

famille plus large de modèle : les modèles individus-centré. Cette famille de modèles regroupe la microsimulation, la théorie des jeux, les automates cellulaires, les simulations orientées-objet et les simulations multi-agents (Amblard, 2003). Les modèles de microsimulation modélisent à un niveau microscopique (à l'échelle de l'individu) un système complexe, ils sont dynamiques lorsqu'ils évoluent dans le temps et stochastiques lorsqu'ils sont composés d'une part d'aléa rendant chaque résultat du modèle "unique". L'intérêt de ce type de modèle est la flexibilité des résultats. En effet, le fait de travailler à une échelle très fine permet d'obtenir des résultats à plusieurs niveaux d'agrégation. Cependant, la microsimulation se voit opposer plusieurs critiques : le volume de données requis, le temps de calcul et la stochasticité. Depuis la première vision d'Orcutt avec *DYNASIM* (Orcutt et al., 1976) de nombreux modèles de microsimulation dynamique ont été proposés tels que *DESTINIE* (INSEE, 1999) ou encore *LifePaths* (Statistics Canada, 2004). Ces modèles permettent d'analyser l'évolution de systèmes complexes en prenant en compte une hétérogénéité dérivée des observations des individus et de leurs interactions. Le modèle *SimVillages* est stochastique incluant des objets hétérogènes (individus, ménages, municipalités, emplois, logements,...) et ses propriétés ne peuvent pas être dérivées analytiquement, nous avons besoin de réaliser un grand nombre de simulations pour comprendre son fonctionnement et ajuster les valeurs de paramètres inconnus pour obtenir une bonne adéquation entre données observées et données simulées.

Le modèle *SimVillages* est un système dynamique à temps discret $X_{t+1} = \mathscr{M}(\theta, \gamma, X_t)$ où $X_t \in \mathbb{R}^n$ est l'état du système, $\gamma = (\gamma_1, ..., \gamma_m)$ les paramètres fixés du modèle et $\theta = (\theta_1, ..., \theta_p)$ les paramètres inconnus du modèle. On observe des trajectoires de ce système dynamique, à partir de conditions intitiales X_0 et pendant un certain nombre de pas de temps T. Dans le modèle *SimVillages* un pas de temps équivaut à un an. Nous pouvons observer sur la Figure 1 que le modèle commence en 1990 et que pour le confronter à la réalité nous disposons de deux dates de recensement, 1999 et 2006. Nous pouvons aussi observer que pour des valeurs fixées de X_0, γ et θ chaque éxecution du modèle donne des trajectoires différentes à cause de la stochasticité.

Il existe deux catégories de paramètres, les paramètres fixés du modèle γ et les paramètres inconnus du modèle θ. Les paramètres fixés du modèle sont à configurer par l'utilisateur, leur valeur est dérivée de valeurs et de distributions de probabilité extraites des données observées à l'aide de méthodes statistiques et de traitement de données. Elles peuvent aussi prendre la forme de scenario intervenant de manière exogène dans la simulation. L'état initial fait aussi partie des paramètres fixés du modèle, il est représenté par une population synthétique construite à partir des données observées, à l'échelle de la région considérée. Chaque individu de cette population est caractérisé par :

- un ménage dont il fait partie, d'une certaine taille (de 1 à 6 ou plus individus) et d'un certain type (personne seule, famille monoparentale, couple avec enfant(s), couple sans enfant(s) et autre ménage),

- un statut au regard de son ménage (chef de famille, partenaire ou enfant),

- une situation au regard de l'emploi (employé, sans-emploi, retraité, inactif ou étudiant),

- une catégorie socio-professionnelle s'il est actif (agriculteur, artisan, profession intermédiaire, cadre, employé et ouvrier),

- un lieu de travail s'il est actif occupé (dans une commune de la région ou à l'extérieur de la région) et un secteur d'activité (agriculture, industriel ou service).

Une représentation des entités composant l'état initial du modèle est proposée Figure 2. Il est important que cet état initial soit statistiquement le plus proche possible de la population observée car il est le point de départ pour la calibration. En effet, l'état initial a un impact sur les évolutions futures du modèles.

Les paramètres inconnus du modèle sont ceux que nous n'avons pas pu directement extraire des données. Nous pouvons observer sur la Figure 3 que les paramètres inconnus du modèle sont extraits des données via une procédure de calibration tandis que les paramètres fixés du modèle sont directement extraits des données. La calibration du modèle *SimVillages* ne peut se faire analytiquement. Pour calibrer le modèle, nous faisons donc varier les paramètres pour minimiser une fonction cible, distance entre des statistiques construites à partir des données observées et des données simulées (population moyenne...). Nous devons pour cela parcourir efficacement l'espace des paramètres afin de trouver le ou les jeux de valeurs de paramètres minimisant la cible. Cela nécessite un grand nombre de simulations du modèle. Par exemple, sur la Figure 1, le but est de trouver des valeurs de θ qui ont au moins une trajectoire "proche" des données observées (représentées par les points verts).

Du point de vue de la dynamique du modèle, à chaque pas de temps, la population des communes évolue, les individus font des choix de vie, d'étude, de carrière, d'union, peuvent avoir des enfants, divorcer, migrer et mourir. Le modèle prend en compte, de manière endogène, les migrations inter-communales, les créations ou les suppressions d'emplois dans les services de proximité en fonction du nombre d'habitant. En plus de ces évolutions endogènes on introduit des scenarios représentant les décisions politiques prises au niveau régional telles que, par exemple, l'implantation d'une entreprise sur une commune. Ces scenarios modifient de manière exogène l'évolution des communes.

La région d'étude modélisée avec le modèle *SimVillages* est le département français du Cantal qui possédait 158 723 habitants répartis en 260 communes en 1990. Le modèle a pour point de départ 1990 et l'estimation de la distribution de valeurs des paramètres a été effectuée en deux points dans le temps, 1999 et 2006 (années correspondant au recensement de la population effectué par l'INSEE). Une simulation sur un ordinateur de bureau prend environ une minute. Une description complète du modèle est détaillée dans Huet et al. (2012a) et sa paramétrisation est détaillée dans Huet et al. (2012b) (disponible en Annexe A).

Résumé des chapitres

Ce travail de thèse se divise en quatre chapitres. Les deux premiers chapitres portent sur l'initialisation du modèle *SimVillages* avec la création d'une population synthétique. Le troisième chapitre concerne un modèle statistique permettant d'estimer le nombre d'emplois dans les services de proximité. Le quatrième chapitre présente une méthode de calcul bayésien approché permettant d'estimer la distribution des valeurs des paramètres inconnus du modèle.

Dans Lenormand and Deffuant (2012), présenté dans le Chapitre 1, nous avons tout d'abord implémenté l'algorithme proposé par Gargiulo et al. (2010) pour créer une population synthétique de l'Auvergne en 1990. Ensuite, nous validons cette population et nous comparons l'algorithme utilisé avec la méthode Iterative Proportional Updating (IPU) proposé par Ye et al. (2009). L'intérêt de l'algorithme proposé dans Gargiulo et al. (2010) est qu'il n'utilise que des données agrégées palliant ainsi l'absence d'un échantillon représentatif de la population. Nous montrons dans le premier chapitre que cet algorithme est plus rapide et qu'il donne de meilleurs résultats que l'autre algorithme utilisant un échantillon. En contre partie il nécessite plus de temps dans la préparation des données.

Pour finaliser la population synthétique il a fallu assigner à chaque individu actif occupé de cette population un lieu de travail lorsqu'il travaillait à l'extérieur de sa commune de résidence. Le réseau formé par les interactions entre communes pour les déplacements domicile-travail s'appelle un réseau de navettage. Les données détaillées étant indisponible en 1990 il a fallu développer un algorithme de génération de réseaux de navettage permettant de simuler un réseau à partir de données agrégées. Ce modèle a été proposé dans Gargiulo et al. (2012) (disponible en Annexe B), cet algorithme construit le réseau progressivement, en attribuant aux navetteurs, un par un, un lieu de travail avec une probabilité d'accepter ce lieu de travail qui augmente avec l'offre d'emploi de ce lieu de travail et diminue avec la distance entre la commune de résidence et la commune de travail candidate. Ce modèle a ensuite été adapté à 34 régions de France (Lenormand et al., 2012b). Dans cet article, disponible en Annexe C, une généralisation du modèle de base est proposée en incluant l'extérieur de la région (possibilité pour les navetteurs de travailler hors de la région d'étude) et en comparant plusieurs fonctions de décisions pour modéliser l'effet de la distance (puissance et exponentielle). Dans Lenormand et al. (2012c), présenté dans le Chapitre 2, nous proposons une loi permettant d'estimer le seul paramètre du modèle en fonction des caractéristiques de la région étudiée, cette loi a été testée et validée sur 80 régions d'Europe et d'Amérique.

Dans le Chapitre 3 nous présentons un modèle statistique permettant d'estimer le nombre d'emplois dans les services de proximité d'une commune en fonction de ses caractéristiques. Dans un premier temps, nous avons essayé d'estimer, pour une commune, la présence ou l'absence de service de proximité mais aussi le nombre d'emplois dans ces différents services en fonction des caractéristiques de la commune (voir Annexe D). Malgré des résultats satisfaisants, ce travail était trop compliqué à mettre en œuvre dans le modèle *SimVillages* car il était difficile de sélectionner les services destinés à la population sachant que certains services ne servent qu'en partie à la popu-

lation locale. Dans Lenormand et al. (2012a), présenté dans le troisième chapitre de la thèse, nous proposons une méthode permettant d'estimer, pour une commune donnée, le nombre d'emplois dans les services de proximité en fonction du nombre d'habitants et de son voisinage en terme de service.

Pour estimer la distribution des valeurs des paramètres inconnus du modèle, nous proposons dans Lenormand et al. (2012d), présenté dans le Chapitre 4, un algorithme de calcul Bayésien approché (ABC) par échantillonage préférentiel. Les méthodes d'échantillonage préférentiel appliquées à l'ABC sont dérivées de méthodes d'échantillonnage classique et elles sont considérées comme étant les plus efficaces en termes de temps de calcul parmi les méthodes ABC. Nous étudions les paramètres de notre algorithme et nous l'avons comparé à trois algorithmes concurrents dans la littérature. Nous montrons qu'avec n'importe quelle paramétrisation de notre algorithme, nous prenons de 2 à 8 fois moins de simulations pour atteindre au moins la même qualité de résultats que les trois autres algorithmes.

Mots-clés : Microsimulation, Modèle Complexe, Modèle Individus Centré, Modèle Stochastique, Calibration, Initialisation, Population Synthétique, Iterative Proportional Updating, Modèle de Réseaux de Navettage, Modèle de Déplacement, Loi de Gravité, Mobilité Humaine, Réseau Spatial, Besoin Minimal, Service de Proximité, Régression Quantile, Municipalité Rurale, Calcul Bayésien Approché, Population Monte Carlo, Sequential Monte Carlo.

Abstract

The purpose of this thesis is to develop statistical tools to initialize and to calibrate dynamic stochastic microsimulation models, starting from their application to the *SimVillages* model (developed within the European PRIMA project). This model includes demographic and economic dynamics applied to the population of a set of rural municipalities. Each individual, represented explicitly in a household living in a municipality, possibly working in another, has its own life trajectory. Thus, model includes rules for the choice of study, career, marriage, birth children, divorce, migration, and death.

We developed, implemented and tested the following models:

- a model to generate a synthetic population from aggregate data, where each individual lives in a household in a municipality and has a status with regard to employment. The synthetic population is the initial state of the model.
- a model to simulate a table of origin-destination commuting from aggregate data in order to assign a place of work for each individual working outside his municipality of residence.
- a sub-model to estimate the number of jobs in local services in a given municipality in terms of its number of inhabitants and its neighbors in terms of service.
- a method to calibrate the unknown *SimVillages* model parameters in order to satisfy a set of criteria. This method is based on a new Approximate Bayesian Computation algorithm using importance sampling. When applied to a toy example and to the *SimVillages* model, our algorithm is 2 to 8 times faster than the three main sequential ABC algorithms currently available.

Keywords: Microsimulation, Complex Model, Individual Based Models, Stochastic Models, Calibration, Initialisation, Synthetic Population, Sample-Free, Iterative Proportional Updating, Network Generation Models, Commuting Patterns, Commuting Networks, Gravity Law, Human Mobility, Spatial Networks, Minimum Requirement, Proximity Service Jobs, Quantile Regression, Rural Municipality, Approximate Bayesian Computation, Population Monte Carlo, Sequential Monte Carlo.

Preamble

This research has been motivated and partly funded by the European project PRIMA (PRototypical policy Impact on Multifonctional Activities in rural municipalities collaborative project, European Union 7th Framework Programme (ENV 2007-1)). This project aimed at developing a method for scaling down the analysis of policy impacts on multifunctional land uses and on the economic activities. It developed a microsimulation model, called *SimVillages*, designed and validated at municipality level, using input from stakeholders. The model address the structural evolution of the populations (appearance, disappearance and change of agents) depending on the local conditions for applying the structural policies on a set of municipality case studies.

This PhD thesis was carried out between January 2010 and December 2012 in the Laboratory of Engineering for Complex System (LISC) in National Research Institute of Science and Technology for Environment and Agriculture (IRSTEA) located in Clermont-Ferrand. The LISC develops individual-based models to study the complexity of social or eco-system dynamics and new methods for assessing the viability or resilience of such systems. In the PRIMA project, the LISC has developed the *SimVillages* model for the Auvergne case study, and this model has then been adapted to other case studies in the UK and Germany.

This PhD thesis was supervised by Guillaume DEFFUANT (Head of LISC) and Sylvie HUET (Engineer).

Each chapter of this PhD thesis is a paper submitted or accepted in a peer reviewed international journal.

This PhD thesis has been funded by the Auvergne region.

Contents

List of Figures

List of Tables

List of Algorithms

Introduction

Contents

Overview

This work aims to develop statistical tools and models to initialize and to calibrate a dynamic stochastic microsimulation model called *SimVillages*. Some of these tools and models are very specific to the *SimVillages* model, while others are more generic.

The microsimulation assumes that by considering the smallest scale brings deeper understanding of the social processes. The idea is to simulate a virtual social system where the virtual simplified individuals evolve and interact. These virtual individuals should be defined with attributes that are statistically similar to the "real" one for the indicators of interest; these indicators being defined in terms of the model objectives. Then, when running this virtual system over time it should replicate past events. Analysing how the model reproduces past events, one can get some assessment of its capacity to anticipate future trends. But developing such an informatic model requires to answer several questions. How to generate a synthetic population without detailed data? If the population is organised in several spatial entities, like municipalities, how to define the relationship between them? How to extract information from data to parameterize the model? If there are unknown model parameters, how to estimate their value? These are the main questions that we address in this work.

In this introduction, I first present the *SimVillages* microsimulation model which motivated the statistical tools developed during my PhD in order to give the reader a more precise idea of the context of the presented methodological developments. Then, I present a detailed outline of the thesis.

The *SimVillages* microsimulation model

The model which motivates the statistical methods developed during this PhD is the dynamic stochastic microsimulation model, *SimVillages*, elaborated within the European PRIMA[1] project. This model couples demographic and economic dynamics ap-

[1] PRototypical policy Impacts on Multifunctional Activities in rural municipalities - EU 7th Framework Research Programme; 2008-2011; https://prima.cemagref.fr/the-project

plied to a population of individuals living in a set of rural municipalities. The dynamics depend, on the one hand, on the spatial interactions between municipalities through commuting flows and services, and on the other hand, on the number of jobs in various activity sectors (supposed exogenously defined by scenarios) and on the jobs in proximity services (supposed dependent on the size of the local population).

The *SimVillages* model belongs to the family of microsimulation models. The origins of the microsimulation approach date back to the late fifties (Orcutt, 1957). It was the first approach taking into account the individual level in the modeling of complex systems, but it is now part of a larger family of model: the individual-based models. This family of models includes the microsimulation, game theory, cellular automata, simulations object-oriented and multi-agent simulations (Amblard, 2003). The microsimulation models represent explicitly each individual of the considered population. They are dynamic when they evolve over time and they are stochastic when they include some random processes making each model run "different". The advantage of this type of model is to provide results at different levels of aggregation. However, microsimulation has several drawbacks: the amount of data required, the computation time, and the stochasticity. Since the first vision of Orcutt with *DYNASIM* (Orcutt et al., 1976) many models of dynamic microsimulation have been proposed such as *DESTINIE* (INSEE, 1999) or *LifePaths* (Statistics Canada 2004). The *SimVillages* model is stochastic and includes several types of dynamics, which make it impossible to derive its properties analytically. Therefore, it is necessary to perform numerous simulations in order to observe its properties. Similarly when calibrating the model, i.e. determining the values of some parameters in order to minimise some error criterion, a systematic exploration of the parameter space is required, leading to a large number of simulations.

SimVillages is a discrete-time dynamical system $X_{t+1} = \mathscr{M}(\theta, \gamma, X_t)$ where $X_t \in \mathbb{R}^n$ is the state of the system, $\gamma = (\gamma_1, ..., \gamma_m)$ the fixed parameters and $\theta = (\theta_1, ..., \theta_p)$ the unknown parameters. We observe the trajectories of the dynamical system from the initial state X_0 and for a number of time steps T. In the *SimVillages* model a time step is set to one year. As we can observe on Figure 1, we start the *SimVillages* model in 1990 and compare to census data of 1999 and 2006. Because of the stochasticity, for fixed values of X_0, γ and θ, each model run gives different trajectories.

There are two types of model parameters - the fixed parameters γ and the unknown parameters θ. The fixed parameters are set by the user or their values are derived from observed data using statistical methods and data analysis. They can also be part of scenarios determined exogenously. The model's initial state can also be considered as a fixed parameter; it is represented by a synthetic population fixed in time and built with observed data. The model is initialized with a synthetic population representing a set of municipalities. Each individual in the population is characterized by:

- a household (to which he belongs) of a certain size (from one to six or more people) and a certain type (single person, single parents, couples with child(ren) located in a municipality of the region,

- a family status (head of household, partner or child),

- a position about employment (employed, unemployed, retired, inactive or student),
- a socio-professional category (farmer, craftsman, intermediate profession, executive, employee or worker) if he is active,
- a place of work (in a municipality of the region or outside of the region) and an activity sector (primary, secondary and tertiary) if he is occupied.

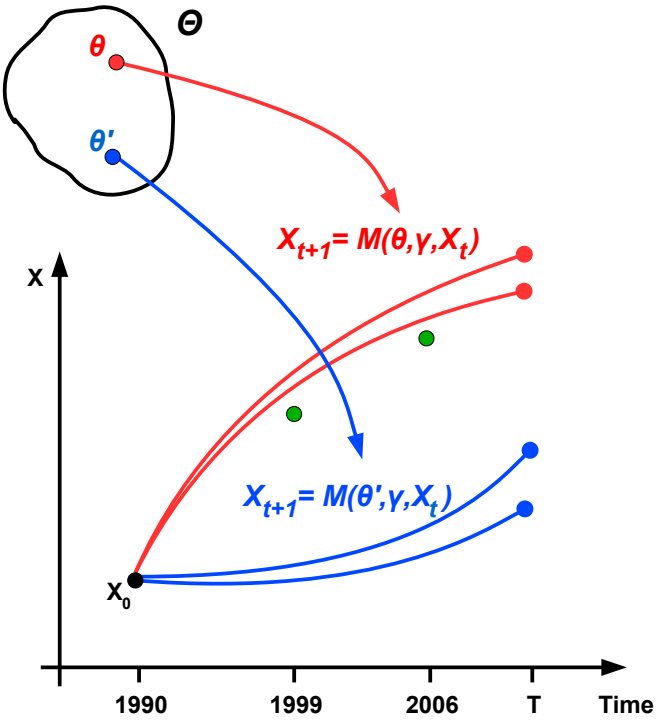


Figure 1: *SimVillages* model schematic representation $X_{t+1} = \mathscr{M}(\theta, \gamma, X_t)$. The trajectories represent four runs of the *SimVillages* model from the initial state X_0 and for a number of time steps T. In red, two trajectories obtained with parameter value θ. In blue, two trajectories obtained with parameter value θ'. The green points represent the observed value in 1999 and 2006.

The main components of the *SimVillages* model are presented in Figure 2. It is important to have a statistically realistic synthetic population as an initial state because it is the starting point for calibration. Indeed, the initial state has an impact on future evolutions of the model.

The unknown model parameters are parameters that we were not able to directly extract from data. We observe in Figure 3 that the unknown parameters are extracted from data through a calibration procedure while the fixed parameters are directly extracted from data to generate the model. The *SimVillages* model calibration cannot be done analytically. Therefore, to calibrate the model, we vary the unknown parameters and we choose the value that minimizes a target function, defined as the distance between statistics constructed from simulated and observed data. To do this, we need to explore

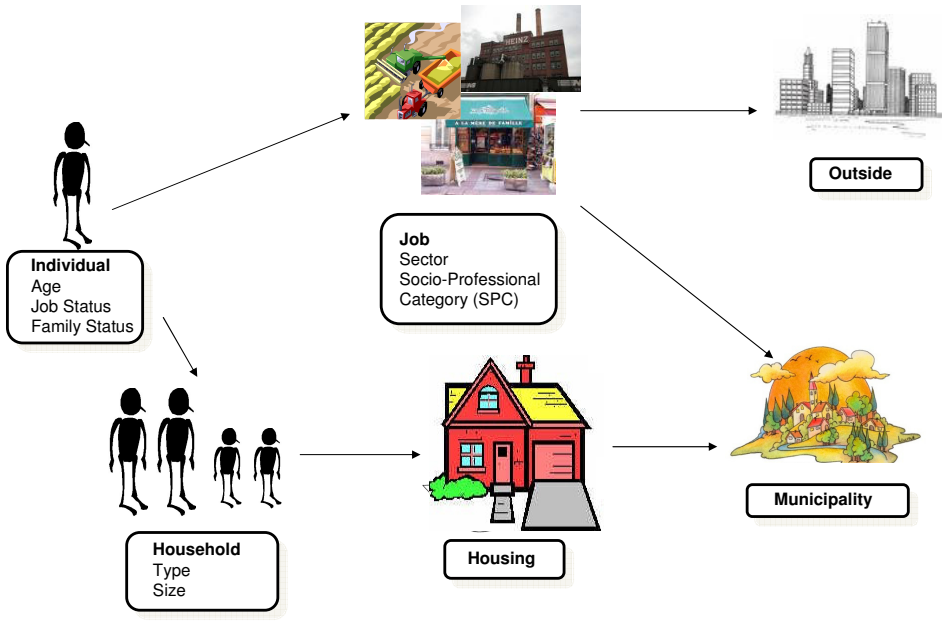
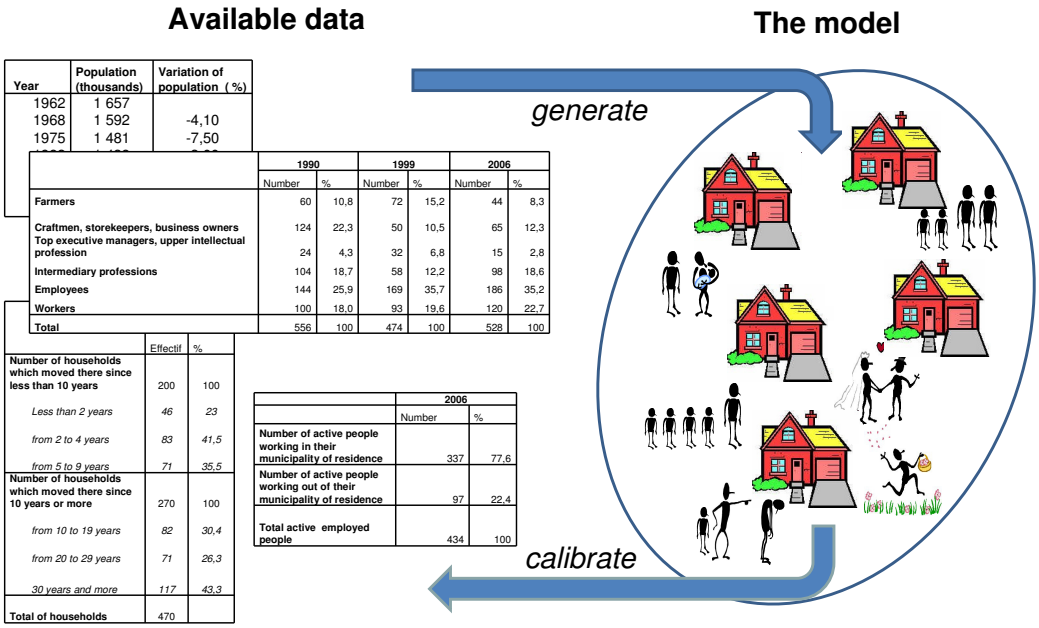


Figure 2: Main components of the *SimVillages* model

efficiently the parameter space to find the parameter values minimizing the target. This requires a large number of model simulations. For example, in Figure 1, we need to find θ values which have at least one trajectory "near" the observed data (represented by the green points).



Year	Population (thousands)	Variation of population (%)
1962	1 657	
1968	1 592	-4,10
1975	1 481	-7,50

	1990		1999		2006	
	Number	%	Number	%	Number	%
Farmers	60	10,8	72	15,2	44	8,3
Craftmen, storekeepers, business owners	124	22,3	50	10,5	65	12,3
Top executive managers, upper intellectual profession	24	4,3	32	6,8	15	2,8
Intermediary professions	104	18,7	58	12,2	98	18,6
Employees	144	25,9	169	35,7	186	35,2
Workers	100	18,0	93	19,6	120	22,7
Total	556	100	474	100	528	100

	Effectif	%
Number of households which moved there since less than 10 years	200	100
Less than 2 years	*46*	*23*
from 2 to 4 years	*83*	*41,5*
from 5 to 9 years	*71*	*35,5*
Number of households which moved there since 10 years or more	270	100
from 10 to 19 years	*82*	*30,4*
from 20 to 29 years	*71*	*26,3*
30 years and more	*117*	*43,3*
Total of households	470	

	2006	
	Number	%
Number of active people working in their municipality of residence	337	77,6
Number of active people working out of their municipality of residence	97	22,4
Total active employed people	434	100

Figure 3: From statistics to individuals

In the dynamics of the model, at each time step, the population of municipalities evolves, individuals make choices in life about study, career, marriage. They may have children, divorce, migrate, and die. The model takes into account endogenously inter-municipal migrations and creations or destructions of jobs in local services based on the number of inhabitants. In addition to these endogenous changes, scenarios are introduced representing the policy decisions taken at the regional level such as the establishment of a company in a municipality. These scenarios exogenously change the evolution of municipalities.

The study area modeled with *SimVillages* is the French department of Cantal (Figure 4), which had 158,723 inhabitants gathered in 260 municipalities in 1990. The model has for its starting point the year 1990 and the model results are evaluated in 1999 and 2006 (years corresponding to the population census conducted by the French Statistical Institute, INSEE [2]). A simulation on a desktop computer takes about a minute. A complete description of the model is available in Huet et al. (2012a) and the parametrization is detailed in Huet et al. (2012b) (available in Appendix A).

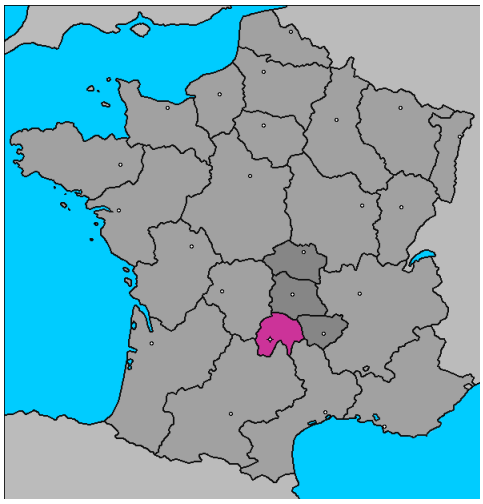


Figure 4: Map to locate the Cantal departement in metropolitan France.

Structure of thesis

This thesis is divided into four chapters. The first two are dedicated to the initialization of the *SimVillages* model. The third presents a statistical model aimed at estimating the number of jobs in proximity services. In the last one, we propose an algorithm using a Bayesian approach for estimating the posterior distribution of the unknown model parameters.

During my PhD, I implemented and validated the algorithm proposed by Gargiulo et al. (2010) that requires only aggregated data to create a synthetic population of the

[2] Institut national de la statistique et des études économiques

Auvergne French region in 1990. In Lenormand and Deffuant (2012), presented in Chapter 1, we compare this sample-free algorithm and a sample-based method, called Iterative Proportional Updating (IPU), proposed by Ye et al. (2009) for generating a synthetic population, organized in households, from various statistics. We generate a reference population for the Auvergne region including 1310 municipalities and measure how both methods approximate it from a set of statistics derived from this reference population. We also perform a sensitivity analysis. The sample-free method better fits the reference distributions of both individuals and households. It also demands less data but it requires more pre-processing. The quality of the results for the sample-based method is highly dependent on the quality of the initial sample.

In order to finalize the synthetic population and to create a socio-economic link with the 1310 Auvergne municipalities we needed to assign a place of work to each individual working outside his municipality of residence. The network of municipalities formed by these journeys to work is called a commuting network. Since detailed data was unavailable in 1990 it was necessary to develop an algorithm to generate commuting networks from aggregated data. This model was proposed in Gargiulo et al. (2012) (available in Appendix B). The model takes as input the number of commuters coming in and out of each municipality and it builds the network progressively, allocating commuters one by one in the different flows. This allocation is made according to probabilities that increase with the number of commuters coming to the destination, and decrease with the distance between the origin and destination. Then the model was adapted to 34 regions of France (Lenormand et al., 2012b). In this paper, available in Appendix C, we propose a generalization of the model including an artificial entity representing the population located outside the considered region (offering the commuters the possibility to work outside of the region) and we propose a comparison between an exponential and a power function to model the effect of the distance. In Lenormand et al. (2012c), presented in Chapter 2, we generate commuting networks on 80 case studies from different regions of the world (Europe and United-States) at different scales (e.g. municipalities, counties, regions). We show that the single parameter of the model follows a law that depends only on the scale of the geographic units (municipality, canton, county). We show that our model significantly outperforms two other approaches proposing a universal commuting model (Balcan et al., 2009; Simini et al., 2012), particularly when the geographic units are small (e.g. municipalities).

For the *SimVillages* model we have also developed a statistical model estimating the number of jobs in proximity services in a municipality. First, we have tried to estimate in a municipality the presence or absence of local services and also the number of jobs in these different services depending on the characteristics of the municipality (see Appendix D). Despite interesting results, this work is weakened by a strong difficulty: it requires to distinguish between local and non-local services in the data and we did not find any rigorous method to perform this task. In Lenormand et al. (2012a), presented in Chapter 3, we use a minimum requirement approach (Ullman and Dacey, 1960) to derive the number of jobs in proximity services per inhabitant in French rural municipalities. We first classify the municipalities according to their time distance in minutes by car to the municipality where the inhabitants go most frequently to obtain services

(called MFM). For each set corresponding to a range of time distance to MFM, we perform a quantile regression estimating the minimum number of service jobs per inhabitant which we interpret as an estimation of the number of proximity jobs per inhabitant. We observe that the minimum number of service jobs per inhabitant is smaller in small municipalities. Moreover, for municipalities of similar sizes, when the distance to the MFM increases, the number of jobs in proximity services per inhabitant increases.

To calibrate the *SimVillages* model, in Lenormand et al. (2012d) (available in Chapter 4) we proposed an approximate Bayesian computation (ABC) algorithm using importance sampling (for a complete review of ABC methods see Marin et al. (2012)). The sampling methods applied to ABC are derived from traditional sampling methods and they are considered as the most efficients of the ABC methods in terms of computation time. This new approximate Bayesian computation algorithm aims at minimizing the number of model runs for reaching a given quality of the posterior approximation. We performed a sensitivity analysis of the parameters of our algorithm and we compared it to the three competing algorithms found in the recent literature. When applied to a toy example and to the *SimVillages* model, our algorithm is two to eight times faster than the three other algorithms in reaching at least the same quality of results.

To conclude, we summarize the results obtained and we present perspectives and open questions.

CHAPTER 1

Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods

Contents

Abstract. We compare a sample-free method proposed by Gargiulo et al. (2010) and a sample-based method proposed by Ye et al. (2009) for generating a synthetic population, organised in households, from various statistics. We generate a reference population for a French region including 1310 municipalities and measure how both methods approximate it from a set of statistics derived from this reference population. We also perform sensitivity analysis. The sample-free method better fits the reference distributions of both individuals and households. It is also less data demanding but it requires more pre-processing. The quality of the results for the sample-based method is highly dependent on the quality of the initial sample.

Manuscript:
Lenormand, M. and Deffuant, G. (2012). Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods. *arXiv:1208.6403v1* (Submitted in *Journal of Artificial Societies and Social Simulation*).

1.1 Introduction

For two decades, the number of microsimulation models, simulating the evolution of large populations with an explicit representation of each individual, has been constantly increasing with the computing capabilities and the availability of longitudinal data. When implementing such an approach, the first problem is initialising properly a large number of individuals with the adequate attributes. Indeed, in most of the cases, for privacy reasons, exhaustive individual data are excluded from the public domain. Aggregated data at various levels (municipality, county,...), guaranteeing this privacy, are hence only available in general. Sometimes, individual data are available on a sample of the population, these data being chosen also for guaranteeing the privacy (for instance omitting the individual's location of residence). This paper focuses on the problem of generating a virtual population with the best use of these data, especially when the goal is generating both individuals and their organisation in households.

Two main methods, both requiring a sample of the population, aim at tackling this problem:

- The synthetic reconstruction method (SR) (Wilson and Pownall, 1976). These methods generally use the Iterative Proportional Fitting (Deming and Stephan, 1940) and a sample of the target population to obtain the joint-distributions of interest (Beckman et al., 1996; Huang and Williamson, 2002; Guo and Bhat, 2007; Arentze et al., 2007; Ye et al., 2009). Many of the SR methods match the observed and simulated households joint-distribution or individual joint-distribution but not simultaneously. To circumvent these limitations Guo and Bhat (2007); Arentze et al. (2007); Ye et al. (2009) proposed different techniques to match both household and individual attributes. Here, we focus on the Iterative Proportional Updating developed by Ye et al. (2009).

- The combinatorial optimization (CO). These methods create a synthetic population by zone using marginals of the attributes of interest and a sub-set of a sample of the target population for each zone (for a complete description see Voas and Williamson (2000); Huang and Williamson (2002)).

Recently, sample-free SR methods appeared (Gargiulo et al., 2010; Barthelemy and Toint, 2012). These methods can be used in the usual situations where no sample is available and one must only use distributions of attributes (of individuals and households). Hence, they overcome a strong limit of the previous methods. It is therefore important to assess if this larger scope of the sample-free method implies a loss of accuracy compared with the sample-based method.

The aim of this paper is contributing to this assessment. With this aim, we compare the sample-based IPU method proposed by Ye et al. (2009) with the sample-free approach proposed by Gargiulo et al. (2010) on an example.

In order to compare the methods, the ideal case would be to have a population with complete data available about individuals and households. It would allow us to measure precisely the accuracy of each method, in different conditions. Unfortunately, we

do not have such data. In order to put ourseleves in a similar situation, we generate a virtual population and then use it as a reference to compare the selected methods as in Barthelemy and Toint (2012).

In the Section 1.2 we formally present the two methods. In the Section 1.3 we present the comparison results. Finally, we discuss our results.

1.2 Details of the chosen methods

1.2.1 Sample-free method

We consider a set of n individuals X to dispatch in a set of m households Y in order to obtain a set of filled households P. Each individual x is characterised by a type t_x from a set of q different individual types T (attributes of the individual). Each household y is characterised by a type u_y from a set of p different household types U (attributes of the household). We define $n_T = \{n_{t_k}\}_{1\leq k\leq q}$ as the number of individuals of each type and $n_U = \{n_{u_l}\}_{1\leq l\leq p}$ as the number of households of each type. Each household y of a given type u_y has a probability to be filled by a subset of individuals L, then the content of the household equals L, which is denoted $c(y) = L$. We use this probability to iteratively fill the households with the individuals of X.

$$\mathbb{P}(c(y) = L|u_y) \tag{1.1}$$

The iterative algorithm used to dispach the individuals into the households according to the Equation 1.1 is described in Algorithm 1.1. The algorithm starts with the list of individuals X and of the households Y, defined by their types. Then it iteratively picks at random a household, and from its type and Equation 1.1, derives a list of individual types. If this list of individual types is available in the current list of individuals X, then this filled household is added to the result, and the current lists of individuals and households are updated. This operation is repeated until one of the lists X or Y is void, or a limit number of iterations is reached.

Algorithm 1.1 The general iterative algorithm

INPUT: X and Y
OUTPUT: P
  Set $P = \varnothing$
  while $Y \neq \varnothing$ **do**
    Pick at random y from Y
    Pick at random L with a probability defined in Equation 1.1
    if $L \subset X$ **then**
      $P \leftarrow P \cup L$
      $Y \leftarrow Y \backslash \{y\}$
      $X \leftarrow X \backslash L$
    end if
  end while

In the case of the generation of a synthetic population, we can replace the selection of the list L by the selection of the individuals one at a time by order of importance in the household. In this case Equation 1.2 replaces Equation 1.1.

$$\begin{aligned} &\mathbb{P}(x_1 \in y | u_y) \times \\ &\mathbb{P}(x_2 \in y | u_y, x_1 \in y) \times \\ &\mathbb{P}(x_3 \in y | u_y, x_1 \in y, x_2 \in y) \times \\ &... \end{aligned} \tag{1.2}$$

The iterative approach algorithm associated with this probability is described in Algorithm 1.2. The principle is the same as previously, it is simply quicker. Instead of generating the whole list of individuals in the household before checking it, one generates this list one by one, and as soon as one of its member cannot be found in X, the iteration stops, and one tries another household.

Algorithm 1.2 The iterative algorithm

INPUT: X and Y
OUTPUT: P
  Set $P = \emptyset$
  while $Y \neq \emptyset$ **do**
    Pick at random y from Y
    Pick at random x_1 with a probability $\mathbb{P}(x_1 \in y | u_y)$
    Pick at random x_2 with a probability $\mathbb{P}(x_2 \in y | u_y, x_1 \in y)$
    Pick at random x_3 with a probability $\mathbb{P}(x_3 \in y | u_y, x_1 \in y, x_2 \in y)$
    ...
    if $\{x_1, x_2, x_3, ...\} \subset X$ **then**
      $P \leftarrow P \cup \{x_1, x_2, x_3, ...\}$
      $Y \leftarrow Y \backslash \{y\}$
      $X \leftarrow X \backslash \{x_1, x_2, x_3, ...\}$
    end if
  end while

In practice this stochastic approach is data driven. Indeed, the types T and U are defined in accordance with the data available and the complexity to extract the distribution of the Equation 1.2 increases with n_T and n_U. The distributions defined in Equation 1.2 are called distributions for affecting individual into household. In concrete applications, it occurs that one needs to estimate n_T, n_U and the probability distributions presented in Equation 1.2. This estimation implies that the Algorithm 1.2 can not converge in a reasonable time because of the stopping criterion ($Y \neq \emptyset$). This stopping criterion is equivalent to an infinite number of "filling" trials by households. In this case, we can replace the stopping criterion by a maximal number of iterations by households and then put the remaining individuals in the remaining households using relieved distributions for affecting individual into household.

In a perfect case where all the data are available and the time infinite, the algorithm would find a perfect solution. When the data are partial and the time constrained, it

is interesting to assess how this method manages to make the best use of the available data.

1.2.2 The sample-based approach

In this approach, proposed by Ye et al. (2009), starts with a sample P_s of P and the purpose is to define a weight w_i associated with each individual and each household of the sample in order to match the total number of each type of individuals in X and households in Y to reconstruct P. The method used to reach this objective is the Iterative Proportional Updating (IPU). The algorithm proposed in Ye et al. (2009) is described in Algorithm 1.3. In this algorithm, for each type of households or individuals j the purpose is to match the weighted sum ws_j with the estimated constraints e_j with an adjustement of the weights. w_i is the weight of household i in the weighted sample and e_j is an estimation of the total number of households or individuals j in P. This estimation is done separetely for each individual and household type using a standard IPF procedure with marginal variables. When the match between the weighted sample and the constraint become stable, the algorithm stops. The procedure then generates a synthetic population by drawing at random the filled households of P_s with probabilities corresponding to the weights. This generation is repeated several times and one chooses the result with the best fit with the observed data.

Table 1.1: The Iterative Proportional Updating Table. The light grey table represents the frequency matrix D showing the household (HH) type U and the frequency of different individual (Ind.) types T within each filled households for the sample P_s. The dimension of D is $|P_s|\times(p+q)$, where $|P_s|$ is the cardinal number of the sample P_s, q the number of individual types and p the number of household types. An element d_{ij} of D represents the contribution of filled household i to the frequency of individual/household type j.

Filled HH ID	**HH Type u_1**	$\cdots$	**HH Type u_p**	**Ind. Type t_1**	$\cdots$	**Ind. Type t_q**	**Weight**
1	d_{11}	$\cdots$	d_{1q}	d_{1q+1}	$\cdots$	d_{1q+p}	$\boldsymbol{w_1}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\boldsymbol{\lvert P_s\rvert}$	$d_{\lvert P_s\rvert 1}$	$\cdots$	$d_{\lvert P_s\rvert q}$	$d_{\lvert P_s\rvert q+1}$	$\cdots$	$d_{\lvert P_s\rvert q+p}$	$\boldsymbol{w_{\lvert P_s\rvert}}$
WS	$\boldsymbol{ws_1}$	$\cdots$	$\boldsymbol{ws_p}$	$\boldsymbol{ws_{p+1}}$	$\cdots$	$\boldsymbol{ws_{p+q}}$	
E	$\boldsymbol{e_1 = \hat{n_{u_1}}}$	$\cdots$	$\boldsymbol{e_p = \hat{n_{u_p}}}$	$\boldsymbol{e_{p+1} = \hat{n_{t_1}}}$	$\cdots$	$\boldsymbol{e_{p+q} = \hat{n_{t_q}}}$	
$\boldsymbol{\delta}$	$\boldsymbol{\delta_1}$	$\cdots$	$\boldsymbol{\delta_p}$	$\boldsymbol{\delta_{p+1}}$	$\cdots$	$\boldsymbol{\delta_{p+q}}$	

1.3 Generating a synthetic population of reference for the comparison

Because we cannot access any population with complete data available about individuals and households, we generate a virtual population and then use it as a reference to compare the selected methods as in Barthelemy and Toint (2012).

Algorithm 1.3 Iterative Proportional Updating algorithm

INPUT: P_s, ϵ
OUTPUT: P
Set $P = \varnothing$
Generate $D \in \mathrm{M}_{|P_s| \times (p+q)}(\mathbb{R})$ described by the light grey table in Table 1.1
Estimate n_T and n_U using the standard IPF procedure and store the resulting estimate into a vector $E = (e_j)_{1 \leq j \leq p+q}$ as in Table 1.1
for $i = 1$ to $|P_s|$ **do**
  Set $w_i = 1$
end for
for $j = 1$ to $p+q$ **do**
  Compute $sw_j = \sum_{i=1}^{|P_s|} d_{ij} w_i$
  Compute $\delta_j = \frac{|sw_j - e_j|}{e_j}$
end for
Compute $\delta = \frac{1}{p+q} \sum_{j=1}^{p+q} \delta_j$
Set $\delta_{\min} = \delta$
Set $\Delta = \epsilon + 1$
while $\Delta > \epsilon$ **do**
  Set $\delta_{\text{prev}} = \delta$
  for $j = 1$ to $p+q$ **do**
    for $i = 1$ to $|P_s|$ **do**
      if $d_{ij} \neq 0$ **then**
        $w_i = \frac{e_j}{ws_j} w_i$
      end if
    end for
    Compute $sw_j = \sum_{i=1}^{|P_s|} d_{ij} w_i$
  end for
  Compute $\delta = \frac{1}{p+q} \sum_{j=1}^{p+q} \delta_j$
  if $\delta < \delta_{\min}$ **then**
    Set $W_{\text{opt}} = (w_i)_{1 \leq i \leq |P_s|}$
    $\delta = \delta_{\min}$
  end if
  $\Delta = |\delta - \delta_{\text{prev}}|$
end while

We start with statistics about the population of Auvergne (French region) in 1990 using the sample-free approach presented above. The Auvergne region is composed of 1310 municipalities, 1,321,719 inhabitants gathered in 515,736 households. In average the municipalities had about 1000 inhabitants with a minimum of 25 and a maximum of 136,180.

1.3.1 Generation of the individuals

For each municipality of the Auvergne region we generate a set X of individuals with a stochastic procedure. For each individual of the age pyramid (distribution 1 in Table 1.2), we randomly choose an age in the bin and then we draw randomly an activity status according to the distribution 2 in Table 1.2.

1.3.2 Generation of the households

For each municipality of the Auvergne region we generate a set Y of households according to the total number of individual $n = |X|$ with a stochastic procedure. We draw at random households according to the distribution 3 in Table 1.2 while the sum of the capacities is below n and then we determine the last household to have n equal to the sum of the size of the households.

1.3.3 Distributions for affecting individual into household

Single

- The age of the individual 1 is determined using the distribution 4 (Table 1.2).

Monoparental

- The age of the individual 1 is determined using the distribution 4 (Table 1.2).
- The ages of the children are determined according to the age of individual 1 (An individual can do a child after 15 and before 55) and the distribution 6 (Table 1.2).

Couple without child

- The age of the individual 1 is determined using the distribution 4 (Table 1.2).
- The age of the individual 2 is determined using the distribution 5 (Table 1.2).

Couple with child

- The age of the individual 1 is determined using the distribution 4 (Table 1.2).
- The age of the individual 2 is determined using the distribution 5 (Table 1.2).
- The ages of the children are determined according to the age of individual 1 and the distribution 6 (Table 1.2).

Other

- The age of the individual 1 is determined using the distribution 4 (Table 1.2).
- The ages of the others individuals are determined according to the age of individual 1.

Table 1.2: Data description

ID	Description	Level
1	Number of individuals grouped by ages	Municipality (LAU2)
2	Distribution of individual by activity statut according to the age	Municipality (LAU2)
3	Joint-distribution of household by type and size	Municipality (LAU2)
4	Probability to be the head of household according to the age and the type of household	Municipality (LAU2)
5	Probability of having a couple according to the difference of age between the partners (from"-16years" to "21years")	National level
6	Probability to be a child (child=live with parent) of household according to the age and the type of household	Municipality (LAU2)

To obtain a synthetic population P with households Y filled by individuals X we use the Algorithm 1.2 where we approximate the Equation 1.2 with the distributions 4, 5 and 6 in Table 1.2. We put no constraint on the number of individuals in the age pyramid, hence the reference population does not give any advantage to the sample-free method.

1.4 Comparing sample-free and sample-based approaches

The attributes of both individuals and households are respectivily described in Table 1.3 and Table 1.4. The joint-distributions of both the attributes for individuals and households give respectively the number of individuals of each individual type $n_T = \{n_{t_k}\}_{1\leq k\leq q}$ and the number of households of each household type $n_U = \{n_{u_l}\}_{1\leq l\leq p}$. In this case, $q = 130$ and $p = 17$. It's important to note that p is not equal to $6 \cdot 5 = 30$ because we remove from the list of household types the inconsistent values like for example single households of size 5. We do the same for the individual types (removing for example retired individuals of age comprised betweeen 0 and 5).

Table 1.3: Individual level attributes

Attribute	Value
Age	[0,5[[5,15[[15,25[[25,35[[35,45[[45,55[[55,65[[65,75[[75,85[85 and more
Activity Statut	Student Active Inactive
Family Statut	Head of a single household Head of a monoparental household Head of a couple without children household Head of a couple with children household Head of an other household Child of a monoparental household Child of a couple with children household Partner Other

Table 1.4: Household level attributes

Attribute	Value
Size	1 individual 2 individuals 3 individuals 4 individuals 5 individuals 6 and more individuals
Type	Single Monoparental Couple without children Couple with children Other

1.4.1 Fitting accuracy measures

We need fitting accuracy measures to evaluate the adequacy between both observed O and estimated E household and individual distributions. The first measure is the Proportion of Good Prediction (PGP) (Equation 1.3), we choose this first indicator for the facility of interpretation. In the Equation 1.3 we multiplied by 0.5 because as we have $\sum_{k=1}^{p} O_k = \sum_{k=1}^{p} E_k$, each misclassified individual or household is counted twice (Harland et al., 2012).

$$PGP = 1 - \frac{1}{2}\frac{\sum_{k=1}^{p} |O_k - E_k|}{\sum_{k=1}^{p} O_k} \tag{1.3}$$

We use the χ^2 distance to perform a statistic test. Obviously the modalities with a zero value for the observed distribution are not included in the χ^2 computation. If we consider a distibution with p modalities different from zero in the observed distribution, the χ^2 distance follows a χ^2 distribution with $p-1$ degrees of freedom.

$$\chi^2 = \frac{\sum_{k=1}^{p} (O_k - E_k)^2}{\sum_{k=1}^{p} O_k} \tag{1.4}$$

For more details on the fitting accuracy measures see Voas and Williamson (2001).

1.4.2 Sample-free approach

To test the sample-free approach, we extract from the reference population, for each municipality, the distributions presented in Table 1.2. Then we use the procedure used for generating the population of reference but now with the constraints on the number of individuals from the age pyramid derived from the reference (remember that we did not have such constraints when generating the reference population). Then we have filled the households with the individuals one at a time using the distributions for affecting individual into household. We limit the number of iterations to 1000 trials by household: If after 1000 trials a household is not filled, we put at random individuals in this household and we change his type for "other". We repeat the process 100 times and we choose, for each municipality, the synthetic population minimizing the χ^2 distance between simulated and reference distributions for affecting individual into household.

In order to assess the robustness of the stochastic sample-free approach, we generate 10 synthetic populations by municipalities, yielding 13,100 synthetic municipality populations in total. For each of them and for each distributions for affecting individual into household we compute the p-value associated to χ^2 distance between the reference and estimated distributions. As we can see in the Figure 1.1a the algorithm is quite robust.

To validate the algorithm we compute the proportion of good predictions for each 13,100 synthetic populations and for each joint-distribution. We obtain an average of 99.7% of good predictions for the household distribution and 91.5% of good predictions for the individual distribution (Figure 1.1b). We have also compute the p-value of the χ^2 distance between the estimated and reference distributions for each of the synthetic

populations and for each joint-distribution. Among the 13,100 synthetic populations 100% are statistically similar to the observed one at a 0.95% level of confidence for the household joint-distribution and 94% for the individual joint-distribution.

In order to understand the effect of the maximal number of iterations by household, we repeat the previous tests for different values of this parameter (1, 10, 100, 500, 1000, 1500 and 2000) and we compute the mean proportion of good predictions obtained for both individual and household. We note that after 100 the quality of the results no longer changes (Figure 1.1c).

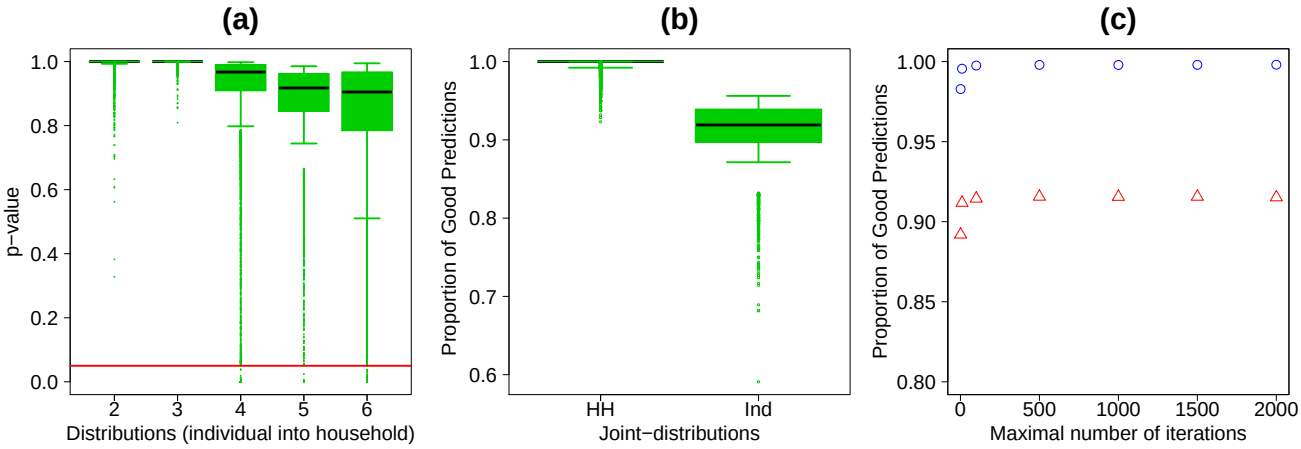


Figure 1.1: (a) Boxplots of the p-values obtained with the χ^2 distance between the estimated distributions and the observed distributions for each distributions for affecting individual into household, municipalities and replications. The x-axis represents the distributions presented in Table 1.2. The red line represents the risk 5% for the χ^2 test. (b) Boxplots of the proportion of good predictions for each joint-distribution, municipalities and replications. (c) Average proportion of good predictions in terms of the number of maximal iteration by households. Blue circles for the households. Red triangles for the individuals.

1.4.3 Iterative Proportional Updating

To use the IPU algorithm we need a sample of filled households and marginal variables. In order to obtain these data we pick at random a significant sample of 25% of households from the reference population P and we also extract from P the two one-dimensional marginals (Size and Type distributions) that we need to build the household joint-distributions with IPF and the three two-dimensional marginals (Age x Activity Statut, Age x Family Statut and Family Statut x Activity Statut joint-distributions) that we need to build the individual joint-distributions with IPF. Then we apply the Algorithm 1.3 using the recommendation of Ye et al. (2009) for the well-know zero-cell and zero-marginal problems to obtain a weighted sample P_s. With this sample we generate 100 times the synthetic population P and choose the one with lowest χ^2 distance between reference and simulated individual joint-distributions.

To check the results obtained with the IPU approach, we generate 10 synthetic pop-

ulations by municipality using different samples of 25% of households randomly selected. For each of these synthetic populations and for each joint-distribution we compute the proportion of good predictions (Figure 1.2a). We obtain an average of 98.6% of good predictions for the household distribution and 86.9% of good predictions for the individual distribution. To determine the error of estimation due to the IPF procedure we compute the proportion of good predictions for the estimated and the IPF-reference distributions. As we can see in Figure 1.2b the results are improved for the household distribution but not for the individual distribution. We also compute the p-value of the χ^2 distance between the estimated and observed distributions for each of the synthetic populations and for each joint-distribution. Among the 13,100 synthetic populations 100% are statistically similar to the observed one at a 0.95% level of confidence for the household joint-distribution and 61% for the individual joint-distribution. We obtained a similarity between the estimated and the IPF-objective distributions of 100% at a 0.95% level of confidence for the household distribution and 64% for the individual distribution.

In order to check the sensitivity of the results to the size of the sample, we plot, on Figure 1.2c, the average proportion of good predictions of the 13,100 household and individuals joint-distributons for different values of the percentage of the reference households drawn at random in the sample (5, 10, 15, 20 ,25, 30, 35, 40, 45 and 50). We note that the results are always good for the household distribution but for the individuals the results are good only from random sample of at least 25% of the reference household population. Not surprisingly, globally the quality of the results increases with the parameter.

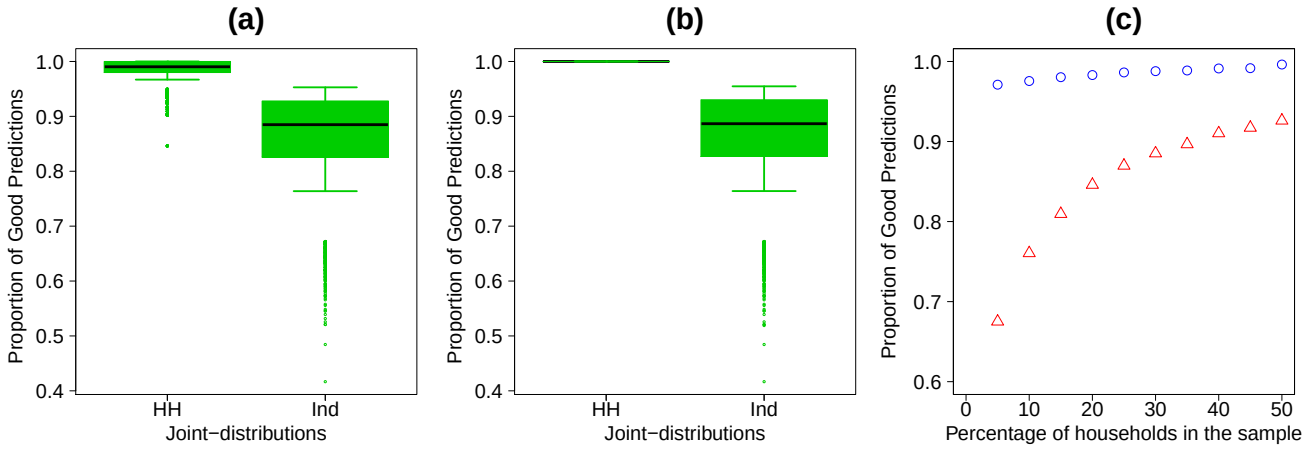


Figure 1.2: (a) Boxplots of the proportion of good predictions for a comparaison between the estimated distribution and the observed distribution for each municipality and replication. (b) Boxplots of the the proportion of good predictions for a comparaison between the estimated distribution and the IPF-objective distribution for each municipality and replication. (c) Average proportion of good predictions in terms of the sample percentage. Blue circles for the households. Red triangles for the individuals.

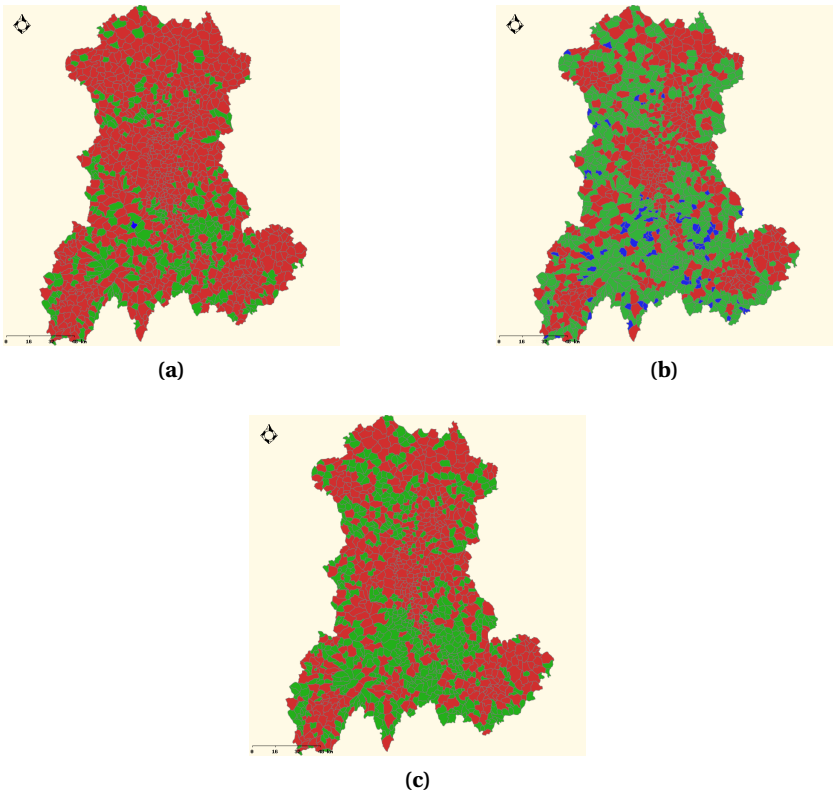
1.5 Discussion

The sample-free method is less data demanding but the data requires much pre-processing. Indeed, this approach requires to extract the distributions for affecting individual into household from data. The sample-free method gives better fit between observed and simulated distribution for both household and individual distribution than the IPU approach. We can observe in Figure 1.3 that, for both methods, the goodness-of-fit is correlated with the number of inhabitants. This observation is especially true for the IPU method because it depends on the number of individuals in the sample. Indeed, the lower is the number of individuals, the higher is the number of sparse cells in the individual distribution. The results obtained with the IPU approach depend on the quality of the initial sample. The execution time on a desktop machine (PC Intel 2.83 GHz) is almost the same for 100 maximal iterations by household for the sample-free method and 25% reference households drawn at random in the sample reference households for the sample-based approach (Table 1.5).

To conclude, the sample-free method gives globally better results in this application on small French municipalities. These results confirm those of Barthelemy and Toint (2012) who compared their sample-free method for working with data from different sources with a sample-based method (Guo and Bhat, 2007), and obtained similar conclusions. Of course, these conclusions cannot be generalized to all sample-free and sample-based methods without further investigation. However, these results confirm the possibility to initialise accurately microsimulation (or agent-based) models, using widely available data (and without any sample of households).

Table 1.5: Average execution time for the two approaches for different parameter values.

IPU		**Iterative**	
Sample size	**Time**	**Iterations**	**Time**
5	13min	1	40min
10	24min	10	41min
15	29min	100	45min
20	38min	500	58min
25	45min	1000	66min
30	53min	1500	78min
40	74min	2000	88min



(a) **(b)**

(c)

Figure 1.3: Maps of the average proportion of good predictions ((a) sample-free and (b) IPU) and the number of inhabitants ((c)) by municipality for the Auvergne case study. For (a)-(b), in blue $0.5 \leq \text{PGP} < 0.75$; In green $0.75 \leq \text{PGP} < 0.9$; In red $0.9 \leq \text{PGP}$. For (c), in green, the number of inhabitants is lower than 350. In red, the number of inhabitants is upper than 350. *Base maps source: Cemagref - DTM - Développement Informatique Système d'Information et Base de Données : F.Bray & A.Torre IGN (Géofla®, 2007).*

References

Arentze, T., Timmermans, H., and Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2014:85–91.

Barthelemy, J. and Toint, P. L. (2012). Synthetic Population Generation Without a Sample. *Transportation Science*.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6 PART A):415–429.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sample Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*, 11:427–444.

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, 5(1).

Guo, J. Y. and Bhat, C. R. (2007). *Population Synthesis for Microsimulating Travel Behavior*. Number 2014 in Transportation Research Record. Transportation Research Board of the National Academies.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of Artificial Societies and Social Simulation*, 15(1):1.

Huang, Z. and Williamson, P. (2002). A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata. Working paper, Departement of Geography, University of Liverpool.

Voas, D. and Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5):349–366.

Voas, D. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2):177–200.

Wilson, A. G. and Pownall, C. E. (1976). A new representation of the urban system for modelling and for the study of micro-level interdependence. *Area*, 8(4):246–254.

Ye, X., Konduri, K., Pendyala, R., Sana, B., and Waddell, P. (2009). Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. In *88th Annual Meeting of the Transportation Research Board*.

CHAPTER 2

A Universal Model of Commuting Networks

Contents

Abstract. We show that a recently proposed model generates accurate commuting networks on 80 case studies from different regions of the world (Europe and United-States) at different scales (e.g. municipalities, counties, regions). The model takes as input the number of commuters coming in and out of each geographic unit and generates the matrix of commuting flows between the units. The single parameter of the model follows a universal law that depends only on the scale of the geographic units. We show that our model significantly outperforms two other approaches proposing a universal commuting model (Balcan et al., 2009; Simini et al., 2012), particularly when the geographic units are small (e.g. municipalities).

Manuscript:
Lenormand, M., Huet, S., Gargiulo, F. and Deffuant, G. A Universal Model of Commuting Networks. *PLoS ONE* 2012, 7(10): e45985.

2.1 Introduction

Billions of people move everyday from home to workplace and generate networks of socio-economic relationships that are the vector of social and economic dynamics such as epidemic outbreaks, information flows, city development and traffic (Ortúzar and Willumsen, 2011; Balcan et al., 2009). Understanding the essential properties of these networks and reproducing them accurately is therefore a crucial issue for public health institutions, policy makers, urban development, infrastructure planners, etc. (De Montis et al., 2007, 2010). This challenge is the subject of an intensive scientific activity (see Barthélemy (2011); Rouwendal and Nijkamp (2004) for reviews), in which the analogy of the gravitational attraction inspires a majority of approaches (Wilson, 1998; Choukroun, 1975): the number of commuters between two geographic units (cities, counties, regions...) is supposed proportional to the product of the "masses" of each geographic unit (the population for example) and inversely proportional to a function of the distance between them. Unfortunately, numerous experiments showed that the optimum function and parameter values vary a lot with the case studies (De Vries et al., 2009; De Montis et al., 2007, 2010; Fotheringham, 1981). This situation is not satisfactory because when one wants to generate a particular commuting network without having the total origin destination matrix of commuting, no practical heuristic is available for choosing the adequate type of function and parameter values. This paper addresses this problem.

We consider a recently proposed model (Gargiulo et al., 2012; Lenormand et al., 2012), differentiating itself from the usual gravity law models in two main features:

- It takes as input the total number of commuters in and out from each geographic unit. With this starting point, the model focuses directly on the influence of the distance between geographic units on the commuting probability. The model is data demanding, but these data are widely available.

- It builds the network progressively, allocating commuters one by one in the different flows, according to probabilities that increase with the number of commuters coming in the destination and decrease with the distance between the origin and destination. These probabilities are updated after each allocation.

Our model is close to the traditional doubly-constrained gravity model (Wilson, 1998; Choukroun, 1975), but it is more flexible and less data demanding. Indeed, the doubly constrained model and the methods used to solve it require a closed network of commuters: they cannot take into account commuting links outside the considered geographical units. Our individual based stochastic approach overcomes this problem and can deal with the usually available data of total number of commuters in and out of geographic units.

We test this model on 80 case-studies with geographic units of different scales. For example in the same case-study the geographic unit can be either the municipality, the canton or the department, (see an example on Figure 2.1). More precisely, the case studies include: Czech Republic (municipality scale, 1 case-study), France (municipality

scale, 34 case-studies), France (canton scale, 15 case-studies including whole France), France (département scale one case-study (whole France), Italy (municipality scale, 10 case-studies), Italy (province scale, 4 case-studies), USA (county level, 15 case-studies including whole USA). For a detailed description of the datasets see the Appendix 2.A.

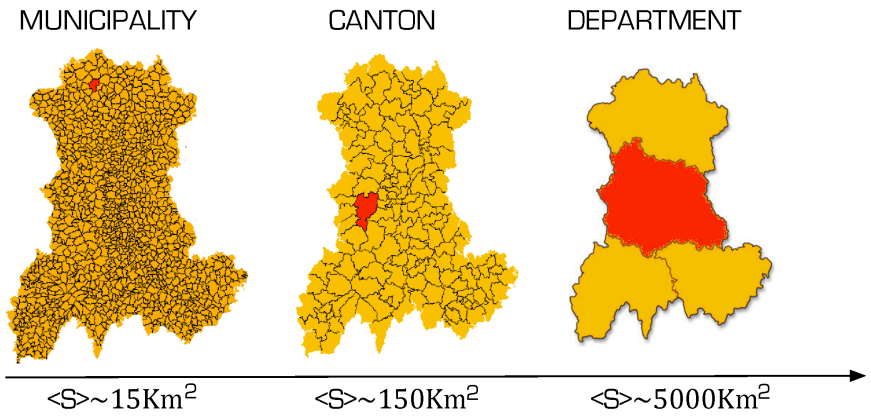


Figure 2.1: Three scales of geographic units (Auvergne region, France)

We show that the single parameter of our model follows a simple universal law that depends only on the average surface of the considered geographic units. This implies that, given the number of commuters in and out of each geographic unit and their average surface, we can derive the whole matrix of flows with a very good confidence.

Two other approaches (Balcan et al., 2009; Simini et al., 2012) claim to catch universal properties of commuting networks. We show that our model yields significantly more accurate results, especially for case-studies with small geographic units (e.g. municipalities).

2.2 The model

We consider the basic double-constrained model setup, without adding any ingredient about the job market characteristics (professions, salary range, etc.). Instead of solving analytically the optimisation problem, we use an individual based procedure that allocates virtual individuals one by one in the different flows between geographic units, according to a probability that is updated after each allocation.

This individual based approach can deal with less constrained data than the doubly-constrained gravity model that requires the total number of commuters in to be equal to the total number of commuters out. In other words the doubly contrained model can only deal with the flows between the considered geographic units; it cannot take into account the commuting links with destinations outside the case study area. This

is a problem when only the numbers of commuters in and out the geographic units are available (and not the complete matrix of the commuting flows), because the data do not distinguish between the flows inside and outside the case study area. It is therefore difficult to estimate the correct data to take as input to the doubly-constrained model in this case. Our approach is more flexible and overcomes this difficulty. It does not require that the total number of commuters in and out to be equal (for more details see Lenormand et al. (2012)), hence it can easily use directly the usually available data on the number of commuters in and out of each geographic unit.

Let s_i^{out} and s_j^{in} be respectively the global number of commuters starting from unit u_i and the global number of commuters arriving in unit u_j. These numbers are initialised from data and then they are progressively modified by the procedure. More precisely, at each step we select unit u_i such that $s_i^{out} > 0$ at random, and we consider a virtual commuter starting from u_i. We draw at random the working place u_{j^*} of this individual among all possible destinations u_j according to probabilities $P_{i\to j}$:

$$P_{i\to j} = \frac{s_j^{in} e^{-\beta D_{ij}}}{\sum_{k=1}^{N} s_k^{in} e^{-\beta D_{ik}}} \tag{2.1}$$

where D_{ij} is the Euclidian distance in meter between units u_i and u_j (computable from the Lambert or GIS coordinates). Having drawn u_{j^*}, we decrement of one s_i^{out} and $s_{j^*}^{in}$. Note that decrementing s^{in} and s^{out} at each step complicates significantly the derivation of an analytical expression of the model. We chose a probability decreasing exponentially with the distance, in accordance with the investigations carried out in Lenormand et al. (2012) and with the literature on commuting network models. The importance of the distance in the commuting choices is embedded in parameter β: for $\beta \to 0$ the probability tends to be independent from the distance, while for high values of β, the probability tends to zero very rapidly when the distance increases, independently from the number of commuters arriving in the units.

To reduce the border effect (see Lenormand et al. (2012)), we consider the job-search basin in an extended area, composed by the n residential units and m units surrounding the area. Thus, we have n units which are commuting origins and $N = n + m$ units that are commuting destinations. The generated network is saved in matrix $\tilde{T} \in M_{n\times N}(\mathbb{N})$ where each entry $\tilde{T}_{ij}$ represents the number of commuters between units u_i and u_j. The algorithm is presented in Algorithm 2.1.

2.3 A universal law ruling parameter β

The model depends on a single parameter ruling the importance of the distance in commuting choice. We show that this parameter can be derived as a function of the scale of the problem, independently from the socio-geographical location of the case study area. This opens the possibility to reconstruct the commuting flows (origin-destination matrix) when they are not provided.

We calibrated parameter β by maximising the common part of commuters (CPC), based on the Sørensen index (Sørensen, 1948).

Algorithm 2.1 Commuting generation model

INPUT: $D \in \mathrm{M}_{n\times N}(\mathbb{R})$, $s^{in} \in \mathbb{N}^N$, $s^{out} \in \mathbb{N}^n$, $\beta \in \mathbb{R}_+$
OUTPUT: $\tilde{T} \in \mathrm{M}_{n\times N}(\mathbb{N})$
  $\tilde{T}_{ij} \leftarrow 0$
  while $\sum_{k=1}^{n} s_k^{out} > 0$ **do**
    Pick at random $i \in |[1,n]|$, such that $s_i^{out} \neq 0$
    Pick at random j from $|[1,N]|$
      with a probability $P_{i\to j}$
    $\tilde{T}_{ij} \leftarrow \tilde{T}_{ij} + 1$
    $s_j^{in} \leftarrow s_j^{in} - 1$
    $s_i^{out} \leftarrow s_i^{out} - 1$
  end while
  return $\tilde{T}$

$$CPC(T,\tilde{T}) = \frac{2NCC(T,\tilde{T})}{NC(T)+NC(\tilde{T})} \tag{2.2}$$

with:

$$NCC(T,\tilde{T}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \min(T_{ij},\tilde{T}_{ij}) \qquad NC(T) = \sum_{i=1}^{n}\sum_{j=1}^{n} T_{ij} \tag{2.3}$$

where T is the observed origin-destination matrix and $\tilde{T}$ is the simulated one. This is a similarity measure based on the Sørensen index in ecology computing which part of the commuting flows is correctly reproduced, on average, by the simulated network. It varies between 0, when no agreement is found, and 1, when the two networks are identical. We priviledged this indicator because of its direct interpretation. Indeed, when $NC(T) \simeq NC(\tilde{T})$ (it is the case for our model), the CPC represents the percentage of commuting connection correctly located (i.e. with the right pair origin - destination). Moreover, we tested on all case studies that the results obtained with the MAE, the RMSE or CPC[1] are equivalent (see the Appendix 2.B for more details). As an example on the FR1 case study, Figure 2.2 shows that the same β value maximizes the CPC and minimizes the MAE. In this figure we can also note that the CPC is very sensitive to β and that its value does not vary much with the different replicas of the stochastic solving process.

Moreover, in order to have an idea of the improvement of the model compared with complete randomness, we have computed the CPC of a random model where the probabilities presented in Equation 2.1 are uniform ($P_{i\to j} = \frac{1}{n}$, where n is the number of units). As shown on the Figure 2.4 we obtained an average CPC around 0.1. For our model, the CPC is always higher than 0.7 with an average around 0.8, which can be interpreted as 70 to 80 % of correctly predicted commuting connections.

Our goal is to derive the value of β from some easily available global characteristics

[1] We have also shown in Gargiulo et al. (2012); Lenormand et al. (2012) that the value of β yielding the maximum CPC also yields the maximum similarity between observed and simulated commuting distance distributions

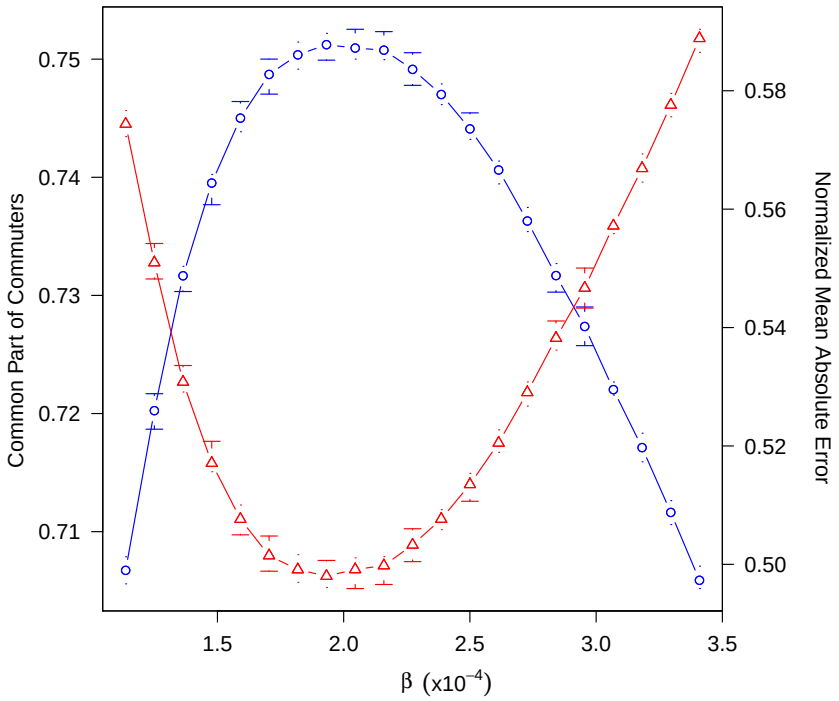


Figure 2.2: Plot of the average CPC (blue circle) and the average NMAE (red triangle) in term of β for 10 replications of the model for the Auvergne case study (FR1). The error bars represent the minimum value and maximum value obtain over the 10 replications.

of the case-study, giving the possibility to reconstruct the commuting flows when they are not available. Figure 2.3 gives strong evidence of such a universal relation.

The x-axis represents the average surface of the geographic units of the case-study ($\langle S \rangle$ in logarithm scale) and the y-axis the optimal β value (in logarithm scale). The linear regression in the log-log plane shows a simple relation:

$$\beta = \alpha \langle S \rangle^{-\nu} \tag{2.4}$$

with $\alpha = 3.15 \cdot 10^{-4}$ and $\nu = 0.177$. We observe that β decreases with the average surface of the units $\langle S \rangle$, meaning that, when $\langle S \rangle$ is small (e.g. for municipalities in France) the distance is more important in the commuting choice than when $\langle S \rangle$ is large (e.g. for regions or counties).

We now evaluate the robustness of our estimation of α and ν using a common statistical procedure: the cross-validation. The cross-validation aims at evaluating the potential error of using the β value derived from the regression model instead of deriving this value by optimisation for a new case study. This procedure repeats a large number of times the following steps: define a sub-sample of the total sample of case studies, derive a regression model of β from this sub-sample, for each case study that do not belong to the sub-sample, derive β from this regression model and compare the corresponding

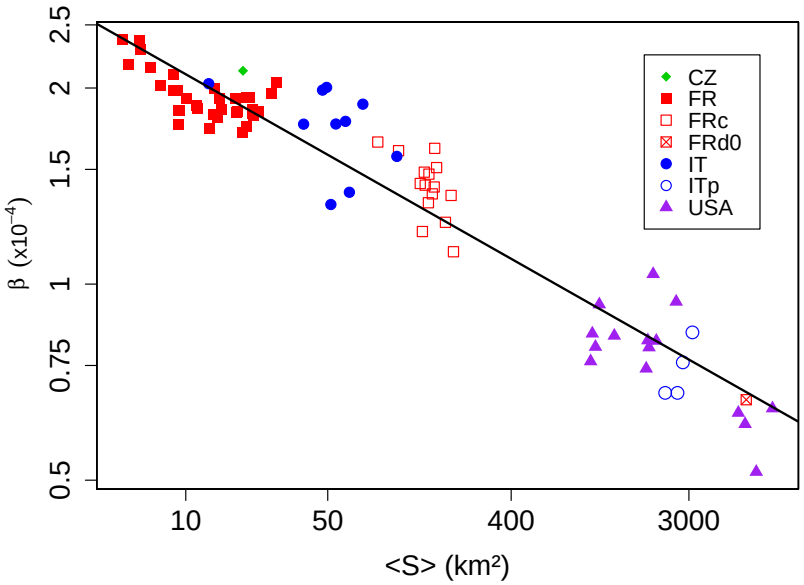


Figure 2.3: Log-log scatter plot of the calibrated β values in terms of average surface of the geographic units for 80 case-studies; the line represents the regression line predicting β.

CPC with the value of β directly calibrated on the complete origin - destination data. The dataset (including 80 case-studies) is randomly cut into two sets, called the training set (comprising 53 case-studies) and the test set (composed of 27 case-studies). We build a regression model on the training set, providing α and ν, from which we derive estimates of β for each of the 27 case-studies of the testing set. We have 27 estimations of β using the relation in Equation 2.4 where α and ν are obtained from the random subsample of 53 case-studies. We repeat this process 10,000 times obtaining 270,000 estimations of β (uniformly distributed over the 80 case-studies) corresponding to about $\frac{270,000}{80} = 3,375$ estimations of β for each case study. Then we calculate the average, minimum and maximum CPC for each of these values of β, and we compare them with the CPC obtained with value of β directly calibrated on the data.

Figure 2.4 shows, for each case-study, the CPC associated with the calibrated β, the average CPC obtained with the β values estimated from the cross-validation and the confidence interval defined by the minimum and the maximum values (but it is too small to be seen in most cases). The CPC obtained with the calibrated β value (black triangle) is almost the same as the average CPC obtained with the estimated β in most cases (red square). Globally, we can conclude that the β estimated with the log-linear model and the calibrated β lead to very similar CPCs and also very similar MAE and the RMSE as shown in the Appendix 2.B. The method appears therefore fairly robust and this gives confidence for using it with the value of β derived from our loglog regression in new cases studies.

2.4 Comparaison with other universal derivations of commuting networks

Two other different approaches, Balcan et al. (2009) and Simini et al. (2012), claim also to provide a universal derivation of commuting networks. The objective of Balcan et al. (2009) is to generate a worldwide commuting network, and the model must deal with the wide variety of populations and surfaces of geographic units for which the data are available. To solve this difficulty, the authors project these data on ad-hoc units defined with a Voronoi diagram. They define their basic unit as a cell approximately equivalent to a rectangle of 25 x 25 kilometers along the Equator. This allows them to calibrate their model because a unit is the same object whatever the country. This is an interesting solution for generating a world-wide commuting network but it leads to an average commuting distance of 250 km which is much larger than the average distance of daily commuting. For example for the USA case study the average distance of daily commuting is about 68 km for the observed network and about 64 km for the simulated network obtained with our algorithm. For the Auvergne (France) case study at municipality scale the average distance of daily commuting is about 12 km for the observed network and about 11 km for the simulated one.

In the radiation model, proposed in Simini et al. (2012), the commuting flow between two geographic units is a function of the cumulated population in a circle at the distance between the two units. The model has an elegant analytical solution and the average flow T_{ij} from unit u_i to unit u_j can be approximated by

$$\langle T_{ij} \rangle = \left(m_i \frac{P_c}{P} \right) \frac{m_i n_j}{\left(m_i + s_{ij}\right)\left(m_i + n_j + s_{ij}\right)} \tag{2.5}$$

where m_i and n_j are respectively the population of units u_i and u_j, P_c is the total number of commuters and P is the total population in the case-study region, and s_{ij} the total population in the circle of radius r_{ij} centred at u_i (excluding the source and destination population).

We implemented their analytical approximation and reproduced the graphs presented in their paper. Figure 2.5 shows the comparison between the radiation model and ours in the US for inter-county commuting and in the French Auvergne region for inter-municipality commuting. We observe that in both cases our approach yields significantly better results. Moreover, as shown on Figure 2.4, the average CPC for the radiation model on all the case studies is around 0.4, and lower for all case studies than the one obtained with our approach.

However, it should be reminded that our model uses more specific data (total number of commuters in and out of each geographic unit) than the radiation model, hence one could expect our results to be more accurate. Therefore, to be fair with the radiation model we implemented a modified version of this model using the number of out and in commuters of each units. This new approximation is presented in Equation 2.6 where s_{ij} the total number of in-commuters in the circle of radius r_{ij} centred at u_i (excluding the source and destination).

$$\langle T_{ij} \rangle = s_i^{out} \frac{s_i^{out} s_j^{in}}{\left(s_i^{out} + s_{ij}\right)\left(s_i^{out} + s_j^{in} + s_{ij}\right)} \tag{2.6}$$

As shown on Figure 2.4, this new model reaches an average CPC around 0.5 which is higher than the original radiation model but still significantly lower than the results obtained with our model. Using the MAE and the RMSE leads to the same conclusions (see the Appendix 2.B for more details).

2.5 Discussion

The power law of our model's single parameter β with the average area of the case study geographic units, is surprising to us because of the high variety in our case studies in terms of scale, number of units, number of commuters and surface areas. For instance the Auvergne region in France is rural with a population density of about 50 hab./km^2 whereas the New York City region is very urban with a population density of about 6500 hab./km^2. As far as we know, this is the first time that a single model is shown to fit such diverse group of datasets.

We show that our approach outperforms the radiation model and that the difference of input data plays a minor role in this superiority. This superiority is not due to our particular treatment of the border effects either. Indeed, we could check our approach outperforms the radiation model also on particular case studies (e.g. on islands such as Corsica) where this border effect does not play. We can conclude that the accuracy of our model comes from a proper use of the number of commuters in and out of each geographic unit and an adequate choice of the function of the distance.

The results of the cross validation procedure give a good confidence in the robustness of this law. However, we have to admit that, despite their diversity, our 80 case studies come all from western industrialised countries. Therefore it will be important to check the validity of our law on case studies coming from other continents and less industrialised countries. Moreover, we use a very rough approximation of the distance between the geographic units with the Euclidian distance between the unit centroids. More accurate approximations of this distance would certainly improve the results. Finally, we also intend to apply our approach to commuting networks inside urban areas because many cities of the world show an impressive growth and an increasing part of commuting takes place within them (Roth et al., 2011). An important issue in our perspective is to check if our law holds at this scale.

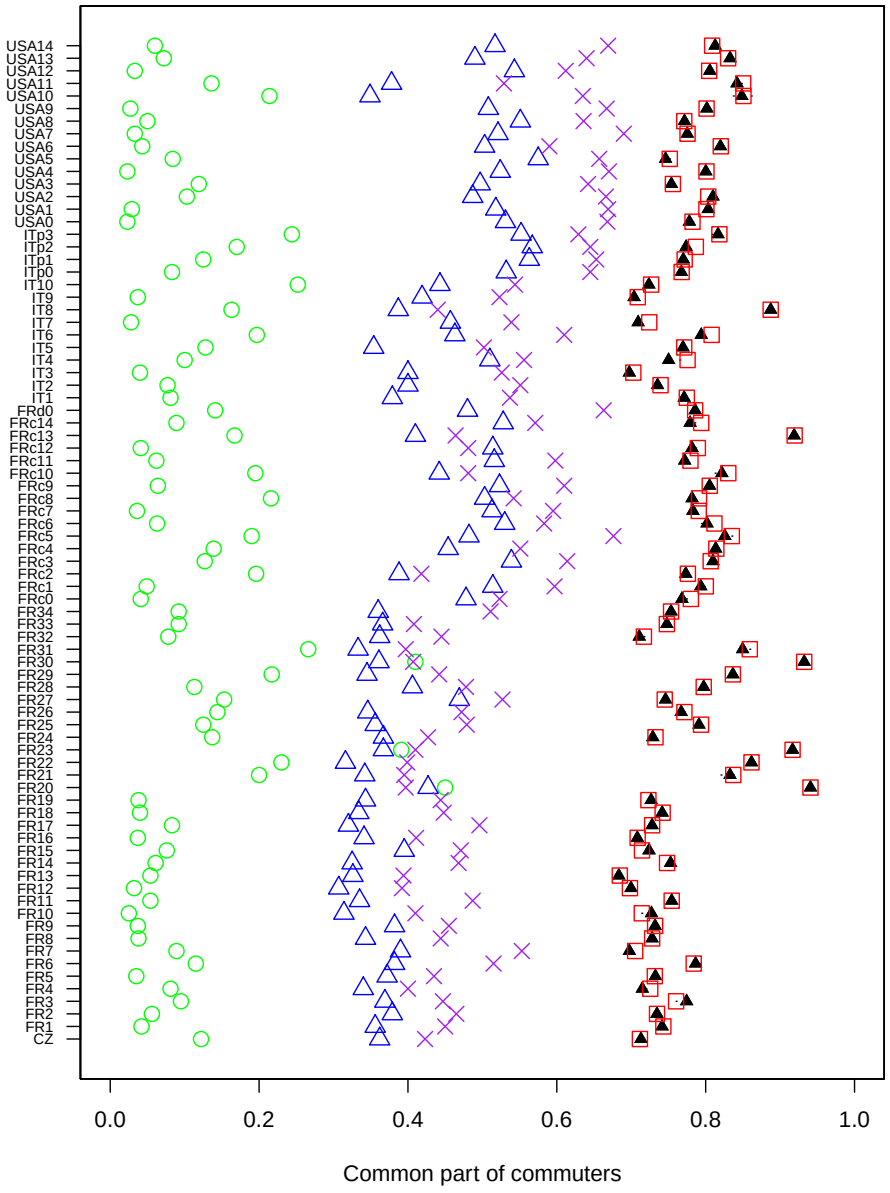


Figure 2.4: Common part of commuters (CPC) for the 80 case-studies. The red squares represent the CPC obtained with the value of β optimised from data on the case-study network. Black plain triangles represent the average CPC obtained with β values estimated with the rule linking β and the average surface of the units obtain with the cross-validation; Dark bars represent the minimum and the maximum CPC obtained with the estimated β but in most cases they are too close to the average to be seen. The green circles represent the CPC obtained with the random model. The blue triangles represent the CPC obtained with the radiation model. The purple crosses represent the CPC obtained with the modified version of the radiation model.

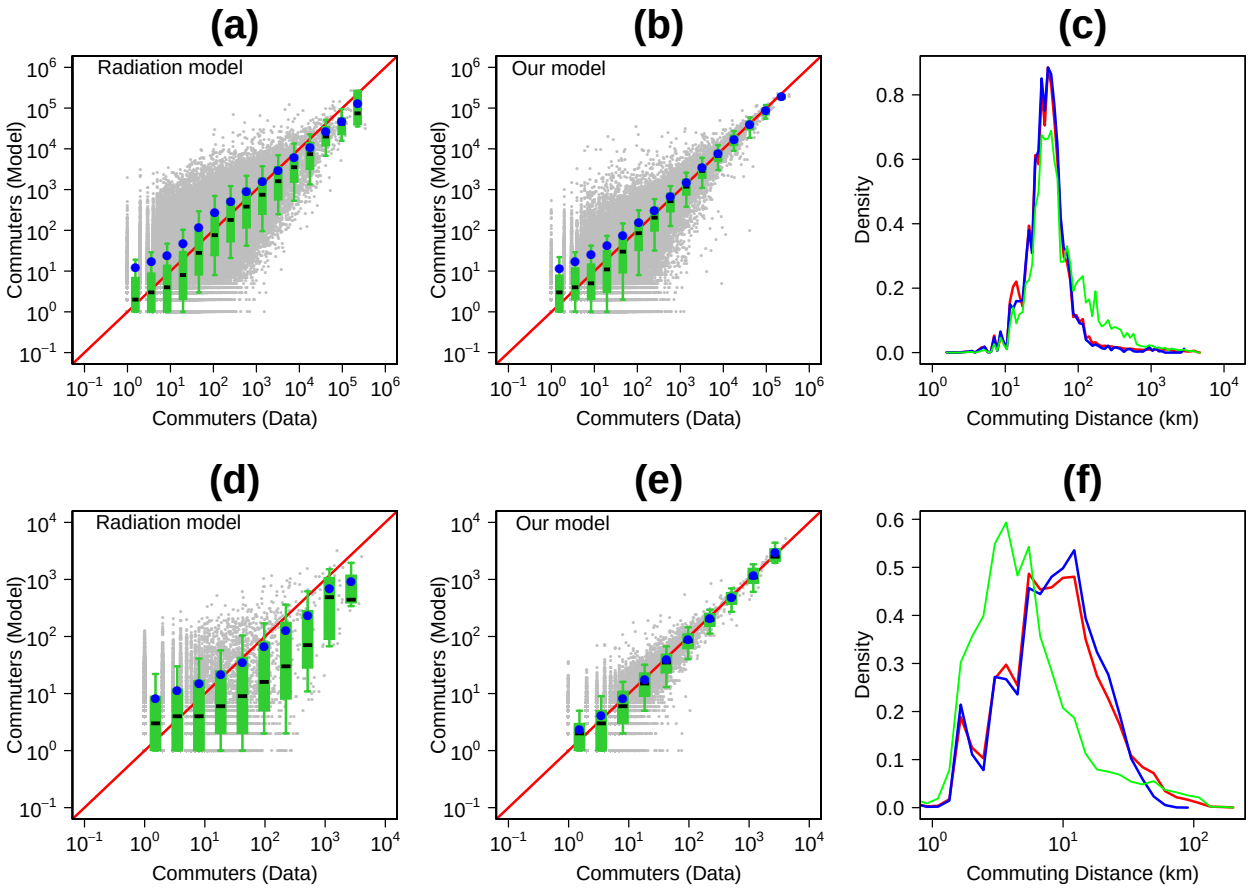


Figure 2.5: Comparing the predictions of the radiation model with ours for two case studies, the first row ((a)-(c)) for USA0 (USA at county scale) and the second row ((d)-(f)) for FR1 (Auvergne region, France at municipality scale). Plots (a), (b) , (d) and (e): Comparison between the observed (Census) and the simulated (model) non-zero flows. Grey points are the scatter plot for each pair of units. The boxplots (D1, Q1, Q2, Q3 and D9) represent the distribution of the number of simulated travelers in different bins of number of observed travelers. The blue circles represent the average number of simulated travelers in the different bins. Plots (c) and (f): Commuting distance distributions (km) (i.e. Probability for a commuters of the region to commut at a distance d). The blue line represents the observed data, the red one the results of our model and the green one the results of the radiation model.

References

Balcan, D., Colizza, V., Goncalves, B., Hud, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.

Barthélemy, M. (2011). Spatial Networks. *Physics Reports*, 499:1–101.

Choukroun, J.-M. (1975). A general framework for the development of gravity-type trip distribution models. *Regional Science and Urban Economics*, 5(2):177–202.

De Montis, A., Barthélemy, M., Chessa, A., and Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924.

De Montis, A., Chessa, A., Campagna, M., Caschili, S., and Deplano, G. (2010). Modeling commuting systems through a complex network analysis: A study of the Italian islands of Sardinia and Sicily. *The Journal of Transport and Land Use*, 2(3):39–55.

De Vries, J., Nijkamp, P., and Rietveld, P. (2009). Exponential or power distance-decay for commuting? An alternative specification. *Environment and Planning A*, 41(2):461–480.

Fotheringham, A. (1981). Spatial structure and distance-decay parameters. *Annals, Association of American Geographers*, 71(3):425–436.

Gargiulo, F., Lenormand, M., Huet, S., and Baqueiro Espinosa, O. (2012). Commuting Network Models: Getting the Essentials. *Journal of Artificial Societies and Social Simulation*, 15(2):6.

Lenormand, M., Huet, S., and Gargiulo, F. (2012). Generating French Virtual Commuting Network at Municipality Level. *arXiv:1109.6759v2*.

Ortúzar, J. and Willumsen, L. (2011). *Modeling Transport*. John Wiley and Sons Ltd, New York.

Roth, C., Kang, S. M., Batty, M., and Barthélémy, M. (2011). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*, 6(1).

Rouwendal, J. and Nijkamp, P. (2004). Living in two worlds: A review of home-to-work decisions. *Growth and Change*, 35(3):287–303.

Simini, F., Gonzalez, M. C., Maritan, A., and Barabasi, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34.

Wilson, A. G. (1998). Land-Use/Transport Interaction Models: Past and Future. *Journal of Transport Economics and Policy*, 32(1):3–26.

Appendix 2.A: Data description

The datasets

Commuting data are usually provided by statistical offices in the form of origin-destination tables. We analyzed 80 case studies from 7 differents datasets and 4 different countries (described in Table 2.1). In these appendices we called outside the m units surrounding the area.

The distances

The distances between units are Euclidean, computed using the Lambert coordinates or the latitude/longitude of the centroid of the units.

The case studies

We define two types of case studies: from administrative regions and from aggregation of small administrative units around a randomly chosen point. Each case study is composed of a region and an outside (the units surrounding the region at a reasonable distance).

To build a case study from an administrative region, we select an administrative region (for example the Auvergne region represented by the dark grey region in Figure 2.6a) and to build the outside we select all the units surrounding the region at a reasonable distance (for the Auvergne region example, the outside is represented by the light grey region in Figure 2.6a).

To build a case study by aggregation of units, firstly, we define the number of desired units and we draw at random a latitude and a longitude (for example the point represented in Figure 2.6b). In a second time we gradually increase the area of a square with as center the starting point until the desired number of units is obtained (Figure 2.6c). To build the outside we select all the units surrounding the defined set of units at a reasonable distance or all the remaining units in the country (it depends of the number of units).

The case studies with an identifier with a 0, for example FRc0, are complete network of the country without outside. Indeed, we have no data for the surrounding countries. When we consider a region in the country we can determine the outside as the units surrounding the region. When we consider as a region the whole country we can't determine an outside, it is the case for FRc0 (all the cantons of France), Frd0 (all the départements of France), Itp0 (all the provincias of Italy) and USA0 (all the counties of USA).

Sources

The 3 French datasets are measured for the 1999 French Census by the French Statistical Institute, INSEE. They were kindly made available by the Maurice Halbwachs Center.

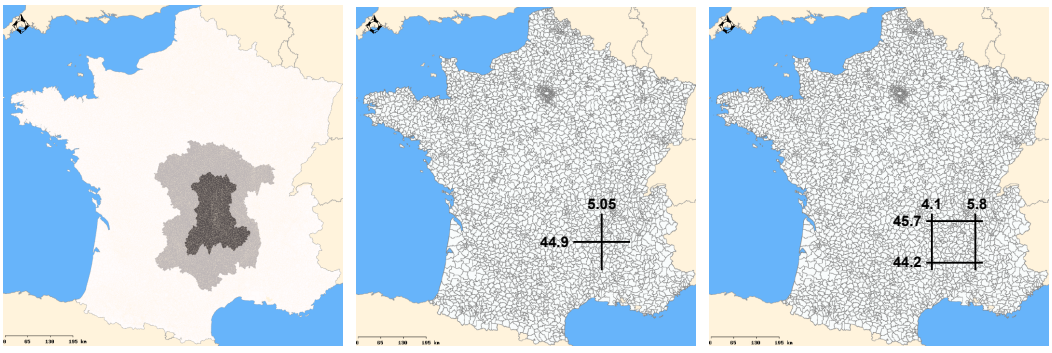
The 2 Italian datasets are measured for the 2001 Italian Census by the National Institute for Statistics, ISTAT.

Table 2.1: Presentation of the datasets

Dataset	Country	Case Study	Distance	Region	Scale	Year	Source
1	Czech Republic	CZ	Latitude Longitude	Administrative	Municipality	2001	*
2	France	FR1 - FR34	Lambert	Administrative	Municipality	1999	INSEE
3	France	FRc0 - FR14	Latitude Longitude	Arbitrary aggregation	Canton	1999	INSEE
4	France	FRd0	Latitude Longitude	Administrative	Département	1999	INSEE
5	Italy	IT1 - IT10	Latitude Longitude	Arbitrary aggregation	Municipality	2001	ISTAT
6	Italy	ITp0 - ITp4	Latitude Longitude	Arbitrary aggregation	Provincia	2001	ISTAT
7	USA	USA0 - USA14	Latitude Longitude	Arbitrary aggregation	County	2000	**

(*) Data are available online at http://www.czso.cz/

(**) Data are available online at http://www.census.gov/geo/www/gazetteer/places2k.html



(a) **(b)** **(c)**

Figure 2.6: Maps to illustrate the build process regions. (a) Administrative; (b) starting point of aggregation and (c) limits of aggregated units. *Base maps source: Cemagref - DTM - Développement Informatique Système d'Information et Base de Données : F.Bray & A.Torre IGN (Géofla®, 2007).*

Table 2.2: Description of the case studies

Case study	Number of units (area)	Number of units (outside)	Surface (km^2)	Average unit surface (km^2)	Standard deviation unit surface (km^2)	Observed number of commuters (area)	Estimated number of commuters (area)
CZ	43	630	35369	822.54	703.23	6585	6847
FR1	1310	3463	26013	19.86	12.49	261822	262452
FR2	1269	1447	27208	21.44	16.14	608587	613363
FR3	419	2809	5762	13.75	8.46	90456	76829
FR4	903	3081	8280	9.17	9.55	409661	402565
FR5	2296	2835	41309	17.99	21.30	679639	657095
FR6	261	3124	5175	19.83	10.46	52921	48681
FR7	185	1859	5167	27.93	18.71	9474	8981
FR8	1464	2467	25810	17.63	12.94	333045	333540
FR9	1842	4718	39151	21.25	14.76	514461	529535
FR10	3020	3845	45348	15.02	15.74	502326	494946
FR11	747	3169	16942	22.68	14.15	118508	117217
FR12	1786	3317	16202	9.07	7.46	239931	236314
FR13	1420	3536	12317	8.67	5.64	396800	402128
FR14	433	3914	6211	14.34	12.41	30175	28729
FR15	515	3808	5874	11.41	9.54	76519	72896
FR16	2339	3067	23547	10.07	7.51	505807	507812
FR17	260	1814	5565	21.40	13.15	17310	17071
FR18	1545	3046	27367	17.71	15.78	354824	354566
FR19	1948	1983	25606	13.14	12.94	333045	329908
FR20	36	1245	176	4.89	3.28	193236	182808
FR21	262	1543	2284	8.72	6.62	226205	206624
FR22	185	1707	1246	6.74	3.83	143938	124185
FR23	47	1234	245	5.21	3.03	143586	121474
FR24	377	2283	3525	9.35	7.44	160294	157123
FR25	195	2338	3718	19.07	17.66	26576	24975
FR26	547	449	4116	7.52	15.87	59709	61324
FR27	163	353	4299	26.37	27.53	145995	148922
FR28	327	2788	4781	14.62	9.76	134048	130910
FR29	102	2031	609	5.97	4.21	22520	20549
FR30	40	783	236	5.90	4.28	139181	125542
FR31	196	1597	1804	9.20	6.04	188855	165505
FR32	463	2588	5229	11.29	8.03	50505	51413
FR33	433	2728	6004	13.87	9.07	69377	63078
FR34	286	2088	5857	20.48	13.36	38141	37197
FRc0	3646	0	540241	171.72	99.90	12193161	12193161
FRc1	1062	2584	173797	163.65	91.23	2229003	2265247
FRc2	523	3123	58366	111.60	114.44	3892543	3922481
FRc3	226	3420	33041	146.20	70.56	548048	558086
FRc4	160	3486	25044	156.52	75.47	320432	323169
FRc5	55	3591	7847	142.67	71.64	61761	60285
FRc6	869	2777	131174	150.95	96.62	1995302	1983097
FRc7	2088	1558	351073	168.14	94.18	4459338	4523902
FRc8	100	3546	20246	202.46	161.41	307744	316592
FRc9	600	3046	113905	189.84	103.57	1078183	1095993
FRc10	302	3344	26627	88.17	77.64	1306425	1274670
FRc11	906	2740	142619	157.42	100.21	2324444	2358580
FRc12	1500	2146	250676	167.12	99.00	3224586	3284517
FRc13	32	3614	6653	207.91	145.33	11959	10634
FRc14	506	3140	75603	149.41	85.63	1311912	1331984
FRd0	94	0	540250	5747.35	1957.11	3548178	3548178
IT1	377	0	24090	63.90	61.89	225351	225351
IT2	395	201	24157	61.16	77.51	409889	408692
IT3	1002	2020	54918	54.81	71.37	1235378	1193338
IT4	201	507	14964	74.45	82.42	246609	248562
IT5	204	1005	10567	51.80	55.68	279014	272310
IT6	51	506	5582	109.45	101.52	57446	51211
IT7	2000	4001	98693	49.35	60.97	2849914	2812238
IT8	186	1023	2412	12.97	15.25	316602	286285
IT9	1510	4004	71167	47.13	58.08	1703944	1702002
IT10	705	3008	26809	38.03	41.62	401998	403307
ITp0	99	0	277220	2800.20	1619.86	1567576	1567576
ITp1	50	49	131773	2635.45	1401.23	742229	727038
ITp2	30	69	93666	3122.21	1599.56	266696	272316
ITp3	20	79	45854	2292.72	1128.38	264824	259988
USA0	3108	0	8070785	2596.78	3437.29	34077841	34077841
USA1	1015	2093	1876151	1848.42	916.86	5855813	5902784
USA2	103	3005	101411	984.57	341.47	527136	535608
USA3	54	3054	306284	5671.93	4488.99	604043	597371
USA4	2011	1097	4169235	2073.21	1786.40	14767588	14926726
USA5	202	2906	404093	2000.46	1994.32	8789633	8893748
USA6	504	2604	949238	1883.41	1041.57	2125887	2155981
USA7	806	2302	4234740	5254.02	5626.18	5003104	5099317
USA8	352	2756	2723212	7736.40	7741.02	4147054	4234376
USA9	1507	1601	2877429	1909.38	1517.28	10099598	10234438
USA10	13	3095	14123	1086.37	343.73	58212	53513
USA11	32	3076	205989	6437.17	4105.95	22496	24085
USA12	1004	2104	1292835	1287.68	563.79	9704950	9735646
USA13	207	2901	207785	1003.79	352.24	1307774	1326018
USA14	301	2807	312955	1039.72	394.71	2054878	2085408

Appendix 2.B: Results with standard indicators of error

We computed the results with standard indicators of error.

- The Normalized Mean Absolute Error:

$$NMAE(T,\tilde{T}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|T_{ij}-\tilde{T}_{ij}|}{\sum_{i=1}^{n}\sum_{j=1}^{n}T_{ij}} \tag{2.7}$$

- Normalized Root Mean Square Error:

$$NRMSE(T,\tilde{T}) = \frac{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(T_{ij}-\tilde{T}_{ij})^2}}{\sum_{i=1}^{n}\sum_{j=1}^{n}T_{ij}} \tag{2.8}$$

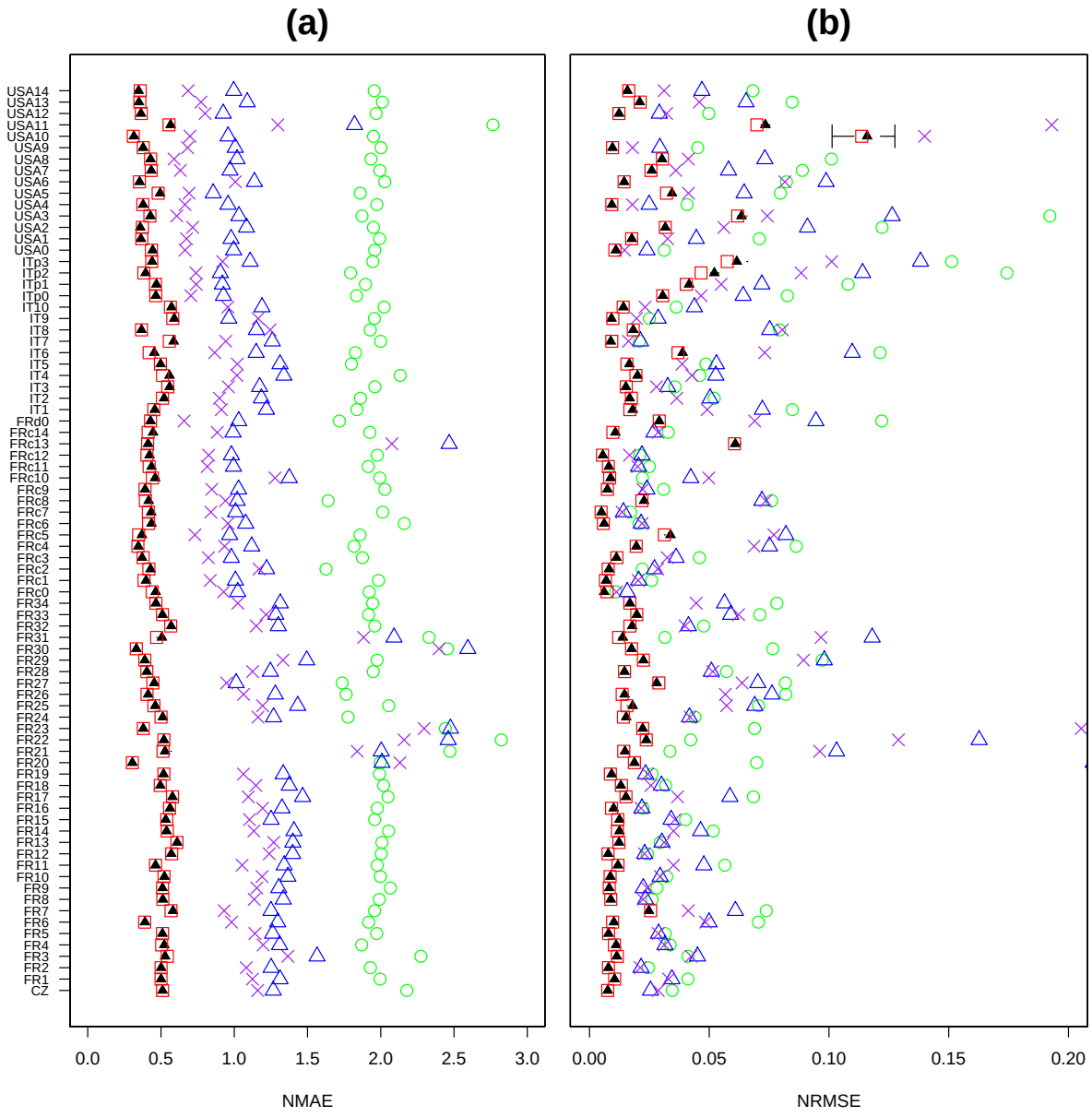


Figure 2.7: Normalized Mean Absolute Error (a) and Normalized Root Mean Square Error (b) for the 80 case-studies. The red squares represent the errors obtained with the value of β optimised from data on the case-study network. Black plain triangles represent the average errors obtained with β values estimated with the rule linking β and the average surface of the units obtain with the cross-validation; Dark bars represent the minimum and the maximum errors obtained with the estimated β but in most cases they are too close to the average to be seen. The green circles represent the errors obtained with the random model. The green circles represent the errors obtained with the random model. The blue triangles represent the value obtained with the radiation model. The purple cross represent the errors obtained with the modified version of the radiation model.

CHAPTER 3

Deriving the Number of Jobs in Proximity Services from the Number of Inhabitants in French Rural Municipalities

Contents

Abstract. We use a minimum requirement approach to derive the number of jobs in proximity services per inhabitant in French rural municipalities. We first classify the municipalities according to their time distance in minutes by car to the municipality where the inhabitants go the most frequently to get services (called MFM). For each set corresponding to a range of time distance to MFM, we perform a quantile regression estimating the minimum number of service jobs per inhabitant that we interpret as an estimation of the number of proximity jobs per inhabitant. We observe that the minimum number of service jobs per inhabitant is smaller in small municipalities. Moreover, for municipalities of similar sizes, when the distance to the MFM increases, the number of jobs of proximity services per inhabitant increases.

Manuscript:
Lenormand, M., Huet, S. and Deffuant, G. Deriving the Number of Jobs in Proximity Services from the Number of Inhabitants in French Rural Municipalities. *PLoS ONE* 2012, 7(7): e40001.

3.1 Introduction

How many service jobs does each inhabitant of a rural municipality generate in his own municipality? This question is important for the modelling work carried out in the PRIMA European project (Huet and Deffuant, 2011)[1], dealing with the evolution of rural areas in Europe. In particular, this model aims at incorporating how the growth or decline of municipalities is enhanced by the creation or destruction of these jobs. Indeed, new approaches based on the residential economy point out that the dynamism of rural areas depends significantly on the demand for locally consumed goods and services. We call proximity service these jobs that are generated by the local demand of the municipality, and this paper proposes a method for assessing their number.

Surprisingly, the literature on the estimation of proximity service job for demographic microsimulation models is very poor. Furthermore the estimation methods proposed are rather crude, for example Brown and Robinson (2006) proposed a threshold function to create service jobs for one hundred new people. For a direct estimation, the main difficulty is that the available data provide the number of jobs in different categories of services (retail, transportations, various services, public administration, teaching, health and social action) without any information about their relation with the local demand. In the same category, some jobs can depend on the very local market (the municipality), whereas others depend on a wider market of surrounding municipalities or even the whole region. Even the same job of service can be partially devoted to the local customers and partially to a larger market. Therefore, the number of jobs in proximity services can only be estimated indirectly.

In this paper, we propose to use the minimum requirement approach (Ullman and Dacey, 1960) to perform this indirect estimation. This method is usually used for estimating the share of jobs in a given activity (Ullman and Dacey, 1960; Brodsky and Sarfaty, 1977), the employment in touristic activities (Dissart et al., 2009; English et al., 2000; Leatherman and Marcouiller, 1996) or to compute the regional multipliers giving the propensity to consume locally produced goods (Rutland and O'Hagan, 2007; Woller and Parsons, 2002; Persky and Wiewel, 1994; Moore, 1975). In our case, the rationale behind choosing this method is that a large set of municipalities of similar proximity service market always includes some municipalities where the services are only devoted to this local market. These municipalities tend to have the minimum number of service jobs, which gives an estimation of the number of proximity service jobs.

We use two variables to characterise the proximity service market: the municipality size (number of inhabitants) and the offer of services in the neighbourhood. Indeed, the municipality size alone is certainly not sufficient to predict the number of jobs in proximity services because, in our data, the average distance between a municipality and its closest neighbour is about 4 km. Hence there are municipalities that can be very dependent on other ones for their proximity services. We describe the neighbouring offer of services with the time distance by car to the most frequented municipality (MFM). The MFM is the municipality where residents from a given municipality usually go to con-

[1] The research leading to these results has received funding from the European Commission's 7th Framework Programme FP7/2007-2013 under grant agreement n° 212345.

sume services, leisure equipment and facilities that they don't find in their own town.

In practice, we defined seven municipality sets corresponding to intervals of tMFM, the time distance to the MFM. In each set, following the minimum requirement approach, we assess the minimum number of jobs per inhabitants with a quantile regression (Koenker and Bassett, 1978), taking as quantile value the first percentile. Indeed, we choose the first percentile (100-quantile) instead of the minimum because the observed data are based on a sample representing a quarter of the population, and the percentile is likely to be more robust to the lack of precision than the minimum. Moreover there is no theoretical justification for using systematically the minimum value (Klosterman, 1990). For each of the seven intervals of tMFM, we obtain a satisfactory regression predicting the first percentile of service jobs per inhabitant. Moreover, the impact of tMFM corresponds to one's expectations: the municipalities which are close to a MFM have the lowest number of jobs in proximity services per inhabitant and, when tMFM increases, the number of jobs in proximity services per inhabitant increases.

The Section 3.2 presents the material and methods used for predicting the number of jobs in proximity services per inhabitant. We finally discuss our results.

3.2 Material and methods

3.2.1 The data from the French statistical office

This work uses data about municipalities of less than 5000 inhabitants coming from the French Census of 1999, 2006 and 2008 managed by the French Statistical Institute, INSEE and from the French Municipal Inventory of 1999. From this collected data, the Maurice Halbwachs Center or the INSEE makes available to all researchers the following data:

- The number of inhabitants for each municipality in 1999, 2006 and 2008;
- The number of jobs in the French tertiary sector (called service jobs) in 1999, 2006 and 2008;
- The time distance in minutes by car to the most frequented municipality (tMFM) in 1999;

The MFM is the municipality where residents from a given municipality usually go to consume services, leisure equipment and facilities that they don't find in their own municipality. This variable was obtained in 1999 by asking the following question to the mayor of each municipality *"Where do you go when you need something unavailable in your municipality?"*. The time distance to the MFM is expressed in minutes by car estimated with a average speed/km.

We observe in Figure 3.1 that the dataset is mostly composed of small municipalities with a small number of service jobs per inhabitant. We note that the minimum number of service jobs per inhabitant can be expressed by a linear relationship with the logarithm of the number of inhabitants. We observe in Figure 3.2 the time distance to the

most frequented municipality is mostly between 0 and 20 minutes. The higher is the tMFM, the more isolated is the municipality. The MFM of a given municipality is assumed to be the same in 2006 and 2008 as in 1999. In order to check the robustness of this assumption we have highlighted, for a given range of values of tMFM, the outliers for the bivariate variable *Number of inhabitants x Number of service jobs* in 1999 and 2008. For different range of values of tMFM, the number of outliers is almost the same in 1999 and in 2008, and there is about 80% of outliers in common between the two time series. We give an example for tMFM$\in]0,5]$ in Figure 3.3. Moreover the 20% "new" outliers in 2008 show a growth of inhabitants that is similar to the one of non-outliers. This does not validate completely the assumption but it reinforces its plausibility.

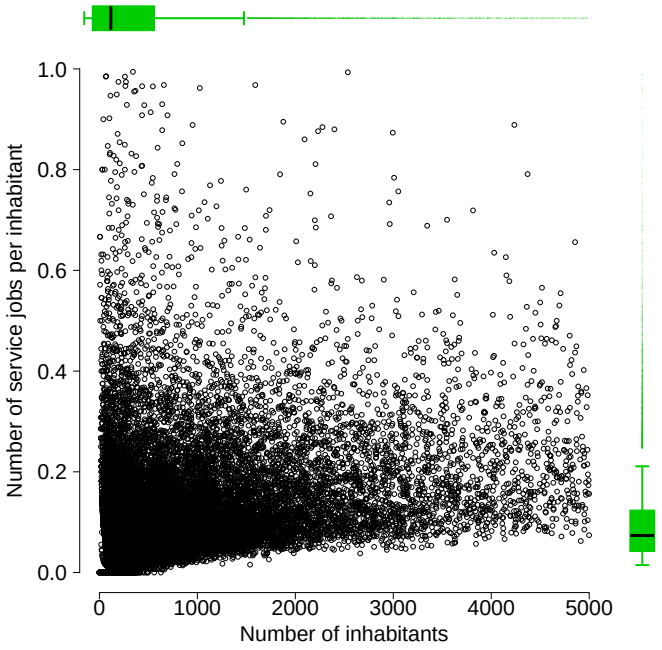


Figure 3.1: Number of service jobs per inhabitant function of the number of inhabitants for each municipality in 1999.

3.2.2 Model estimate of the number of jobs in proximity services per inhabitant

In this section, we present the model estimation of the number of jobs in proximity services per inhabitant based on a minimum requirement approach applied to several tMFM intervals. We assume that the number of jobs in proximity services per inhabitant in a municipality depends not only on the number of inhabitants but also on tMFM. Therefore, we define seven sets of municipalities corresponding to intervals of tMFM (values expressed in minutes): tMFM $\in]0,5]$, tMFM $\in]5,10]$, tMFM $\in]10,15]$, tMFM $\in]15,20]$, tMFM $\in]20,25]$, tMFM $\in]25,30]$ and tMFM > 30. For each of these sets

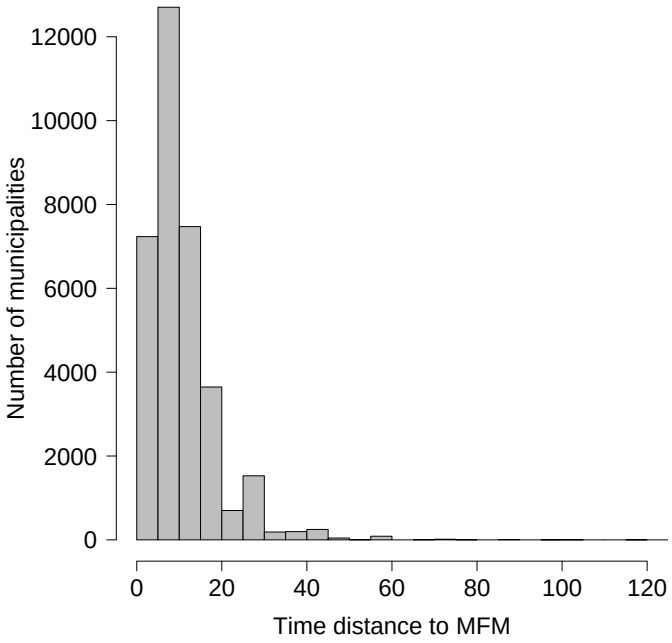


Figure 3.2: Histogram of the tMFM in minutes by car in 1999.

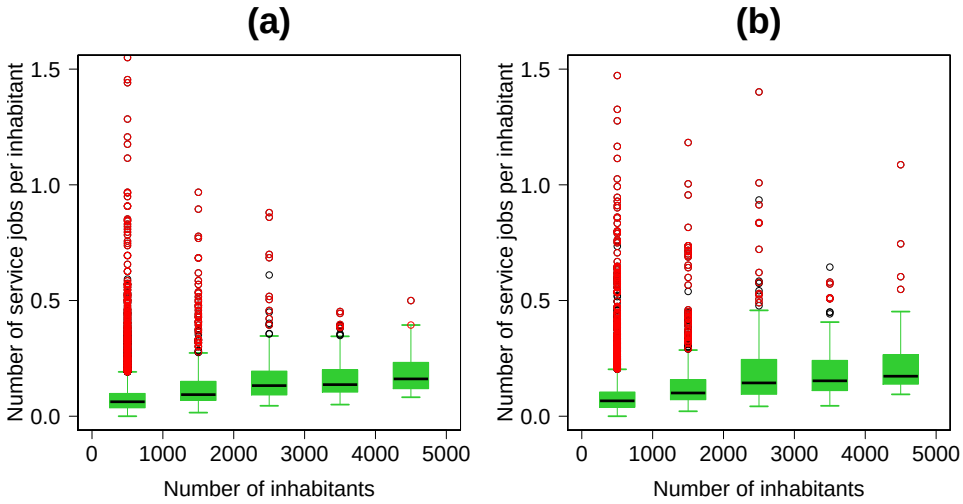


Figure 3.3: Box-and-whisker plot of the number of service jobs per inhabitant function of the number of inhabitants for tMFM $\in]0,5]$. The red points represent the common outliers 1999 and 2008. (a) 1999; (b) 2008.

of municipalities we apply a method derived from the minimum requirement approach to estimate the number of jobs in proximity services per inhabitant as a function of the municipality size.

In general, the minimum requirement approach computes minima on subsets of municipalities of similar sizes, which requires to define these subsets with an appropriate clustering method. We choose to use a quantile regression (Koenker and Bassett, 1978), which does not require to perform this clustering, and yields directly a function estimating the minimum (or a quantile). We choose the first-percentile ($\tau = 0.01$) in the regression because our data on the number of service jobs are derived from a sample representing a quarter of the population, and we expect the first percentile to be more robust than the minimum to this lack of precision.

Let E be the number of service jobs per inhabitant and P the number of inhabitants. We consider the following quantile regression model:

$$E = \beta_0 + \beta_1 \ln P + \epsilon$$

where β_0 and β_1 are parameters and ϵ the residual vector.

With this method, we estimate the number of jobs in proximity services per inhabitant as a function of the municipality size, for each interval of tMFM.

3.3 Results

In this section, we present the results obtained when applying the method on the data from 1999, 2006 and 2008.

The coefficients of the quantile regression for each set of tMFM obtain with the regression quantile $\tau = 0.01$ and the 1999 data are presented in Table 3.1. All the coefficients are significant and the associated standard deviations are quite low. The quantile regression model is significant with one percentile but also with five percentile, ten percentile and the median but we choose the focus on the results for one percentile because we want to be as close as possible to the minimum. Figure 3.4 shows the relation given by the model for 1999 for tMFM $\in]0,5]$ and tMFM > 30. As we can see on the scatter plots, we obtained a good fit of the model. To assess changes over time in the relationship we have repeated the procedure in 2006 and 2008 (using tMFM from 1999). We note that, for all the tMFM intervals, the slope is positive, and it is the highest for tMFM > 30. This implies that the number of proximity service jobs created (or destroyed) is higher in big municipalities than in a small one, when the population evolves, and even higher for municipalities that are far from their MFM.

Figure 3.5 shows the results for 2006 and 2008 with 1999 for a 500 and a 3000 inhabitants municipality. For each tMFM interval we observe that the number of proximity service jobs per inhabitant tends to increase with time. One can see that the number of proximity service jobs per inhabitant is smaller for tMFM < 15 and then increases. It is coherent with the results presented in Mordier (2010) which shows the number of service providers is higher in isolated rural area than in suburbs of rural center. The same author shows the number of service providers in rural suburbs is smaller than the on in rural centres (defined as having at least 1500 jobs). The whole form a curve is also coherent with Hubert (2009) who shows that in the rural and weakly urban areas the

Table 3.1: Parameter values and standard deviations (in brackets) of the quantile regression predicting the number of proximity services jobs per inhabitant for the different intervals of tMFM in minutes by car in 1999.

tMFM	Intercept	Slope
$]0,5]$	-0.084 (0.0031)	0.016 (0.0006)
$]5,10]$	-0.083 (0.0024)	0.016 (0.0005)
$]10,15]$	-0.079 (0.0014)	0.015 (0.0003)
$]15,20]$	-0.094 (0.0025)	0.018 (0.0005)
$]20,25]$	-0.097 (0.0021)	0.019 (0.0007)
$]25,30]$	-0.099 (0.0055)	0.019 (0.0012)
>30	-0.112 (0.0067)	0.021 (0.0020)

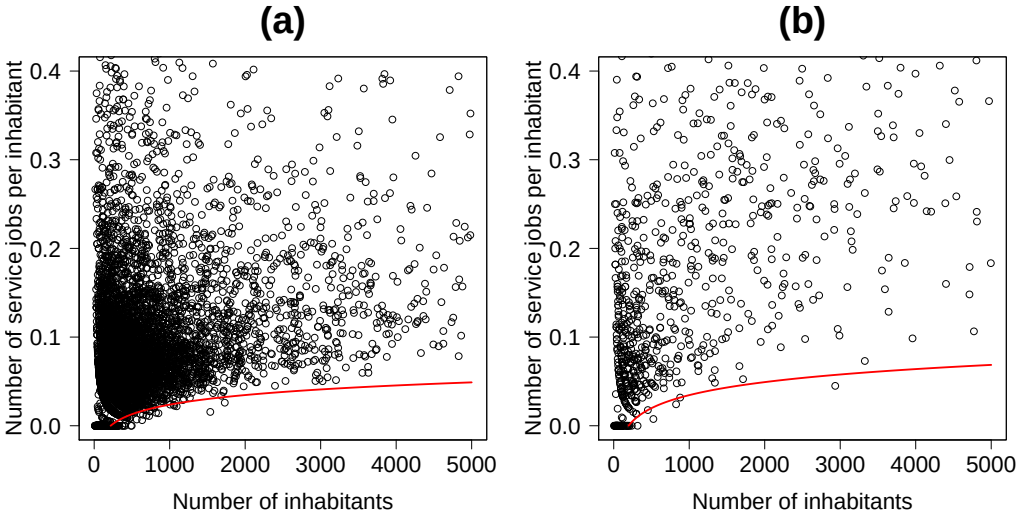


Figure 3.4: Number of service jobs per inhabitant function of the number of inhabitants for each municipality in 1999. The line represents the quantile regression line for $\tau = 0.01$. (a) tMFM $\in]0,5]$; (b) tMFM >30.

average daily moving time is 16 minutes in 1994 and 17 minutes in 2008 in France (for those moving by car).

Finally, within municipalities of 3000 inhabitants, the ones which are tMFM >30 have about 0.02 proximity job services per inhabitant more than municipalities close to MFM (tMFM <15), while this difference is about 0.005 within muncipalities of 500 inhabitants. This suggests that the same population changes in municipalities of 3000 inhabitants, have a significantly higher impact on the proximity service jobs in municipalities far from MFM than in municipalities close to MFM. In municipalities of 500 inhabitants tMFM seems to have a weaker impact.

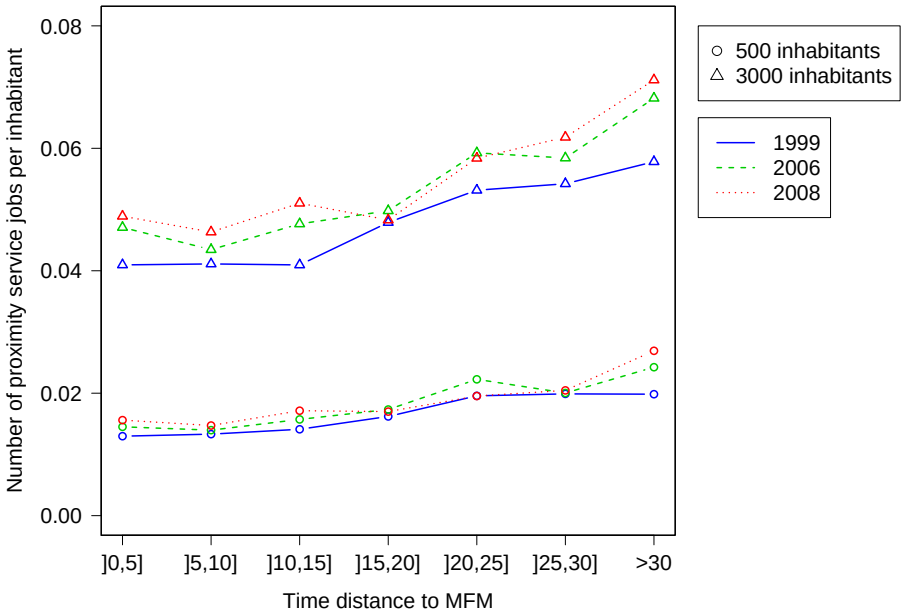


Figure 3.5: Number of proximity service jobs per inhabitant function of tMFM interval (min.) (tMFM $\in]0,5]$, tMFM $\in]5,10]$, tMFM $\in]10,15]$, tMFM $\in]15,20]$, tMFM $\in]20,25]$, tMFM $\in]25,30]$ and tMFM > 30). Blue solid line for 1999; Green dashed line for 2006; Red dotted line for 2008. Circles: municipality of 500 inhabitants; Triangles: municipality of 3000 inhabitants.

3.4 Discussion

We choose the minimum requirement approach for deriving the number of proximity services jobs per inhabitant in French rural municipalities, because it seems reasonable that, in a sufficiently large set of municipalities, some of them have only service jobs for the municipality population itself. Indeed, one can postulate that the long range services are located only in some privileged municipalities. However, we had to adapt the minimum requirement to our problem on three aspects:

- Instead of considering the share of jobs in a given activity, we considered the number of jobs per inhabitant. This corresponds better to our assumption that the proximity service jobs depend on the local population.

- We performed a series of minimum requirement procedures, corresponding to intervals of time distance to the most frequented municipality.

- Instead of using a discrete model based on a clustering of the municipalities by sizes as in the usual minimum requirement approach, we use a quantile regression (Koenker and Bassett, 1978) with as quantile value the first-percentile ($\tau = 0.01$).

The model yields accurate predictions of the first percentile. It suggests that big municipalities (close to 5000 inhabitants) generate (or destroy) significantly more proximity service jobs than small ones (around 500 inhabitants), for the same growth (or decline) of their population. Moreover, the impact of the time to the most frequented muncipality (MFM) corresponds to one's expectations: The municipalities which are close to a MFM have the lowest number of jobs in proximity services per inhabitant, and when the municipality gets farther from the MFM, its number of jobs in proximity services per inhabitant increases. Finally, this impact of tMFM on the number of proximity service jobs per inhabitant is significantly higher on big municipalities than on small ones.

We believe that such results can be interesting for policy makers, who have to make choices for distributing incentives to maintain employment and population in some rural areas. According to our results, the policies will have higher leverage effects in the big municipalities of our sample, especially the one with tMFM > 30. Moreover, our results suggest that in municipalities which are close to MFM, the population changes are likely to impact also the service jobs in the MFM.

Acknowledgments

We wish to thank Olivier Aznar and Solenn Tanguy for their help and Alexandre Kych from the Maurice Halbwachs Center for his kindness.

References

Brodsky, H. and Sarfaty, D. E. (1977). Measuring the urban economic base in a developing country. *Land Economics*, 53:445–454.

Brown, D. G. and Robinson, D. T. (2006). Effects of Heterogeneity in Residential Preferences on an Agent-Based Model of Urban Sprawl. *Ecology And Society*, 11(1):46.

Dissart, J.-C., Aubert, F., and Truchet, S. (2009). An estimation of tourism dependence in French rural areas. In Matias A., Sarmento M., N. P., editor, *Advances in Modern Tourism Research II*, chapter 17. Springer.

English, D., Marcouiller, D., and Cordell, H. (2000). Tourism dependence in rural America: Estimates and effects. *Society & Natural Resources*, 13(3):185–202.

Hubert, J. P. (2009). Dans les grandes agglomérations, la mobilité quotidienne des habitants diminue, et elle augmente ailleurs. *Insee première*, (1252).

Huet, S. and Deffuant, G. (2011). An Abstract Modelling Framework implemented through a Data-Driven approach to study the Impact of Policies at the Municipality level. *ESSA 2011 Conference, september 2011*, page 22.

Klosterman, R. (1990). *Community analysis and planning techniques*. Rowman & Littlefield.

Koenker, R. and Bassett, G. J. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.

Leatherman, J. C. and Marcouiller, D. W. (1996). Estimating tourism's share of local income from secondary data sources. *Review of Regional Studies*, 26(3):x5–339.

Moore, C. L. (1975). A New Look at the Minimum Requirements Approach to Regional Economic Analysis. *Economic Geography*, 51(4):350–356.

Mordier, B. (2010). Les services marchands aux particuliers s'implantent dans l'espace rural. *Insee première*, (1307).

Persky, J. and Wiewel, W. (1994). The growing localness of the global city. *Economic Geography*, 70(2):129–143.

Rutland, T. and O'Hagan, S. (2007). The growing localness of the Canadian City, or, on the continued (ir)relevance of economic base theory. *Local Economy*, 22(2):163–185.

Ullman, E. and Dacey, M. (1960). The Minimum Requirement Approach to the Urban Economic Base. *Papers and Proceedings of the Regional Science Assn.*, 6:192.

Woller, G. and Parsons, R. (2002). Assessing the community economic impact of nongovernmental development organizations. *Nonprofit and Voluntary Sector Quarterly*, 31(3):419–428.

CHAPTER 4

Adaptive Approximate Bayesian Computation for Complex Models

Contents

Abstract. We propose a new approximate Bayesian computation (ABC) algorithm that aims at minimizing the number of model runs for reaching a given quality of the posterior approximation. This algorithm automatically determines its sequence of tolerance levels and makes use of an easily interpretable stopping criterion. Moreover, it avoids the problem of particle duplication found when using a MCMC kernel. When applied to a toy example and to a complex social model, our algorithm is 2 to 8 times faster than the three main sequential ABC algorithms currently available.

Manuscript:
Lenormand, M. et Deffuant, G. (2012). Adaptive Approximate Bayesian Computation for Complex Models. *arXiv:1111.1308v3* (Submitted in *Computational Statistics*).

4.1 Introduction

Approximate Bayesian computation (ABC) techniques appear particularly relevant for calibrating stochastic models because their very principle includes stochasticity but they are applicable to any model. They generate a sample of model parameter values $(\theta_i)_{i=1,..,N}$ (often also called particles) from the prior distribution $\pi(\theta)$ and select the θ_i values leading to model outputs $x \sim f(x|\theta_i)$ satisfying a proximity criterion with the target data y ($\rho(S(x),S(y)) \leq \epsilon$, $\rho(\cdot)$ expressing a distance, $S(\cdot)$ expressing a summary statistic and ϵ being a tolerance level). The selected sample of parameter values approximates the posterior distribution of parameters, leading to model outputs with the expected quality of approximation. However, in practise, running these techniques is very demanding computationally because sampling the whole space of parameters requires a number of simulations which grows exponentially with the number of parameters to identify. This tends to limit the application of these techniques to easily computable models (Beaumont, 2010). In this paper, our goal is minimizing the number of model runs for reaching a given quality of posterior approximation, and thus to make the approach applicable to a larger set of models.

ABC is the subject of intense scientific researches and several improved versions of the original scheme are available, such as using local regressions to improve parameter inference (Beaumont et al., 2002; Blum and François, 2010), automatically selecting informative summary statistics (Joyce and Marjoram, 2008; Fearnhead and Prangle, 2012), coupling to Markov chain Monte Carlo (Marjoram et al., 2003; Wegmann et al., 2009) or improving sequentially the posterior distributions with sequential Monte Carlo methods (Sisson et al., 2007; Toni et al., 2009; Beaumont et al., 2009). This last class of methods approximates progressively the posterior, using sequential samples $S^{(t)} = (\theta_i^{(t)})_{i=1,..,N}$ derived from sample $S^{(t-1)}$, and using a decreasing set of tolerance levels $\{\epsilon_1,...,\epsilon_T\}$. This strategy focuses the sampling effort in parts of the parameter space of high likelihood, avoiding to spend much computing time in systematically sampling the whole parameter space.

The first sequential method applied to ABC was proposed by Sisson et al. (2007) with the ABC-PRC (Partial Rejection Control). This method is based on a theoritical work of Del Moral et al. (2006) to ABC. However, Beaumont et al. (2009) has shown that this method leads to a bias in the approximation of the posterior. Beaumont et al. (2009); Toni et al. (2009) proposed a new algorithm, called Population Monte Carlo ABC in Beaumont et al. (2009) and hereafter called PMC. This algorithm, corrects the bias by affecting to each particle a weight corresponding to the inverse of its importance in the sample. It is particularly interesting in our perspective because it provides with a rigorous framework to the sequential sample idea, which seems a good way for minimizing the number of runs. In this approach, the problem is then defining the sequence of tolerance levels $\{\epsilon_1,...,\epsilon_T\}$. Drovandi and Pettitt (2011) and Del Moral et al. (2012) solve partly this problem by deriving the tolerance level at a given step from values of $\rho(S(x),S(y))$ of the previously selected sample. However, a difficulty remains: when to stop? If the final tolerance level ϵ_T is too large, the final posterior will be of bad quality. Inversely, a too small ϵ_T leads to a posterior that could have been obtained with less

model runs.

Moreover, the MCMC kernel used in Drovandi and Pettitt (2011) and Del Moral et al. (2012) to sample new values $\theta_j^{(t)}$, despite its mathematical elegance, has a significant drawback in our view: it can lead to particle duplications. Indeed, each time the MCMC jumps from a particle to a new one which is not accepted, the initial particle is kept in the new sample of particles, hence when this occurs several times with the same initial particle, this particle appears several times in the new sample. As long as these duplications are very few compared with the size of the sample, their effect can be neglected, but it can easily happen that their number grows to a very significant part of the sample, then strongly deteriorating the quality of the posterior, as illustrated below. To solve this problem, Drovandi and Pettitt (2011) proposed to perform R MCMC jump trials instead of one, while Del Moral et al. (2012) proposed to resample the parameter values when too many are duplicated. Del Moral et al. (2012) also proposed to run the model M times for each particle, in order to decrease the variance of the acceptance ratio of the MCMC jump. However, these solutions increase the number of model runs, going against the initial benefit of using sequential samples.

In this paper, we propose a modification of the PMC algorithm that we call adaptive population Monte Carlo ABC (hereafter called APMC). This new algorithm determines by itself the sequence of tolerance levels as in Drovandi and Pettitt (2011) and Del Moral et al. (2012), and it also provides a stopping criterion. Furthermore, our approach avoids the problem of duplications. We prove that the computation of the weights associated to the particles in this algorithm lead to the intended posterior distribution and we also prove that the algorithm stops whatever the chosen value of the stopping parameter. We show that our algorithm, applied to a toy example and to an individual-based social model, requires significantly less simulations to reach a given quality level of the posterior distribution than the PMC algorithm of Beaumont et al. (2009), the replenishment SMC ABC algorithm of Drovandi and Pettitt (2011) (hereafter called RSMC) and the adaptive SMC ABC algorithm of Del Moral et al. (2012) (hereafter called SMC). These algorithms are detailed in Appendix 4.A.

4.2 Adaptive population Monte-Carlo approximate Bayesian computation

4.2.1 Overview of the APMC algorithm

The APMC algorithm follows the main principles of the sequential ABC, and defines on-line the tolerance level at each step like in Drovandi and Pettitt (2011) and Del Moral et al. (2012). For each tolerance level ϵ_t, it generates a sample $S^{(t)}$ of particles and computes their associated weights. This weighted sample approximates the posterior distribution, with an increasing approximation quality as ϵ_t decreases. We say that a parameter value $\theta_i^{(t)}$, satisfies the tolerance level ϵ_t, if when running the model we get $x \sim f(x|\theta_i^{(t)})$, such that its distance $\rho_i^{(t)} = \rho(S(x), S(y))$ to the target data y, is below ϵ_t. Suppose the APMC reached step $t-1$, with a sample $S^{(t-1)}$ of $N_\alpha = \lfloor \alpha N \rfloor$ particles and

their associated weights $(\theta_i^{(t-1)}, w_i^{(t-1)})_{i=1,..,N_\alpha}$, the main features of the APMC are (see Algorithm 4.6 for details):

- the algorithm generates $N - N_\alpha$ particles $(\theta_j^{(t-1)})_{j=N_\alpha+1,..,N}$ where $\theta_j^{(t-1)} \sim \mathcal{N}(\theta_j^*, \sigma^2_{(t-1)})$, the seed θ_j^* is randomly drawn from the weighted set $(\theta_i^{(t-1)}, w_i^{(t-1)})_{i=1,..,N_\alpha}$ and the variance $\sigma^2_{(t-1)}$ of the Gaussian kernel $\mathcal{N}(\theta_j^*, \sigma^2_{(t-1)})$ is twice the empirical variance of the weighted set $(\theta_i^{(t-1)}, w_i^{(t-1)})_{i=1,..,N_\alpha}$, following Beaumont et al. (2009).

- the weights $w_j^{(t-1)}$ of the new particles $(\theta_j^{(t-1)})_{j=N_\alpha+1,..,N}$ are computed so that these new particles can be combined with the sample $S^{(t-1)}$ of the previous step without causing a bias in the posterior distribution. These weights are given by Equation 4.2 (see below).

- the algorithm concatenates the N_α previous particles $(\theta_i^{(t-1)})_{i=1,..,N_\alpha}$ with the $N - N_\alpha$ new particles $(\theta_j^{(t-1)})_{j=N_\alpha+1,..,N}$, together with their associated weights and distances to the data. This constitutes a new set noted $S^{(t)}_{temp} = (\theta_i^{(t)}, w_i^{(t)}, \rho_i^{(t)})_{i=1,..,N}$.

- the next tolerance level ϵ_t is determined as the first $\alpha-$quantile of the $(\rho_i^{(t)})_{i=1,..,N}$.

- the new sample $S^{(t)} = (\theta_i^{(t)}, w_i^{(t)})_{i=1,..,N_\alpha}$ is then constituted from the N_α particles of $S^{(t)}_{temp}$ satisfying the tolerance level ϵ_t.

- if the proportion p_{acc} of particles satisfying the tolerance level ϵ_{t-1} among the $N - N_\alpha$ newly generated particles is below a chosen value $p_{acc_{min}}$, the algorithm stops, and its result is $(\theta_i^{(t)})_{i=1,..,N_\alpha}$ with their associated weights.

Note that in our algorithm, to get a number N_α of retained particles for the next step, the choice of ϵ_t is heavily constrained: it has to be at least equal to the first $\alpha-$quantile of the $(\rho_i^{(t)})_{i=1,..,N}$ and smaller than the immediately superior $(\rho_i^{(t)})$ value. We chose to fix it to the first $\alpha-$quantile for simplicity. This choice also ensures that the tolerance level decreases from one iteration to the next: in the worst case where $p_{acc} = 0$ (no newly simulated particles accepted), $\epsilon_t = \epsilon_{t-1}$. Our algorithm does not use a MCMC kernel and avoids duplicating particles. It requires a reweighting step in $O(N_\alpha^2)$ instead of $O(N_\alpha)$ in Drovandi and Pettitt (2011), but in our perspective, this computational cost is supposed negligible compared with the cost of running the model.

4.2.2 Weights correcting the kernel sampling bias

As pointed out by Beaumont et al. (2009), the newly generated particles $\theta_i^{(t)}$ in a sequential procedure are no more drawn from the prior distribution but from a specific probability density $d_i^{(t)}$ that depends on the particles selected at the previous step and on the chosen kernel. This introduces a bias in the procedure. This bias should be corrected by attributing a weight equal to $\pi(\theta_i^{(t)})/d_i^{(t)}$ to each newly generated particle $\theta_i^{(t)}$.

The density of probability $d_i^{(t)}$ to generate particle $\theta_i^{(t)}$ at step t is given by the sum of the probabilities to reach $\theta_i^{(t)}$ from one of the N_α particles of the previous step times their respective weights:

$$d_i^{(t)} = \sum_{j=1}^{N_\alpha} \frac{w_j^{(t-1)}}{\sum_{k=1}^{N_\alpha} w_k^{(t-1)}} \sigma_{t-1}^{-1} \varphi\left(\sigma_{t-1}^{-1}(\theta_i^{(t)} - \theta_j^{(t-1)})\right) \tag{4.1}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the kernel function.

This yields the expression of the weight $w_i^{(t)}$ to be attributed to the newly drawn particle $\theta_i^{(t)}$:

$$w_i^{(t)} = \frac{\pi(\theta_i^{(t)})}{\sum_{j=1}^{N_\alpha} \left(w_j^{(t-1)} / \sum_{k=1}^{N_\alpha} w_k^{(t-1)}\right) \sigma_{t-1}^{-1} \varphi\left(\sigma_{t-1}^{-1}(\theta_i^{(t)} - \theta_j^{(t-1)})\right)} \tag{4.2}$$

This formula differs from the scheme of Beaumont et al. (2009) where the weights need only to be proportional to Equation 4.2 at each step. Since we want to concatenate particles obtained at different steps of the algorithm (while Beaumont et al. (2009) generate the sample at step t from scratch), we need the scaling of weights to be consistent across the different steps of the algorithm. Using the weight of Equation 4.2 guarantees the correction of the sampling bias throughout the APMC procedure and ensures that the N_α weighted particles $\theta_i^{(t)}$ produced at the t-th iteration follow the posterior distribution $\pi\left(\theta | \rho(S(x), S(y)) < \epsilon_t\right)$.

4.2.3 The stopping criterion

We stop the algorithm when the proportion of "accepted" particles (Equation 4.3) among the $N - N_\alpha$ new particles is below a predetermined threshold $p_{acc_{min}}$. This choice of stopping rule ensures that additional simulations would only marginally change the posterior distribution. Note that this stopping criterion will be achieved even if $p_{acc_{min}} = 0$, this ensures that the algorithm converges. We present a formal proof of this assertion in Appendix 4.B.

$$p_{acc}(t) = \frac{1}{N - N_\alpha} \sum_{k=N_\alpha+1}^{N} \mathbb{1}_{\rho_k^{(t-1)} < \epsilon_{t-1}} \tag{4.3}$$

4.3 Experiments on a toy example

We consider four algorithms: APMC, PMC, the SMC and the RSMC. Their implementations in R (R Development Core Team, 2011) are available [1]. We compare them on the toy example studied in Sisson et al. (2007) where $\pi(\theta) = \mathcal{U}_{[-10,10]}$ and $f(x|\theta) \sim \frac{1}{2}\phi\left(\theta, \frac{1}{100}\right) + \frac{1}{2}\phi(\theta, 1)$ where $\phi\left(\mu, \sigma^2\right)$ is the normal density of mean μ and variance

[1] http://motive.cemagref.fr/people/maxime.lenormand/script_r_toyex

σ^2. In this example, we consider that $y = 0$ is observed, so that the posterior density of interest is proportional to $\left(\phi\left(0, \frac{1}{100}\right) + \phi(0,1)\right)\pi(\theta)$.

We structure the comparisons on two indicators: the number of simulations performed during the application of the algorithms, and the $\mathbb{L}_2$ distance between the exact posterior density and the histogram of particle values obtained with the algorithms. This $\mathbb{L}_2$ distance is computed on the 300-tuple obtained by dividing the support $[-10, 10]$ into 300 equally-sized bins.

We choose $N = 5000$ particles and a target tolerance level equal to 0.01. For the PMC algorithm we use a decreasing sequence of tolerance levels from $\epsilon_1 = 2$ down to $\epsilon_{11} = 0.01$. For the SMC algorithm, we use 3 different values for α: $\{0.9, 0.95, 0.99\}$ and $M = 1$ as in Del Moral et al. (2012). For the RSMC algorithm we use $\alpha = 0.5$ as in Drovandi and Pettitt (2011). To explore our algorithm, we test 9 different values for α: $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and 4 different values for $p_{acc_{min}}$: $\{0.01, 0.05, 0.1, 0.2\}$. In each case, we perform 50 times the algorithm, and compute the average and standard deviation of the two indicators: the total number of simulations and the $\mathbb{L}_2$ distance between the exact posterior density and the histogram of particle values. We used as kernel transition a normal distribution parameterized with twice the weighted variance of the previous sample, as in Beaumont et al. (2009).

We report below the effects of varying α and $p_{acc_{min}}$ on the performance of our algorithm, and compare it with the PMC, SMC and RSMC algorithms.

4.3.1 Particle duplication in SMC and RSMC

The number of distinct particles decreases during the course of the SMC algorithm whatever the value of α, as shown on Figure 4.1a-c. The oscillations of the number of distinct particles are caused by the resampling step in the SMC algorithm (see Del Moral et al. (2012)), but they are not sufficient to counterbalance the overall decrease. This decrease deteriorates the posterior approximation as shown on Figure 4.2. For the RSMC algorithm, the number of distinct particles is maintained at a reasonably high level (Figure 4.1c), but this has a cost in terms of the number of required model runs (see Figure 4.2). Note that the APMC and the PMC algorithms keep N distinct particles.

4.3.2 Influence of parameters on APMC

The values of α and $p_{acc_{min}}$ have an impact on the studied indicators. We find that smaller α and $p_{acc_{min}}$ improve the quality of the approximation (smaller $\mathbb{L}_2$ distance), and increase the total number of model runs, with $p_{acc_{min}}$ having the largest effect (Figure 4.2). With a large α, the tolerance levels decrease slowly and there are numerous steps before the algorithm stops. In this toy example, our simulations show that all explored sets of $(\alpha\ ,\ p_{acc_{min}})$ such that $p_{acc_{min}} < 0.1$ give good results for the criterion *Number of simulations* $\times\ \mathbb{L}_2^2$ (Figure 4.3b). Large α provide slightly better results for small $p_{acc_{min}}$ while small α provide slightly better results for large $p_{acc_{min}}$ (Figure 4.3b). On this toy example it appears that intermediate values of α and $p_{acc_{min}}$ ($0.3 \leq \alpha \leq 0.7$ and $0.01 \leq p_{acc_{min}} \leq 0.05$), present a good compromise between number of model runs

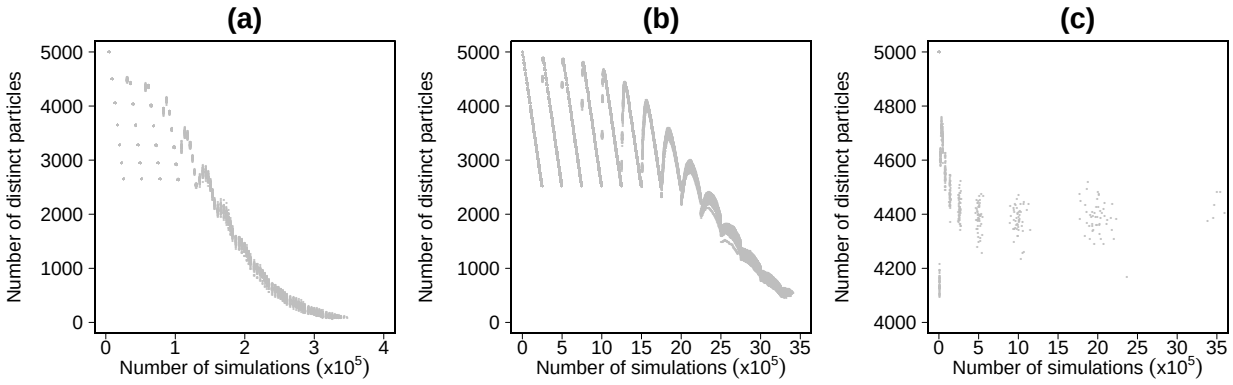


Figure 4.1: Number of distinct particles in a sample of $N = 5000$ particles during the course of the SMC and RSMC algorithms applied to the toy example. (a) SMC with $\alpha = 0.9$ and $M = 1$; (b) SMC with $\alpha = 0.99$ and $M = 1$; (c) RSMC with $\alpha = 0.5$. In all three panels, the tolerance target is equal to 0.001.

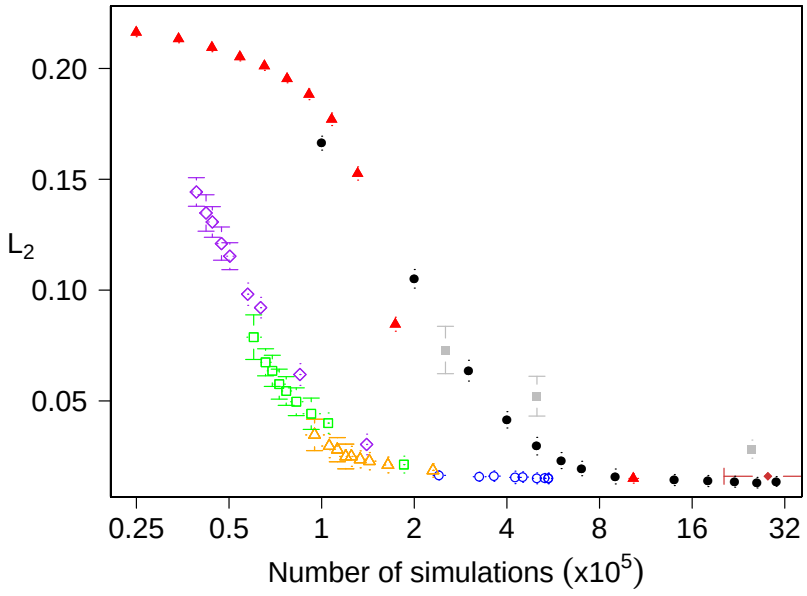


Figure 4.2: Posterior quality ($\mathbb{L}_2$) versus computing cost (number of simulations) averaged over 50 replicates. Vertical and horizontal bars represent the standard deviations among replicates. Algorithm parameters used for APMC: α in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $p_{acc_{min}}$ in $\{0.01, 0.05, 0.1, 0.2\}$. Blue circles are used for $p_{acc_{min}} = 0.01$, orange triangles for $p_{acc_{min}} = 0.05$, green squares for $p_{acc_{min}} = 0.1$, and purple diamonds for $p_{acc_{min}} = 0.2$. PMC: red plain triangles for a sequence of tolerance levels from $\epsilon_1 = 2$ down to $\epsilon_{11} = 0.01$. SMC: grey plain square for α in $\{0.9, 0.95, 0.99\}$ (from left to right), $M = 1$ and a ϵ target equal to 0.01. RSMC: brown plain diamond for $\alpha = 0.5$ and a ϵ target equal to 0.01. Results obtained with a standard rejection-based ABC algorithm are depicted with black plain circles.

and the quality of the posterior approximation.

4.3.3 Comparing performances

Whatever the value of α and $p_{acc_{min}}$, the APMC algorithm always yields better results than the other three algorithms. It requires between 2 and 8 times less simulations to reach a given posterior quality $\mathbb{L}_2$ (Figure 4.2). Furthermore, good approximate posterior distributions are very quickly obtained (Figure 4.2). The compromise between simulation speed and convergence level can also be illustrated using the criterion *Number of simulations* $\times$ $\mathbb{L}_2^2$ (Glynn and Whitt, 1992). This criterion is smaller for the APMC algorithm (Figure 4.3a).

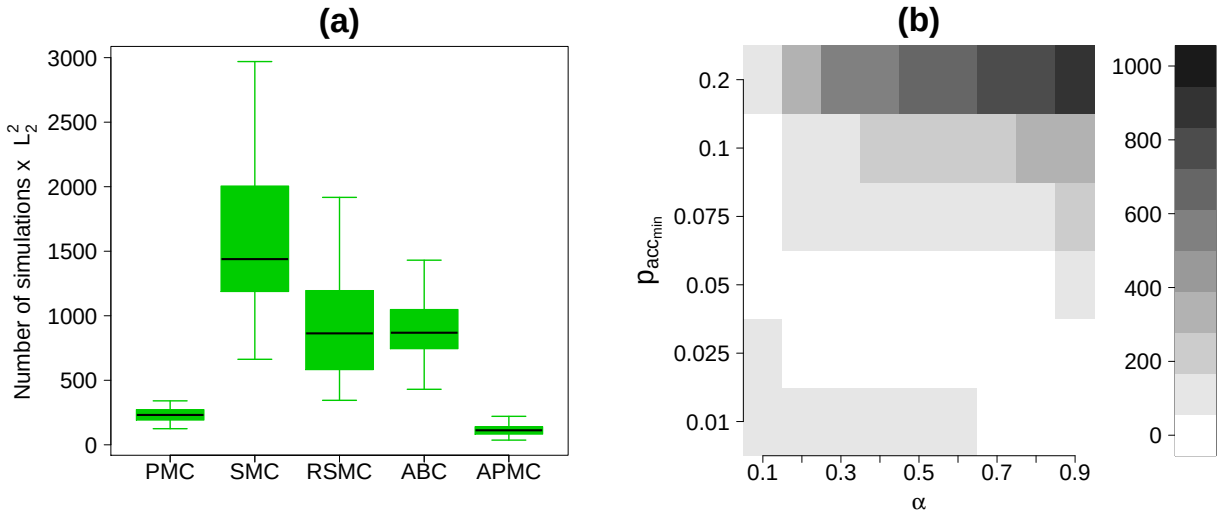


Figure 4.3: (a) Boxplot of the criterion "squared $\mathbb{L}_2$ distance times the number of simulations" for the different ABC algorithms. APMC: for α in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $p_{acc_{min}} = 0.01$; SMC: for α in $\{0.9, 0.95, 0.99\}$, $M = 1$ and a ϵ target equal to 0.01; RSMC: for $\alpha = 0.5$ and a ϵ target equal to 0.01; ABC: for a ϵ target equal to 0.01; PMC: for a sequence of tolerance levels from $\epsilon_1 = 2$ to $\epsilon_{11} = 0.01$. (b) Criterion "squared $\mathbb{L}_2$ distance times the number of simulations" in the APMC algorithm for the different values of α and $p_{acc_{min}}$. Each cell depicts the average of the criterion over the 50 performed replicates of the APMC.

4.4 Application to the model *SimVillages*

In this section, we check if our algorithm still performs better than the PMC, the RSMC and the SMC when applied to an individual-based social model developed during the European project PRIMA[2]. The aim of the model is to simulate the effect of a

[2] PRototypical policy Impacts on Multifunctional Activities in rural municipalities - EU 7th Framework Research Programme; 2008-2011; https://prima.cemagref.fr/the-project

scenario of job creation (or destruction) on the evolution of the population and activities in a network of municipalities.

4.4.1 Model and data

The model simulates the dynamics of virtual individuals living in seven interconnected villages in a rural area of Auvergne (a region of Central France). A single run of the model *SimVillages* with seven rural municipalities takes about 1.4 seconds on a desktop machine (PC Intel 2.83 GHz). The dynamics include demographic change (aging, marriage, divorce, births and deaths), activity change (change of jobs, unemployment, inactivity, retirement), and movings from one municipality to another or outside of the set. The model also includes a dynamics of creation / destruction of jobs of proximity services, derived from the size of the local population. More details on the model can be found in Huet et al. (2012). The individuals (about 3000) are initially generated using the 1990 census data of the National Institute of Statistics and Economic Studies (INSEE), some of them are given a job type and a location for this job (in a municipality of the set or outside), they are organised in households living in a municipality of the set. The model dynamics is mostly data driven, but four parameters cannot be directly derived from the available data. They are noted θ_p for $1 \leq p \leq 4$, described in Table 4.1.

We use our algorithm to identify the distribution of the four parameters for which the simulations, initialized with the 1990 census data, satisfy matching criteria with the data of the 1999 and 2006 census. The set of summary statistics $\{S_m\}_{1\leq m\leq M}$ and the associated discrepancy measure used ρ_m are described in Table 4.2. We note S_m the simulated summary statistics and $S_m^{'}$ the observed statistics. The eight summary statistics are normalized (variance equalization) and they are combined using the infinity norm (Equation 4.4):

$$\|(\rho_m(S_m, S_m^{'}))_{1\leq m\leq M}\|_\infty = \sup_{1\leq m\leq M} \rho_m(S_m, S_m^{'}) \tag{4.4}$$

We first generate a sample of length N from the prior $\mathcal{U}_{[a,b]}$, where $[a,b]$ is available for each parameter in Table 4.1, with a Latin hypercube (Carnell, 2009) and we select the best N_α particles. To move the particles, we use as kernel transition a multivariate normal distribution parameterized with twice the weighted variance-covariance matrix of the previous sample (Filippi et al., 2011).

As in the Section 4.3, we perform a parameter study and compare APMC with its three competitors. For APMC, α varies in $(\{0.3, 0.5, 0.7\})$ and $p_{acc_{min}}$ in $(\{0.01, 0.05, 0.1, 0.2\})$, and we set $N_\alpha = 5000$ particles. For the PMC, SMC and RSMC we also set $N = 5000$ particles and a tolerance level target equal to 1.4. The tolerance value $\epsilon = 1.4$ corresponds to the average final tolerance value we obtain with APMC for $p_{acc_{min}} = 0.01$. Note that otherwise this final tolerance is difficult to set properly and a worse choice for this value would have lead to worse performances of these algorithms. For the PMC algorithm, we use the decreasing sequence of tolerance levels $\{3, 2.5, 2, 1.7, 1.4\}$. For the SMC algorithm, we use 3 different values for the couple (α, M): $\{(0.9, 1), (0.99, 1), (0.9, 15)\}$. For the RSMC algorithm we use $\alpha = 0.5$, as in Drovandi and Pettitt (2011). For each algorithm and parameter setting, we perform 5 replicates.

We approximated posterior density (unknown in this case) with the original rejection-based ABC algorithm, starting with $N = 10,000,000$, selecting 7890 particles below the tolerance level $\epsilon = 1.4$.

To compute the $\mathbb{L}_2$ distance between posterior densities, we divided each parameter support into 4 equally sized bins, leading to a grid of $4^4 = 256$ cells, and we computed on this grid the sum of the squared differences between histogram values.

Table 4.1: *SimVillages* parameter descriptions

Parameters	Description	Range
θ_1	Average number of children per woman	$[0,4]$
θ_2	Probability to accept a new residence for a household	$[0,1]$
θ_3	Probability to make couple for two individuals	$[0,1]$
θ_4	Probability to split for a couple in a year	$[0,0.5]$

Table 4.2: Summary statistic descriptions

Summary statistic	Description	Measure of discrepancy
S_1	Number of inhabitants in 1999	$\mathbb{L}_1$ distance
S_2	Age distribution in 1999	χ^2 distance
S_3	Household type distribution in 1999	χ^2 distance
S_4	Net migration in 1999	$\mathbb{L}_1$ distance
S_5	Number of inhabitants in 2006	$\mathbb{L}_1$ distance
S_6	Age distribution in 2006	χ^2 distance
S_7	Household type distribution in 2006	χ^2 distance
S_8	Net migration in 2006	$\mathbb{L}_1$ distance

4.4.2 Study of APMC result

APMC yields a unimodal approximate posterior distribution for the model *SimVillages* (Figure 4.4). Interestingly, parameters θ_1 and θ_4 are slightly correlated (Figure 4.4c). This is logical since they have contradictory effects on the number of child in the population. What is less straightforward is that we are able to partly tease apart these two effects with the available census data, since we get a peak in the approximate posterior distribution instead of a ridge.

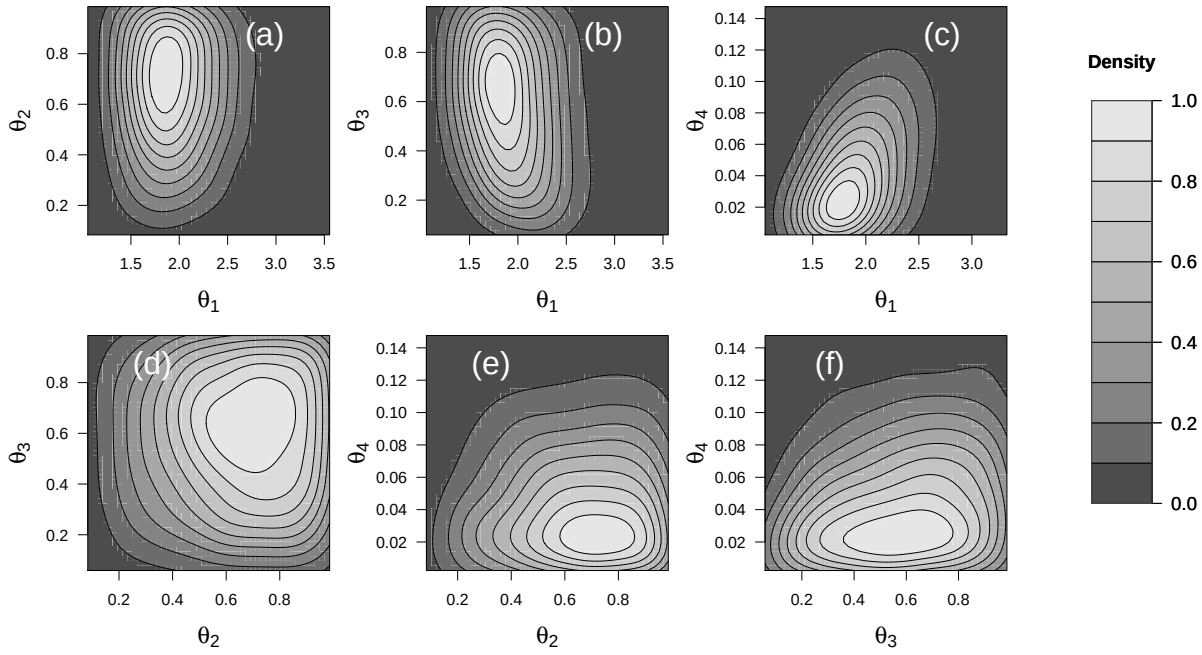


Figure 4.4: Contour plot of the bivariate joint densities of θ_i and θ_j obtained with our algorithm, and with $\alpha = 0.5$ and $p_{acc_{min}} = 0.01$; (a) θ_1 and θ_2; (b) θ_1 and θ_3; (c) θ_1 and θ_4; (d) θ_2 and θ_3; (e) θ_2 and θ_4; (f) θ_3 and θ_4.

4.4.3 Influence of parameters on APMC

As for the toy example, we find that the intermediate values of $(\alpha, p_{acc_{min}})$ that we used lead to similar results (Figure 4.5c). In practice, we therefore recommend to use $\alpha = 0.5$ and $p_{acc_{min}}$ between 0.01 and 0.05 depending on the wished level of convergence.

4.4.4 Comparing performances

APMC requires between 2 and 7 times less simulations to reach a given posterior quality than the other algorithms $\mathbb{L}_2$ (Figure 4.5a). Again, the gain in simulation number is progressive during the course of the algorithm. The *Number of simulations* $\times$ $\mathbb{L}_2^2$ criterion is again smaller for the APMC algorithm (Figure 4.5b).

4.5 Discussion

The good performances of APMC should of course be confirmed on other examples. Nevertheless we argue that they are due to the main assets of our approach:

- We choose an appropriate reweighting process instead of a MCMC kernel, which corrects the sampling bias without duplicating particles;
- We define an easy to interpret stopping criterion that automatically defines the number of sequential steps.

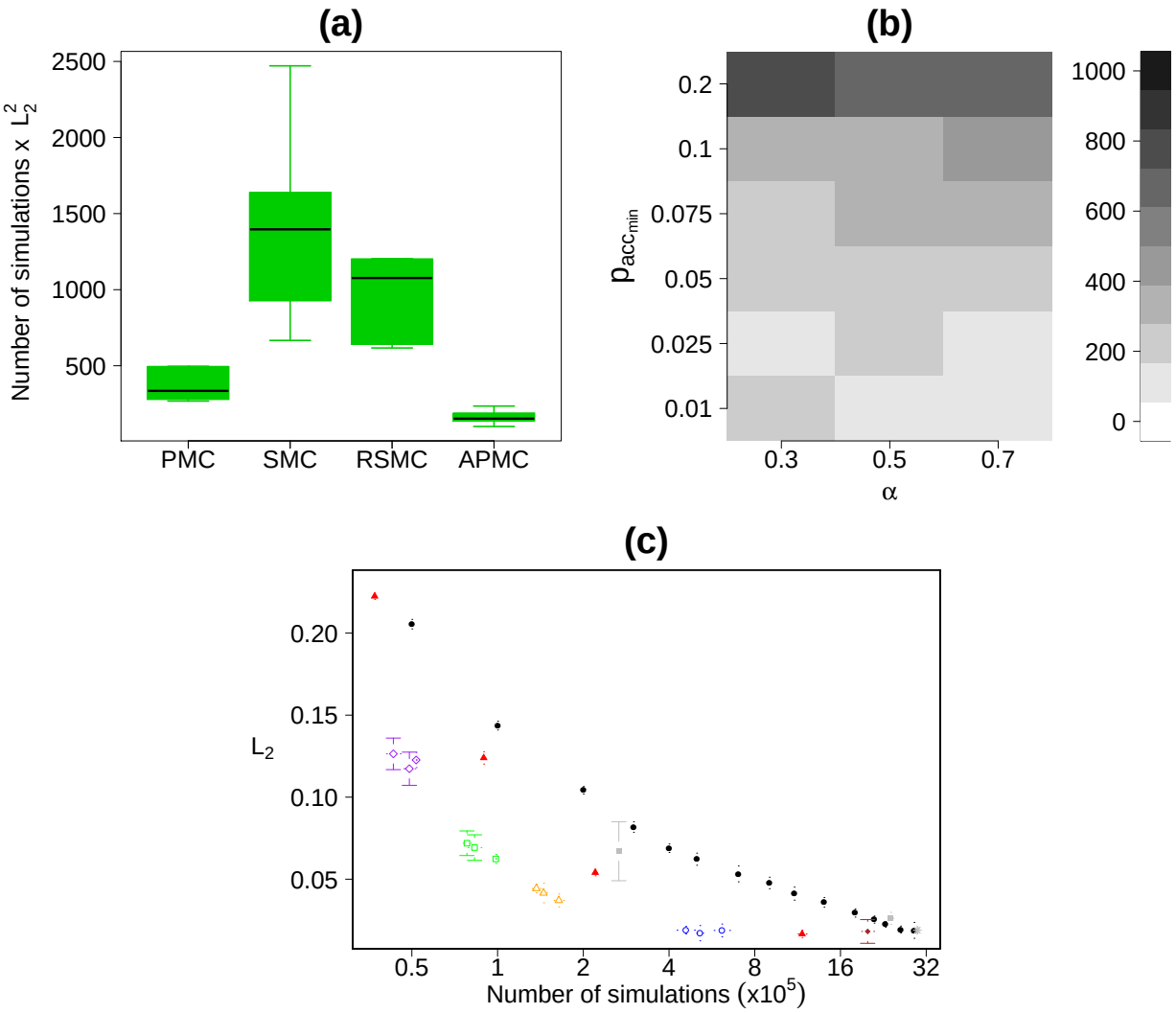


Figure 4.5: (a) Boxplot of the criterion "squared $\mathbb{L}_2$ distance times the number of simulations" for the different algorithms. APMC: for α in $\{0.3, 0.5, 0.7\}$ and $p_{acc_{min}} = 0.01$; SMC: for (α, M) in $\{(0.9,1),(0.99,1),(0.9,15)\}$ and a ϵ target equal to 0.01; RSMC: for $\alpha = 0.5$ and a ϵ target equal to 0.01; ABC: for a ϵ target equal to 1.4; PMC: for a sequence of tolerance levels from $\epsilon_1 = 3$ to $\epsilon_5 = 1.4$. (b) Criterion "squared $\mathbb{L}_2$ distance times the number of simulations" in the APMC algorithm for the different values of α and $p_{acc_{min}}$. Each cell depicts the average of the criterion over the 5 performed replicates of the APMC. (c) Posterior quality ($\mathbb{L}_2$) versus computing cost (number of simulations) averaged over 5 replicates. Vertical and horizontal bars represent the standard deviations among replicates. Algorithm parameters used for APMC: α in $\{0.3, 0.5, 0.7\}$ and $p_{acc_{min}}$ in $\{0.01, 0.05, 0.1, 0.2\}$. Blue circles are used for $p_{acc_{min}} = 0.01$, orange triangles for $p_{acc_{min}} = 0.05$, green squares for $p_{acc_{min}} = 0.1$, and purple diamonds for $p_{acc_{min}} = 0.2$. PMC: red plain triangles for a sequence of tolerance levels from $\epsilon_1 = 3$ to $\epsilon_5 = 1.4$. SMC: grey plain square for (α, M) in $\{(0.9,1),(0.99,1)\}$, grey star for $(\alpha, M) = (0.9, 15)$ and a ϵ target equal to 1.4. RSMC: brown plain diamond for $\alpha = 0.5$ and a ϵ target equal to 1.4. Results obtained with a standard rejection-based ABC algorithm are depicted with black plain circles.

Therefore, we can have some confidence in the good performances of APMC on other examples.

In the future, it would be interesting to evaluate this algorithm on models involving a larger number of parameters and/or multi-modal posterior distributions. Moreover, APMC could benefit from other improvements, in particular by performing a semi-automatic selection of informative summary statistics after the first ABC step (Joyce and Marjoram, 2008; Fearnhead and Prangle, 2012) and by using local regressions for post-processing the final posterior distribution (Beaumont et al., 2002; Blum and François, 2010). We did not perform such combinations in the present contribution, so that our algorithm is directly comparable with the three other sequential algorithms we looked at. However, they would be straightforward, because the different improvements concern different steps of the ABC procedure.

References

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Beaumont, M. A., Cornuet, J., Marin, J., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Blum, M. G. B. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73.

Carnell, R. (2009). lhs: Latin Hypercube Samples. *R package version 0.5*.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society.Series B: Statistical Methodology*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5):1009–1020.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation. *Biometrics*, 67(1):225–233.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society.Series B: Statistical Methodology*, 74(3):419–474.

Filippi, S., Barnes, C., and Stumpf, M. P. H. (2011). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *arXiv:1106.6280v3*.

Glynn, P. and Whitt, W. (1992). The Asymptotic Efficiency of Simulation Estimators. *Operations Research*, 40(3):505–520.

Huet, S., Dumoulin, N., Deffuant, G., Gargiulo, F., Lenormand, M., Baqueiro Espinosa, O., and Ternès, S. (2012). Micro-simulation model of municipality network in the Auvergne case study. Technical report, PRIMA Project, IRSTEA(Cemagref) LISC.

Joyce, P. and Marjoram, P. (2008). Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–15328.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1760–1765.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. *Genetics*, 182(4):1207–1218.

Appendix 4.A: Description of the algorithms

Algorithm 4.1 Likelihood-free rejection sampler 1

Given N the number of particles
for $i = 1$ to N **do**
  repeat
    Generate $\theta^* \sim \pi(\theta)$
    Simulate $x \sim f(x|\theta^*)$
  until $S(x) = S(y)$
  Set $\theta_i = \theta^*$
end for

Algorithm 4.2 Likelihood-free rejection sampler 2

Given N the number of particles
for $i = 1$ to N **do**
  repeat
    Generate $\theta^* \sim \pi(\theta)$
    Simulate $x \sim f(x|\theta^*)$
  until $\rho(S(x), S(y)) < \epsilon$
  Set $\theta_i = \theta^*$
end for

Algorithm 4.3 Population Monte Carlo ABC (PMC)

Given N the number of particles and a decreasing sequence of tolerance level $\epsilon_1 \geq .. \geq \epsilon_T$,
For $t = 1$,
for $i = 1$ à N **do**
  repeat
    Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $x \sim f(x|\theta_i^{(1)})$
  until $\rho(S(x), S(y)) < \epsilon_1$
  Set $w_i^{(1)} = \frac{1}{N}$
end for
Take σ_2^2 as twice the weighted empirical variance of $(\theta_i^{(1)})_{1 \leq i \leq N}$
for $t = 2$ to T **do**
  for $i = 1$ to N **do**
    repeat
      Sample θ_i^* from $\theta_j^{(t-1)}$ with probabilities $w_j^{(t-1)}$
      Generate $\theta_i^{(t)}|\theta_i^* \sim \mathcal{N}(\theta_i^*, \sigma_t^2)$ and $x \sim f(x|\theta_i^{(t)})$
    until $\rho(S(x), S(y)) < \epsilon_t$
    Set $w_i^{(t)} \propto \frac{\pi(\theta_i^{(t)})}{\sum_{j=1}^{N} w_j^{(t-1)} \sigma_t^{-1} \varphi(\sigma_t^{-1}(\theta_i^{(t)} - \theta_j^{(t-1)}))}$
  end for
  Take σ_{t+1}^2 as twice the weighted empirical variance of $(\theta_i^{(t)})_{1 \leq i \leq N}$
end for
Where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Algorithm 4.4 Sequential Monte Carlo ABC Replenishment (RSMC)

Given N, ϵ_1, ϵ_T, c, $\alpha \in [0,1]$ and $N_\alpha = \lfloor \alpha N \rfloor$,
for $i = 1$ to N **do**
  repeat
    Simulate $\theta_i \sim \pi(\theta)$ and $x \sim f(x|\theta_i)$
    $\rho_i = \rho(S(x), S(y))$
  until $\rho_i \leq \epsilon_1$
end for
Sort (θ_i, ρ_i) by ρ_i
Set $\epsilon_{MAX} = \rho_N$
while $\epsilon_{MAX} > \epsilon_T$ **do**
  Remove the N_α particles with largest ρ
  Set $\epsilon_{NEXT} = \rho_{N-N_\alpha}$
  Set $i_{acc} = 0$
  Compute the parameters of the proposal MCMC $q(\cdot,\cdot)$ with the $N - N_\alpha$ particles.
  for $j = 1$ to N_α **do**
    Simulate $\theta_{N-N_\alpha+j} \sim (\theta_i)_{1 \leq i \leq N-N_\alpha}$
    for $k = 1$ à R **do**
      Generate $\theta^* \sim q(\theta^*, \theta_{N-N_\alpha+j})$ et $x^* \sim f(x^*|\theta^*)$
      Generate $u < \mathscr{U}_{[0,1]}$
      if $u \leq 1 \wedge \dfrac{\pi(\theta^*) q(\theta_{N-N_\alpha+j}, \theta^*)}{\pi(\theta_{N-N_\alpha+j}) q(\theta^*, \theta_{N-N_\alpha+j})} \mathbb{1}_{\rho(S(x^*),S(y)) \leq \epsilon_{NEXT}}$ **then**
        Set $\theta_{N-N_\alpha+j} = \theta^*$
        Set $\rho_{N-N_\alpha+j} = \rho(S(x^*), S(y))$
        $i_{acc} \leftarrow i_{acc} + 1$
      end if
    end for
  end for
  Set $p_{acc} = \dfrac{i_{acc}}{R N_\alpha}$
  Set $R = \dfrac{\log(c)}{\log(1 - p_{acc})}$
end while

Algorithm 4.5 Adaptive Sequential Monte Carlo ABC (SMC)

Given N, M, $\alpha \in [0,1]$, $\epsilon_0 = \infty$, ϵ and N_T,
For $t = 0$,
for $i = 1$ to N **do**
  Simulate $\theta_i^{(0)} \sim \pi(\theta)$
  for $k = 1$ à M **do**
    Simulate $X_{(i,k)}^{(0)} \sim f(\cdot|\theta_i^{(0)})$
  end for
  Set $W_i^{(0)} = \dfrac{1}{N}$
end for
We have $ESS((W_i^{(0)}), \epsilon_0) = N$ where $ESS((W_i^{(0)}), \epsilon_0) = \left(\sum_{i=1}^{N} (W_i^{(0)})^2\right)^{-1}$
Set $t = 1$
while $\epsilon_{t-1} > \epsilon$ **do**
  Determine ϵ_t resolving $ESS((W_i^{(t)}), \epsilon_t) = \alpha ESS((W_i^{(t-1)}), \epsilon_{t-1})$ where
  $W_i^{(t)} \propto W_i^{(t-1)} \dfrac{\sum_{k=1}^{M} \mathbb{1}_{A_{\epsilon_{t-1},y}}(X_{(i,k)}^{(t-1)})}{\sum_{k=1}^{M} \mathbb{1}_{A_{\epsilon_{t-1},y}}(X_{(i,k)}^{(t-1)})}$ et $A_{\epsilon,y} = \{x | \rho(S(x), S(y)) < \epsilon\}$
  if $\epsilon_t < \epsilon$ **then**
    $\epsilon_n = \epsilon$
  end if
  if $ESS((W_i^{(t)}), \epsilon_t) < N_T$ **then**
    for $i = 1$ to N **do**
      Simulate $(\theta_{(i)}^{(t-1)}, X_{(i,1:M)}^{(t-1)})$ in $(\theta_{(j)}^{(t-1)}, X_{(j,1:M)}^{(t-1)})$ with probabilities $W_j^{(t)}$, $1 \leq j \leq N$
      Set $W_i^{(t)} = \frac{1}{N}$
    end for
  end if
  for $t = 1$ to N **do**
    if $W_j^{(t)} > 0$ **then**
      Generate $\theta^* \sim K(\theta^*|\theta_{(i)}^{(t-1)})$
      for $k = 1$ to M **do**
        Simulate $X_{(*,k)} \sim f(\cdot|\theta^*)$
      end for
      Generate $u < \mathscr{U}_{[0,1]}$
      if $u \leq 1 \wedge \dfrac{\sum_{k=1}^{M} \mathbb{1}_{A_{\epsilon_t,y}}(X_{(*,k)}) \pi(\theta^*) K_t(\theta_{(i)}^{(t-1)}|\theta^*)}{\sum_{k=1}^{M} \mathbb{1}_{A_{\epsilon_t,y}}(X_{(i,k)}^{(t-1)}) \pi(\theta_{(i)}^{(t-1)}) K_t(\theta^*|\theta_{(i)}^{(t-1)})}$ **then**
        Set $(\theta_{(i)}^{(t)}, X_{(i,1:M)}^{(t)}) = (\theta^*, X_{(*,1:M)})$
      else
        Set $(\theta_{(i)}^{(t)}, X_{(i,1:M)}^{(t)}) = (\theta_{(i)}^{(t-1)}, X_{(i,1:M)}^{(t-1)})$
      end if
    end if
  end for
end while

Algorithm 4.6 Adaptive Population Monte Carlo ABC (APMC)

Given N, $N_\alpha = \lfloor \alpha N \rfloor$ the number of particles to keep at each iteration among the N particles ($\alpha \in [0,1]$) and $p_{acc_{min}}$ the minimal acceptance rate.

for $t = 1$ **do**

  for $i = 1$ to N **do**

    Simulate $\theta_i^{(0)} \sim \pi(\theta)$ and $x \sim f(x|\theta_i^{(0)})$

    Set $\rho_i^{(0)} = \rho(S(x), S(y))$

    Set $w_i^{(0)} = 1$

  end for

  Let $\epsilon_1 = Q_{\rho^{(0)}}(\alpha)$ the first α-quantile of $\rho^{(0)}$ where $\rho^{(0)} = \left\{\rho_i^{(0)}\right\}_{1 \leq i \leq N}$

  Let $\left\{(\theta_i^{(1)}, w_i^{(1)}, \rho_i^{(1)})\right\} = \left\{(\theta_i^{(0)}, w_i^{(0)}, \rho_i^{(0)}) | \rho_i^{(0)} \leq \epsilon_1,\ 1 \leq i \leq N\right\}$

  Take σ_1^2 as twice the weighted empirical variance of $\{(\theta_i^{(1)}, w_i^{(1)})\}_{1 \leq i \leq N_\alpha}$

  Set $p_{acc} = 1$

  $t \leftarrow t + 1$

end for

while $p_{acc} > p_{acc_{min}}$ **do**

  for $i = N_\alpha + 1$ to N **do**

    Pick θ_i^* from $\theta_j^{(t-1)}$ with probability $\frac{w_j^{(t-1)}}{\sum_{k=1}^{N_\alpha} w_k^{(t-1)}}$, $1 \leq j \leq N_\alpha$

    Generate $\theta_i^{(t-1)} | \theta_i^* \sim \mathcal{N}(\theta_i^*, \sigma_{(t-1)}^2)$ and $x \sim f(x|\theta_i^{(t-1)})$

    Set $\rho_i^{(t-1)} = \rho(S(x), S(y))$

    Set $w_i^{(t-1)} = \dfrac{\pi(\theta_i^{(t-1)})}{\sum_{j=1}^{N_\alpha} (w_j^{(t-1)} / \sum_{k=1}^{N_\alpha} w_k^{(t-1)}) \sigma_{t-1}^{-1} \varphi(\sigma_{t-1}^{-1}(\theta_i^{(t-1)} - \theta_j^{(t-1)}))}$

  end for

  Set $p_{acc} = \frac{1}{N - N_\alpha} \sum_{k=N_\alpha+1}^{N} \mathbb{1}_{\rho_i^{(t-1)} < \epsilon_{t-1}}$

  Let $\epsilon_t = Q_{\rho^{(t-1)}}(\alpha)$ where $\rho^{(t-1)} = \left\{\rho_i^{(t-1)}\right\}_{1 \leq i \leq N}$

  Let $\left\{(\theta_i^{(t)}, w_i^{(t)}, \rho_i^{(t)})\right\} = \left\{(\theta_i^{(t-1)}, w_i^{(t-1)}, \rho_i^{(t-1)}) | \rho_i^{(t-1)} \leq \epsilon_t,\ 1 \leq i \leq N\right\}$

  Take σ_t^2 as twice the weighted empirical variance of $\{(\theta_i^{(t)}, w_i^{(t)})\}_{1 \leq i \leq N_\alpha}$

  $t \leftarrow t + 1$

end while

Where $\forall u \in [0,1]$ and $X = \{x_1, ..., x_n\}$, $Q_X(u) = \inf\{x \in X | F_X(x) \geq u\}$ and $F_X(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{x_k \leq x}$.

Where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Appendix 4.B: Proof that the algorithm stops

We know that there exists $\epsilon_\infty > 0$ such that $\epsilon_t \underset{t\to+\infty}{\longrightarrow} \epsilon_\infty$ because, by construction of the algorithm (ϵ_t) is a positive decreasing sequence and it is bounded by 0.
For each $\theta \in \Theta$, we consider the distance $(\rho(S(x), S(y))|\theta)$ as a random variable $\rho(\theta)$. Let $f_{\rho(\theta)}$ be the probability density function of $\rho(\theta)$.
The probability $\mathbb{P}[\rho(\theta) \geq \epsilon_t]$ that the drawn distance associated to parameter θ is higher than the current tolerance ϵ_t satisfies:

$$\begin{aligned}\mathbb{P}[\rho(\theta) \geq \epsilon_t] &= 1 - \mathbb{P}[(\rho(\theta) < \epsilon_t] \\ &= 1 - \int_{\epsilon_\infty}^{\epsilon_t} f_{\rho(\theta)}(x) d_x\end{aligned}$$

We define:

$$\mathbb{P}_{max} = \sup_{\theta\in\Theta} \left\{ \sup_{x\in\mathbb{R}^+} \left\{ f_{\rho(\theta)}(x) \right\} \right\}$$

We have:

$$\mathbb{P}[\rho(\theta) \geq \epsilon_t] \geq 1 - \mathbb{P}_{max}(\epsilon_t - \epsilon_\infty)$$

The $N - N_\alpha$ particles are independent and identically distributed from π_{t+1} the density defined by the algorithm, hence the probability $\mathbb{P}[p_{acc}(t+1) = 0]$ that no particle is accepted at step $t+1$ is such that:

$$\mathbb{P}[p_{acc}(t+1) = 0] \geq (1 - \mathbb{P}_{max}(\epsilon_t - \epsilon_\infty))^{N-N_\alpha}$$

If $\mathbb{P}_{max} < +\infty$, because $\epsilon_t - \epsilon_\infty \underset{t\to+\infty}{\longrightarrow} 0$, we have:

$$\mathbb{P}[p_{acc}(t+1) = 0] \underset{t\to+\infty}{\longrightarrow} 1$$

We can conclude that $p_{acc}(t)$ converges in probability towards 0 if $\mathbb{P}_{max} < +\infty$. This ensures that the algorithm stops, whatever the chosen value of $p_{acc_{min}}$.

Summary and Perspectives

Contents

In this thesis, we have developed statistical tools matching various needs for elaborating the microsimulation model *SimVillages*. Indeed, this model is "data driven" and it requires the use of statistics from its construction to its validation. With the increase of data sources and the storage capacities, we can expect that integrated multi-formalism data driven models will become more and more common. Therefore, the type of approach we have adopted in the thesis will correspond to an increasing need.

Our work shows that it is generally difficult to use directly existing methods and that specific needs of the model pose new research problems. For example, to generate a synthetic population or a commuting network without detailed data we needed to create new methods (see Chapter 1 and Chapter 2). To estimate the number of proximity service jobs we have adapted the Minimum Requirement method to our problem (Chapter 3) and to calibrate the model we have adapted the Population Monte Carlo ABC algorithm to speed it up (Chapter 4).

In this chapter, we propose a brief summary and we present some perspectives and open questions for each contribution of the thesis. Several avenues exist to pursue the research carried out in this thesis. Indeed, the three first contributions developed in this thesis need to be tested on more practical applications and/or new case studies. Regarding the calibration of microsimulation models and the Approximate Bayesian Computation methods, questions and problems abound.

1 Generating a synthetic population

1.1 Summary of my contribution

In the Chapter 1, we compare a sample-free method proposed by Gargiulo et al. (2010) with a sample-based method proposed by Ye et al. (2009) for generating a synthetic population, organised in households, from various statistics. We generate a reference population for a French region including 1310 municipalities and measure how both methods approximate it from a set of statistics derived from this reference population. We also perform a sensitivity analysis. The sample-free method better fits the reference distributions of both individuals and households. It is also less data demanding but it requires more pre-processing. The quality of the results for the sample-based method is highly dependent on the quality of the initial sample.

1.2 Perspectives and open questions

Sensitivity to the uncertainty on the initial population An interesting problem which has not been addressed in this thesis is about the sensitivity to the uncertainty on the initial population of a stochastic dynamic microsimulation model. Indeed, when the intial state comes from a stochastic model (as for the *SimVillages* model) we need to choose one result of synthetic population from this model. Such as, for instance, the "best" synthetic population among a set of synthetic populations according to the goodness of fit to the observed data (the 1990 census for the *SimVillages* model). Then, it is important to study the degree to which the initial state affects the calibration process and the model outputs. This requires to use a set of different synthetic populations to study the impact of the change of initial state on the estimation of unknown model parameter values and also to study the propagation of the initial state uncertainty on the model outputs.

Comparison between sample-free and sample-based methods In order to refine the comparison, it would be interesting to further compare sample-free and sample-based methods on other case studies.

2 A universal model of commuting networks

2.1 Summary of my contribution

In the Chapter 2, we show that a recently proposed model generates accurate commuting networks on 80 case studies from different regions of the world (Europe and United-States) at different scales (e.g. municipalities, counties, regions). The model takes as input the number of commuters coming in and out of each geographic unit and generates the matrix of commuting flows between the units. The single parameter of the model follows a universal law that depends only on the scale of the geographic

units. We show that our model significantly outperforms two other approaches proposing a universal commuting model (Balcan et al., 2009; Simini et al., 2012), particularly when the geographic units are small (e.g. municipalities).

2.2 Perspectives and open questions

Validation of the universal model of commuting networks To carry on the validation of the commuting network generation model it would be interesting to test the model on new case studies with different scales, different cultures and at different years. We could also validate the model using an epidemic model such as a SIR model (Susceptible Infected Recovered). Indeed, we could measure the impact of the use of the simulated commuting network instead of the observed one on the epidemic spread of infectious diseases.

3 Deriving the number of jobs in proximity services

3.1 Summary of my contribution

In the Chapter 3, we use a minimum requirement approach to derive the number of jobs in proximity services per inhabitant in French rural municipalities. We first classify the municipalities according to their time distance in minutes by car to the municipality where the inhabitants go the most frequently to get services (called MFM). For each set corresponding to a range of time distance to MFM, we perform a quantile regression estimating the minimum number of service jobs per inhabitant that we interpret as an estimation of the number of proximity jobs per inhabitant. We observe that the minimum number of service jobs per inhabitant is smaller in small municipalities. Moreover, for municipalities of similar sizes, when the distance to the MFM increases, the number of jobs of proximity services per inhabitant increases.

3.2 Perspectives and open questions

Apply the method to other case studies To highlight the differences between societies in terms of proximity services frequence in rural municipalities it would be interesting to apply the method to other countries and at different years.

4 Adaptive approximate Bayesian computation for complex models

4.1 Summary of my contribution

In the Chapter 4, we propose a new approximate Bayesian computation (ABC) algorithm that aims at minimizing the number of model runs for reaching a given quality of the posterior approximation. This algorithm automatically determines its sequence of tolerance levels and makes use of an easily interpretable stopping criterion. Moreover,

it avoids the problem of particle duplication found when using a MCMC kernel. When applied to a toy example and to a complex social model, our algorithm is 2 to 8 times faster than the three main sequential ABC algorithms currently available.

4.2 Perspectives and open questions

How to use the parameter posterior distribution estimated with ABC? We have decided to use ABC rather than a heuristic optimisation method to calibrate the *SimVillages* model because of the theory behind ABC which gives a rigorous mathematical framework of the estimated quantities. Indeed, with ABC, we estimate the conditional distribution over parameters given observed data know as posterior parameter distribution. The posterior distribution can be used in two ways.

First, it is interesting to use this distribution to perform an uncertainty analysis. The uncertainty analysis determines the level of uncertainty in the model outputs resulting from the uncertainty in the model inputs or on the parameters. It provides information on the uncertainty associated with model results. The uncertainty analysis is based on the distribution of the model results when running it using all the parameters drawn from the posterior distribution. These model results can be the same as the ones used for estimating the parameter posterior distribution and thus help evaluate the quality of this estimation. The model results can also be different from the ones used for the calibration, and in particular results at future time steps. In the latter case, one can study how the uncertainty increases with time.

Second, it is interesting to use the posterior distribution to estimate and to extract the most likely parameter values with, for example, a kernel density estimation. Then, we can use these parameter values to calibrate the model, to explore the model dynamics and then use these values to discuss the potential future trends that are given by the model.

ABC versus optimisation The goal of the optimization and ABC are different. Indeed, with optimization method like Particle Swarm Optimization, we want to find a set of parameter values for which the error (distance between the observed data and simulated data) is less than a threshold. With ABC, we can theoretically make optimization, by taking the maximum of the parameter values distribution as such an optimum. It would be interesting to compare the results obtained with both methods in terms of fit to observed data and number of simulations.

Choosing summary statistics An important issue in ABC is the choice of summary statistics. The summary statistics are a set of conditions supposed to be sufficient to summarize the data. Many good answers have been proposed to select these statistics (see for example Joyce and Marjoram (2008); Fearnhead and Prangle (2012)). However, all these methods propose to select the summary statistics before performing the ABC. However, the links summary statistics-parameters change depending on where we are in the parameter space. It would be interesting, especially in the case of sequential methods to weight the summary statistics according to the parameter sensitivity and

thus to modify the importance of summary statistics in each iteration (even sometimes remove useless statistics). For example, the weights may be determined with sensitivity indices calculated with the simulations of each iteration.

High performance computing The computational cost of running the *SimVillages* model on the Cantal departement is about one minute by simulation. Therefore, to perform an execution of the APMC algorithm with $N_\alpha = 1000$ particles, $\alpha = 0.5$ and a stopping criterion $p_{acc_{min}} = 0.01$ we need 80 iterations so 80×1000 minutes (about 56 days). To overcome this limitation, we need to parallelize simulations at each iteration. To do this, we use OpenMOLE (Open MOdeL Experiment) (Reuillon et al., 2010), a generic workflow engine for experimenting on simulation models using distributed computing. OpenMOLE allows us, among other things, to separate tasks and to parallelize some of them. We used OpenMOLE to parallelize simulations at each iteration with a cluster of 24 nodes. Now, to perform one execution of the APMC algorithm we need about 5 days. We plan to use a computational grid composed of 2000 elements to study the sensitivity of N_α to the parameter posterior distribution.

Bibliography

Alonso, W. (1964). *Location and land use: toward a general theory of land rent*. Publication of the Joint Center for Urban Studies. Harvard University Press.

Amblard, F. (2003). *Comprendre le fonctionnement de simulations sociales individus-centrées : Application à des modèles de dynamiques d'opinions*. PhD thesis, Université Blaise-Pascal.

Arentze, T., Timmermans, H., and Hofman, F. (2007). Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2014:85–91.

Aubert, F., Dissart, J. C., and Lépicier, D. (2009). Facteurs de localisation de l'emploi résidentiel en France. In *XLVIème Colloque de l'Association de Science Régionale de Langue Française (ASRDLF), 6-8 juillet, Clermont-Ferrand, France, 27*.

Balcan, D., Colizza, V., Goncalves, B., Hud, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.

Ballas, D., Clarke, G., Dorling, D., and Rossiter, D. (2007). Using Simbritain to model the geographical impact of national government policies. *Geographical Analysis*, 39(1):44–77.

Ballas, D., Clarke, G. P., and Wiemers, E. (2005). Building a dynamic spatial microsimulation model for Ireland. *Population, Space and Place*, 11(3):157–172.

Baqueiro-Espinosa, O., Unay-Gailhard, I., Raley, M., and Huet, S. (2011). Two adaptations of a Microsimulation Model to Study the Impact of Policies at the Municipality level. Technical report, PRIMA European project.

Barabási, A. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.

Barrat, A., Barthélemy, M., and Vespignani, A. (2005). The effects of spatial constraints on the evolution of weighted complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 11(5):49–68.

Barthelemy, J. and Toint, P. L. (2012). Synthetic Population Generation Without a Sample. *Transportation Science*.

Barthélemy, M. (2011). Spatial Networks. *Physics Reports*, 499:1–101.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Beaumont, M. A., Cornuet, J., Marin, J., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6 PART A):415–429.

Berger, T. and Schreinemachers, P. (2006). Creating agents and landscapes for multiagent systems from random samples. *Ecology and Society*, 11(2).

Bernstein, D. (2003). Transportation planning. In *The Civil Engineering Handbook*. Boca Raton, London, New York, Washington D.C.: CRC Press LLC.

Birkin, M. and Clarke, M. (2011). Spatial Microsimulation Models: A Review and a Glimpse into the Future. In Stillwell, J. and Clarke, M., editors, *Population Dynamics and Projection Methods*, Understanding Population Trends and Processes, chapter 9, pages 193–208. Springer.

Birkin, M. and Wu, B. (2012). A Review of Microsimulation and Hybrid Agent-Based Approaches. In Heppenstall, A. J., Crooks, A. T., See, L. M., and Batty, M., editors, *Agent-Based Models of Geographical Systems*, pages 51–68. Springer Netherlands.

Blanc, M., Ambiaud, r., and Schmitt, B. (2007). Orientation économique et croissance locale de l'emploi dans les bassins de vie des bourgs et petites villes. *Économie et Statistique*, 402(1):57–74.

Blum, M. G. B. and François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1):63–73.

Bousquet, F. and Le Page, C. (2004). Multi-agent simulations and ecosystem management: A review. *Ecological Modelling*, 176(3–4):313 – 332.

Bozon, M. and Héran, F. (1987). La découverte du conjoint. I. Evolution et morphologie des scènes de rencontre. *Population*, 42(6):943–985.

Bozon, M. and Héran, F. (1988). La découverte du conjoint. II. Les scènes de rencontre dans l'espace social. *Population*, 43(1):121–150.

Brodsky, H. and Sarfaty, D. E. (1977). Measuring the urban economic base in a developing country. *Land Economics*, 53:445–454.

Brown, D. G., Aspinall, R., and Bennett, D. A. (2006). Landscape models and explanation in landscape ecology - A space for generative landscape science? *Professional Geographer*, 58(4):369–382.

Brown, D. G. and Robinson, D. T. (2006). Effects of Heterogeneity in Residential Preferences on an Agent-Based Model of Urban Sprawl. *Ecology And Society*, 11(1):46.

Carnell, R. (2009). lhs: Latin Hypercube Samples. *R package version 0.5*.

Choukroun, J.-M. (1975). A general framework for the development of gravity-type trip distribution models. *Regional Science and Urban Economics*, 5(2):177–202.

Clark, W. A. V., Huang, Y., and Withers, S. (2003). Does commuting distance matter?: Commuting tolerance and residential change. *Regional Science and Urban Economics*, 33(2):199 – 221.

Coulombel, N. (2011). Residential choice and household behavior : State of the Art. *SustainCity Working Paper, 2.2a, ENS Cachan*.

Cörvers, F., Hensen, M., and Bongaerts, D. (2009). Delimitation and coherence of functional and administrative regions. *Regional Studies*, 43(1):19–31.

Davezies, L. (2009). L'économie locale "résidentielle". *Géographie Economie Société*, 11(1):47–53.

De Montis, A., Barthélemy, M., Chessa, A., and Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924.

De Montis, A., Chessa, A., Campagna, M., Caschili, S., and Deplano, G. (2010). Modeling commuting systems through a complex network analysis: A study of the Italian islands of Sardinia and Sicily. *The Journal of Transport and Land Use*, 2(3):39–55.

De Vries, J., Nijkamp, P., and Rietveld, P. (2009). Exponential or power distance-decay for commuting? An alternative specification. *Environment and Planning A*, 41(2):461–480.

Deffuant, G. (2001). Rapport final du projet FAIR 3 2092 IMAGES : Modélisation de la diffusion de l'adoption de mesures agri-environnementales par les agriculteurs (1997-2001). Technical report, Cemagref.

Deffuant, G., Amblard, F., Weisbuch, G., and Faure, T. (2002). How can extremism prevail? A study based on the relative agreement model. *The Journal of Artificial Societies and Social Simulation*, 5(4):27.

Deffuant, G., Huet, S., and Amblard, F. (2005). An individual-based model of innovation diffusion mixing social value and individual payoff dynamics. *American Journal of Sociology*, 110(4):41–69.

Deffuant, G., Huet, S., and Skerratt, S. (2008). *An agent based model of agri-environmental measure diffusion: What for ?* INSISOC, Valladolid, ESP.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society.Series B: Statistical Methodology*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5):1009–1020.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sample Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*, 11:427–444.

Dissart, J.-C., Aubert, F., and Truchet, S. (2009). An estimation of tourism dependence in French rural areas. In Matias A., Sarmento M., N. P., editor, *Advances in Modern Tourism Research II*, chapter 17. Springer.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation. *Biometrics*, 67(1):225–233.

Dubuc, S. (2004). Dynamisme rural : l'effet des petites villes. *L'Espace Géographique*, 1:69–85.

English, D., Marcouiller, D., and Cordell, H. (2000). Tourism dependence in rural America: Estimates and effects. *Society & Natural Resources*, 13(3):185–202.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society.Series B: Statistical Methodology*, 74(3):419–474.

Felemou, M. (2011). *Analyse de données relatives à l'évolution des communes d'Auvergne pour la sélection de communes prototypiques.* PhD thesis, Université Blaise-Pascal.

Fernandez, L. E., Brown, D. G., Marans, R. W., and Nassauer, J. I. (2005). Characterizing location preferences in an exurban population: Implications for agent-based modeling. *Environment and Planning B: Planning and Design*, 32(6):799–820.

Fik, T. J. and Mulligan, G. F. (1990). Spatial flows and competing central places: Towards a general theory of hierarchical interaction. *Environment & Planning A*, 22(4):527–549.

Filippi, S., Barnes, C., and Stumpf, M. P. H. (2011). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *arXiv:1106.6280v3*.

Fontaine, C. M. and Rounsevell, M. D. A. (2009). An agent-based approach to model future residential pressure on a regional landscape. *Landscape Ecology*, 24(9):1237–1254.

Fotheringham, A. (1981). Spatial structure and distance-decay parameters. *Annals, Association of American Geographers*, 71(3):425–436.

Gargiulo, F., Lenormand, M., Huet, S., and Baqueiro Espinosa, O. (2012). Commuting Network Models: Getting the Essentials. *Journal of Artificial Societies and Social Simulation*, 15(2):6.

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, 5(1).

Gitlesen, J. P., Kleppe, G., Thorsen, I., and Ubøe, J. (2010). An empirically based implementation and evaluation of a hierarchical model for commuting flows. *Geographical Analysis*, 42(3).

Glynn, P. and Whitt, W. (1992). The Asymptotic Efficiency of Simulation Estimators. *Oper. Res.*, 40(3):505–520.

Goux, D. (2003). Une histoire de l'enquête emploi. *Economie et statistique*, 362(1):41–57.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., and Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, 221(23):2760–2768.

Guo, J. Y. and Bhat, C. R. (2007). *Population Synthesis for Microsimulating Travel Behavior*. Number 2014 in Transportation Research Record. Transportation Research Board of the National Academies.

Harland, K., Heppenstall, A., Smith, D., and Birkin, M. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of Artificial Societies and Social Simulation*, 15(1):1.

Haynes, K. E. and Fotheringham, A. S. (1984). *Gravity and Spatial Interaction Models*. Sage Publications, Beverly Hills.

Holmes, E., Holme, K., Mäkilä, K., Kauppi, M. M., and Mörtvik, G. (2002). *The sverige spatial microsimulation model, content, validation, and example applications*. gerum kulturgeografi, Umea university.

Huang, Z. and Williamson, P. (2002). A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata. Working paper, Departement of Geography, University of Liverpool.

Hubert, J. P. (2009). Dans les grandes agglomérations, la mobilité quotidienne des habitants diminue, et elle augmente ailleurs. *Insee première*, (1252).

Huet, S. and Deffuant, G. (2011a). An Abstract Modelling Framework implemented through a Data-Driven approach to study the Impact of Policies at the Municipality level. *ESSA 2011 Conference, september 2011*, page 22.

Huet, S. and Deffuant, G. (2011b). Common Framework for the Microsimulation Model in PRIMA project. Technical report, Cemagref LISC.

Huet, S., Dumoulin, N., Deffuant, G., Gargiulo, F., Lenormand, M., Baqueiro Espinosa, O., and Ternès, S. (2012a). Micro-simulation model of municipality network in the Auvergne case study. Technical report, PRIMA Project, IRSTEA(Cemagref) LISC.

Huet, S., Lenormand, M., Deffuant, G., and Gargiulo, F. (2012b). Parameterisation of individual working dynamics. In Smajgl, A. and Barreteau, O., editors, *Empirical Agent-Based Modeling: Parameterization Techniques in Social Simulations*, chapter ??, page 22. Springer.

INSEE (1999). Le Modèle de Microsimulation Dynamique, DESTINIE. Document de travail, G9913, INSEE.

Joyce, P. and Marjoram, P. (2008). Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1).

Klosterman, R. (1990). *Community analysis and planning techniques*. Rowman & Littlefield.

Koenker, R. and Bassett, G. J. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.

Konjar, M., Lisec, A., and Drobne, S. (2010). Method for delineation of functional regions using data on commuters. In *13th AGILE International Conference on Geographic Information Science (Guimarães, Portugal)*, Guimarães, Portugal.

Leatherman, J. C. and Marcouiller, D. W. (1996). Estimating tourism's share of local income from secondary data sources. *Review of Regional Studies*, 26(3):x5–339.

Lemercier, C. and Rosental, P.-A. (2008). Les migrations dans le Nord de la France au XIXe siècle. In *Nouvelles approches, nouvelles techniques en analyse des réseaux sociaux*, Lille France.

Lenormand, M. and Deffuant, G. (2012). Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods. *arXiv:1208.6403v1*.

Lenormand, M., Huet, S., and Deffuant, G. (2012a). Deriving the Number of Jobs in Proximity Services from the Number of Inhabitants in French Rural Municipalities. *PLoS ONE*, 7(7):e40001.

Lenormand, M., Huet, S., and Gargiulo, F. (2012b). Generating French Virtual Commuting Network at Municipality Level. *arXiv:1109.6759v2*.

Lenormand, M., Huet, S., Gargiulo, F., and Deffuant, G. (2012c). Universal Commuting Network Model. *PLoS ONE*, 7(10):e45985.

Lenormand, M., Jabot, F., and Deffuant, G. (2012d). Adaptive approximate Bayesian computation for complex models. *arXiv:1111.1308v2*.

Marin, J.-M., Pudlo, P., Robert, C., and Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing*.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–15328.

Moeckel, R., Spiekermann, K., Schürmann, C., and Wegener, M. (2003). Microsimulation of land use. *International Journal of Urban Sciences*, 71(1):14–31.

Moore, C. L. (1975). A New Look at the Minimum Requirements Approach to Regional Economic Analysis. *Economic Geography*, 51(4):350–356.

Morand, E., Toulemon, L., Pennec, S., Baggio, R., and Billari, F. (2010). Demographic modelling: The state of the art. *SustainCity Working Paper, 2.1a, Ined, Paris*.

Mordier, B. (2010). Les services marchands aux particuliers s'implantent dans l'espace rural. *Insee première*, (1307).

Müller, K. and Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen und Eisenbahnbau (IVT).

Orcutt, G., Caldwell, S., and Wertheimer, R. (1976). *Policy exploration through microanalytic simulation*. Governance in Europe Series. Urban Institute.

Orcutt, G. H. (1957). A New Type of Socio-Economic System. *The Review of Economics and Statistics*, 39(2):pp. 116–123.

Ortúzar, J. and Willumsen, L. (2011). *Modeling Transport*. John Wiley and Sons Ltd, New York.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., and Deadman, P. (2003). Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review. *Annals of the Association of American Geographers*, 93(2):314–337.

Pastor-Satorras, R. and Vespignani, A. (2004). *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, New York, NY, USA.

Patuelli, R., Reggiani, A., Gorman, S. P., Nijkamp, P., and Bade, F. (2007). Network analysis of commuting flows: A comparative static approach to German data. *Networks and Spatial Economics*, 7(4):315–331.

Perrier-Cornet, P. (2001). La dynamique des espaces ruraux dans la société française : un cadre d'analyse. *Territoires 2020*, 3:61–74.

Persky, J. and Wiewel, W. (1994). The growing localness of the global city. *Economic Geography*, 70(2):129–143.

Polhill, J., Parker, D., Brown, D., and Grimm, V. (2008). Using the ODD Protocol for Describing Three Agent-Based Social Simulation Models of Land-Use Change. *Journal of Artificial Societies and Social Simulation*, 11(2):3.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reggiani, A. and Rietveld, P. (2010). Networks, commuting and spatial structures: An introduction. *The Journal of Transport and Land Use*, 2(3):1–4.

Reggiani, A. and Vinciguerra, S. (2007). Network connectivity models: An overview and empirical applications. In Friesz, T. L., editor, *Network Science, Nonlinear Science and Infrastructure Systems*, volume 102 of *International Series in Operations Research & Management Science*, pages 147–165. Springer US.

Reuillon, R., Chuffart, F., Leclaire, M., Faure, T., Dumoulin, N., and Hill, D. (2010). Declarative task delegation in OpenMOLE. In *High Performance Computing and Simulation (HPCS), 28/06/2010-02/07/2010, Caen, France*, pages 55–62.

Rindfuss, R. R., Walsh, S. J., Turner, B. L., Fox, J., and Mishra, V. (2004). Developing a science of land change: Challenges and methodological issues. *Proceedings of the National Academy of Sciences*, 101(39):13976–13981.

Roth, C., Kang, S. M., Batty, M., and Barthélémy, M. (2011). Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE*, 6(1).

Rouwendal, J. and Nijkamp, P. (2004). Living in two worlds: A review of home-to-work decisions. *Growth and Change*, 35(3):287–303.

Rutland, T. and O'Hagan, S. (2007). The growing localness of the Canadian City, or, on the continued (ir)relevance of economic base theory. *Local Economy*, 22(2):163–185.

Simini, F., Gonzalez, M. C., Maritan, A., and Barabasi, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1760–1765.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34.

Soumagne, J. (2003). Les services en milieu rural, enjeu d'aménagement territorial. *Revista da Faculdade de Letras - Geografia I série, XIX*.

Stillwell, J. and Duke-Williams, O. (2007). Understanding the 2001 UK census migration and commuting data: The effect of small cell adjustment and problems of comparison with 1991. *Journal of the Royal Statistical Society.Series A: Statistics in Society*, 170(2):425–445.

Thorsen, I. and Gitlesen, J. P. (1998). Empirical evaluation of alternative model specifications to predict commuting flows. *Journal of Regional Science*, 38(2):273–292.

Thorsen, I., Ubøe, J., and Nævdal, G. (1999). A network approach to commuting. *Journal of Regional Science*, 39(1):73–101.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187.

Ullman, E. and Dacey, M. (1960). The Minimum Requirement Approach to the Urban Economic Base. *Papers and Proceedings of the Regional Science Assn.*, 6:192.

Van Den Berg, G. J. and Gorter, C. (1997). Job search and commuting time. *Journal of Business and Economic Statistics*, 15(2):269–281.

Verburg, P., Schulp, C., Witte, N., and Veldkamp, A. (2006). Downscaling of land use change scenarios to assess the dynamics of european landscapes. *Agriculture, Ecosystems and Environment*, 114(1):39 – 56.

Verburg, P. H., Schot, P. P., Dijst, M. J., and Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *GeoJournal*, 61(4):309–324.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451.

Voas, D. and Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5):349–366.

Voas, D. and Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2):177–200.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M., and Ulfarsson, G. (2003). Microsimulation of urban development and location choices: Design and implementation of Urbansim. *Networks and Spatial Economics*, page 2003.

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009). Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood. *Genetics*, 182(4):1207–1218.

Williams, I. (1976). A comparison of some calibration techniques for doubly constrained models with an exponential cost function. *Transportation Research*, 10(2):91–104.

Wilson, A. G. (1998). Land-Use/Transport Interaction Models: Past and Future. *Journal of Transport Economics and Policy*, 32(1):3–26.

Wilson, A. G. and Pownall, C. E. (1976). A new representation of the urban system for modelling and for the study of micro-level interdependence. *Area*, 8(4):246–254.

Woller, G. and Parsons, R. (2002). Assessing the community economic impact of non-governmental development organizations. *Nonprofit and Voluntary Sector Quarterly*, 31(3):419–428.

Ye, X., Konduri, K., Pendyala, R., Sana, B., and Waddell, P. (2009). Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. In *88th Annual Meeting of the Transportation Research Board*.

APPENDIX A

Parameterisation of Individual Working Dynamics

Manuscript:
Huet, S., Lenormand, M., Deffuant, G. and Gargiulo, F. Parameterisation of individual working dynamics. *Accepted in Empirical Agent-Based Modeling: Parametrization Techniques in Social Simulations*, 2012, Chapter ??, 22 pages, A. Smagl and O. Barreteau eds, Springer.

Parameterisation of Individual Working Dynamics

S. Huet, M. Lenormand, G. Deffuant, F. Gargiulo

***Laboratoire d'Ingénierie pour les Systèmes Complexes,**

Irstea

How do European rural areas evolve? While for decades the countryside in many regions of Europe was synonymous with inevitable decline, nowadays, some areas experience a “rebirth, even in areas where until recently development was not considered possible" (Y. Champetier, 2000). A recent EPSON (European Observation Network for Territorial Development and Cohesion) project report (Johansson and Rauhut, 2007), concludes that "since the 1970s a global process of counter-urbanization has become increasingly manifest". However, this general rebirth of the countryside hides deep heterogeneities. That can be observed in the Cantal "département" in France where the population remains stable after having been depopulated with some subgroups of its municipalities have an increasing population while others have a decreasing one. Our modelling effort aims at better understanding these heterogeneities.

Micro modelling (Gilbert and Troitzch 2005) is a very relevant paradigm to study the evolution of areas composed from various objects appearing as very heterogeneous. It includes three different approaches: cellular automata change (Ballas et al. 2005, 2006; Ballas et al. 2007; Brown et al. 2006; Coulombel 2010; Moeckel et al. 2003; Rindfuss et al. 2004; P.H. Verburg et al. 2002; P.H. Verburg et al. 2004; P. H. Verburg et al. 2006), microsimulation (Orcutt 1957) (INSEE 1999), (Holme et al. 2004), (Turci et al. 2010) (Morand et al. 2010) and agent-based models (Bousquet and Le Page 2004; Brown and Robinson 2006; Deffuant and al. 2001; Deffuant et al. 2002; Deffuant et al. 2005; Deffuant et al. 2008; Fontaine and Rounsevell 2009; Parker et al. 2003) (Fontaine and Rounsevell 2009) which have been already used to study problem close to ours. However, recent reviews recommend a hybrid approach (Birkin and Clarke 2011; Birkin and Wu 2012), particularly coupling microsimulation and agent-based modelling. Thus, trying to develop an approach which is as close to the data as we can, we decide to use microsimulation and agent approaches allowing us to address some complex individual dynamics, largely unknown and for which no data are available, such as the residential location decision (Coulombel 2010).

The problem of such modelling approach is the link to data. If it is obvious in the basic microsimulation, that is not so easily manageable in dynamic microsimulation with a “real” evolution time after time of the individual. Indeed the dynamic microsimulation remains rare (Birkin and Wu 2012): the most common way to introduce change of the demographic structure is to apply static ageing techniques consisting in reweighting the age class according to external information. That is to avoid considering functions of evolution of the behaviour of the individual and their parameterisation. Regarding the multiagent modelling, (Berger and Schreinemachers 2006) argue it “holds the promise of providing an enhanced collaborative framework in which planners, modellers, and stakeholders may learn and interact. The fulfilment of this promise, however, depends on the empirical parameterization of multiagent models. Although multiagent models have been widely applied in experimental and hypothetical settings, only few studies have strong linkages to empirical data and the literature on methods of empirical parameterization is still limited.” An example can be read in (Fernandez et al. 2005) which initialise individual preference from analyses of the data coming from an ad hoc survey but don’t consider a possible change in the preference of an individual.

In our model[1], we tried to have a strong linkage to data both in the definition of the initial population and the one of the individual behaviour. This model implements virtual individuals, members of households located in municipalities and their state transitions corresponding to demographic and changing activity events: birth, finding a partner, moving, changing job, quitting their partner, retiring, dying … The virtual municipalities offer jobs and dwellings which constrain the possible state transitions. Because we are interested in understanding better the dynamics leading to the development or, on the contrary, to the decline and possible disappearance of municipalities and settlements, two sets of cruxes can be identified in the model: The individual dynamics

[1] This work has been funded under the PRIMA (Prototypical policy impacts on multifunctional activities in rural municipalities) collaborative project, EU 7th Framework Programme (ENV 2007-1), contract no. 212345

which determine the needs for residence and jobs; the dwelling and the job offers exogenous and endogenous dynamics at the local (i.e. municipality) level.

The present paper focuses on how to make such a model close enough to the data to guarantee a good understanding of the dynamics of population/depopulation based on "real" situations, and a real utility for policy makers. As the developed model is very large, taking into account many dynamics, we are going to focus on the design and the parameterisation of the individual dynamics regarding the labour market.

After a summary of the whole model, presented in details in (Huet et al. 2011), we present how we have conceived and parameterised the submodel of the individual activity dynamics. The final section tries to explain what we have learnt from such an exercise. In particular, we want to stress out the necessity not to only consider the objectives of the model during the design phases, but also since the very beginning censing the existing data sources and studying the implicit model beside the databases.

1 MODEL DESCRIPTION

We have adopted a micro-modelling approach. The presentation of the model globally follows the requirements of the ODD (Overview, Design concepts, and Details) framework (Grimm et al. 2006). Indeed, this recently updated protocol (Grimm et al. 2010) has proved its utility to describe properly complex individual-based models, for example in (Polhill et al. 2008).

The purpose of the model is to study how the population of rural municipalities evolves. We assume that this evolution depends, on the one hand, on the spatial interactions between municipalities through commuting flows and service, and on the other hand, on the number of jobs in various activity sectors (supposed exogenously defined by scenarios) and on the jobs in proximity services (supposed dependent on the size of the local population). Indeed, in the literature, the most cited explanation for the evolution of the rural municipalities is what is called the residential economy (Blanc and Schmitt 2007; Davezies 2009). It argues that rural areas dynamics is linked to the money transfers between production areas and residence locations. These money transfers are for instance performed by commuters, or by retirees who move from the urban to the rural areas. Indeed migrations from urban to rural areas are also considered as a very important strand for rural areas evolution (Perrier-Cornet 2001). The residential economics studies particularly how an increasing local population (and money transfers) increases the employment in local services. The geographic situation plays also a role in the municipality evolution (Dubuc 2004). To summarise, existing literature stresses the importance of the different types of mobility between municipalities, commuting, residential mobility (short range distance), migration (long range distance) (Coulombel 2010) and the local employment offer generated by the presence of the local population.

These two aspects have to be properly taken into account in our model, since our objective is to study through simulations the dynamics of rural areas. Obviously, it appears also essential to model the demographic evolution of the municipality considering the strands explaining the local natural balance.

1.1 MAIN ENTITIES, STATE VARIABLES AND SCALES

The model represents a network of municipalities and their population. The distances between municipalities are used to determine the flows of commuting individuals (for job or services). Each municipality comprises a list of households, each one defined as a list of individuals. The municipalities also include the offers of jobs, of residences and their spatial coordinates. Here is the exhaustive list of the main model entities with their main attributes and dynamics.

1.1.1 MUNICIPALITYSET

The set of municipalities can be of various sizes. It can represent a region of type NUTS 2 or NUTS 3[2], or more LAU or intermediate sets of municipalities such as "communauté de communes" in France. In the present paper, the set corresponds to the Cantal “département” in France composed of 260 municipalities.

Parameter: a threshold distance called "proximity" between two municipalities; beyond this distance the municipalities are considered too far from each other, to allow commuting between them without considering to move for instance (parameterised at 25 km).

1.1.2 MUNICIPALITY

It corresponds to LAU2[3]. The municipality is the main focus of the model. It includes:

- A set of households living in the municipality. The household corresponds to the nuclear family[4]. It includes a list of individuals who have an occupation located inside or outside the municipality).
- The set of jobs existing on the municipality and available for the population of the model (i.e. subtracting the jobs occupied by people living outside the modelling municipality set).
- The distribution of residences, or lodgings, on the municipality.

There is a particular municipality, called "Outside": it represents available jobs accessible from municipalities of the considered set, but which are not in the considered set. The job offer of Outside is infinite and the occupation is defined by a probability of individuals to commute outside the set (see 2.3.3 for details).

Parameters:

- An initial population of households composed of individuals with their attribute value and their situation on the labour market
- A residence offer: available number of residences for each type. A type corresponds to the number of rooms
- A job offer: number of jobs offered by the municipality for each type of job; the exogenously defined part of job offers is distinguished from the endogenously defined part in order to update this last part easily
- The laws ruling the proximity of municipalities: each municipality has rings of ‘nearby’ municipalities (practically every 3 Euclidian kilometres) with a maximum distance of 51 Euclidian km. The accessibility of each ring varies depending on the process (commuting, looking for a residence, looking for a partner) following appropriate probability distribution laws.
- Spatial coordinates

As said earlier, in the case of special municipality called "Outside", all variables, except job offer and job occupation, are empty.

1.1.3 THE JOB AND THE RESIDENCE

A job has two attributes, a profession and an activity sector in which this profession can be practiced. It is available in a municipality and can be occupied by an individual. The profession is an attribute of the individual and can take six various values (see 1.1.5 for details) at the same time it defines a job. There are four activity sectors: Agriculture, Forestry and Fishing; Industry; Building; Services and Commerce. Overall, considering the six professions for four activity sectors, we obtain 24 jobs to describe the whole diversity of jobs in the region we study (i.e. the Cantal "département", called only Cantal later in this chapter).

The residence has a type which is classically its size expressed in number of rooms. A residence is available in a municipality and can be occupied by 0, one or more households. Indeed several households can live in one residence for instance when a couple splits up and one of the partner remains in the common residence for a while. It is also the case in some European countries where it is customary for several generations to live under the same roof.

---

[2] Eurostat defines the NUTS (Nomenclature of Territorial Units for Statistics) classification as a hierarchical system for dividing up the EU territory: NUTS 1 for the major socio-economic regions; NUTS 2 for the basic regions for the application of regional policies; NUTS 3 as small regions for specific diagnoses; LAU (Local Administrative Units 1 and 2) has been added more recently to allow local level statistics

[3] consists of municipalities or equivalent units

[4] A nuclear family corresponds to the parents and the children; that is a reductive definition of the family corresponding on the most common way to define the family in Europe nowadays.

1.1.4 HOUSEHOLD

Table 1. Attributes defining the household state

Name	Type	Values
Members	List of Individuals	
Couple	Boolean	True, false
Leader	Individual	
Residence	Residence	
Residence need	Boolean	True, false
Municipality of residence	Municipality	

For the initialisation, residences are associated randomly with households. Then, new households are created when new couples are formed or when people from outside the set of municipalities migrate into the municipality. Households are eliminated when their members die, or when the couple splits up, or when they simply migrate outside the municipality set. When a behavior of an individual has an impact on the household, a leader is assigned randomly, or designed depending on the process. This leader will be the one deciding for the household. That is for example the case when an individual finds a job very far: she becomes the leader to make the household moving and finding a residence close to her new job.

1.1.5 INDIVIDUAL

The individual is instantiated via one of the adults of a household having the "couple" status in the birth method, or directly from the initialisation of the population, or by immigration.
The age to die, the age the person will enter the labour market, and the age of retirement are attributed to the individual when it is created. These ages are assigned by a probability method. The activity status defines the situation of the individual regarding employment, especially whether or not she is looking for a job. The individual can quit a job, search for and change jobs …
The profession is an attribute of the individual indicating at the same time her skills, level of education and the occupation she can aspire to. Professions take the value of the French socio-professional categories categorised in six modalities that define at the same time a kind of occupation, an average level of education and an approximate salary.

Table 2. Attributes defining the state of an individual

	Type	Values
Activity status	Enum	student, inactive, retired, employed, unemployed (only the two last can search a job)
Profession	Enum	farmers; craftsmen, storekeepers, business owners; top executive managers, upper intellectual profession (senior executives); intermediary professions; employees; workers.
Job	Couple of values	24 couples (profession, activity sector) (see 1.1.3 for details)
Place of work	Municipality	Nil or a Municipality
Household status	Enum	Adult, Child
Age to die	Integer	Drawn from a distribution
Age in labour market	Integer	Drawn from a distribution
Age of retirement	Integer	Drawn from a distribution

1.2 PROCESS OVERVIEW AND SCHEDULING

1.2.1 THE MAIN LOOP

The main loop calls processes ruling demographic evolution, the migrations, the job changes, and their impact on some endogenously created services and/or jobs. First, the scenarios are applied to the municipalities. Then, endogenously available jobs and services are updated in municipalities. Finally, demographic changes are applied to the list of households. The following pseudo code sums-up the global dynamics:

```
At each time step:
   For each municipality
```

```
   municipality.update external forcings: offer of jobs, residence
   municipality.update endogenous job offer for services to residents
   municipality.compute in-migration
For each household:
   household.members.job searching decision (this process can make free some
      jobs from people becoming retired or inactive)
For each household:
   household.members.searching for a job
   household.members events (coupling, divorce, birth, death)
   household.residential migration
   household.members.individual ages
```

Time is discrete with time steps corresponding to years. The households are updated in a random order during a time step. We shall calibrate the model on the first 16 years and study its evolution on the next 24 years.

1.2.2 DYNAMICS OF OFFER FOR JOBS, SERVICES AND LODGING

In the municipality objects, jobs, services and dwelling offers are ruled. Changes in dwelling offers are specified in scenarios. Various sizes are considered in order to match the needs of households.

The job offer process is twofold: one part defined through scenarios which specify the increase or decrease of jobs in different sectors, and a second part concerning the proximity of service jobs, which are derived by a specific statistical model.

Indeed, numerous are the researches pointing out the importance of services for the rural areas dynamism (Aubert et al. 2009; Dubuc 2004; Fernandez et al. 2005; Soumagne 2003). Also the residential economics shows the importance of the presence of the population in rural municipalities (Davezies 2009). Practically, we distinguish the proximity services which rely directly on the presence of population from the services which are decided according to other factors (assets of the location, political will at different levels, etc.). We integrated the dynamics of creation and destruction of proximity services jobs in the micro-simulation model, using a statistical model derived from the data of the region. Starting from the classical minimum requirement approach proposed by (Ullman and Dacey 1960), (Lenormand et al. 2011c) we propose a model which takes into account the distance between a municipality and its closest centre of services (i.e. most frequented municipality, called MFM). This new model has been grounded on detailed data related to jobs and poles of services (Lenormand et al. 2011a). Therefore, we use the extracted statistical relation to adjust the number of jobs in proximity services in the municipalities of the model.

It is $E = \beta_0 + \beta_1 \ln P + \varepsilon$ with E = minimum employment offer in the municipality to satisfy the need for services of one resident; P = the population of the municipality; β_0 and β_1 = parameters

For each municipality, this function is computed every year in order to update the service sector job offer depending on the distance of the municipality to the closest pole of service (called MFM). The form of the function for different municipality sizes with various distances to the MFM indicates that:

- in any case, the job offer is higher in the pole of services and decreases in the surrounding;
- however further from the pole of services, the number of jobs increases again until reaching a plateau at a distance higher than 10 minutes;
- the larger is the municipality, the higher is the number of jobs in proximity services.

The other creations and destructions of jobs are ruled by scenarios.

Parameters: distances to the Most Frequented Municipality of every municipality of the Cantal (given by the French Municipal Inventory of 1999); class of distance to the most frequented municipality (MFM) for every municipality and regression coefficients β_0 and β_1 extracted of the analysis of the French Census of 1990, 1999 and 2006 (see (Lenormand et al. 2011a) for more explanations).

Classes of distance in minutes to the Most Frequented Municipality	β_0	β_1
0	-0.170901146	0.033121263
]0,5]	-0.130158882	0.025111874
]5,10],	-0.141049558	0.026983278
>10	-0.162030187	0.031165605

Table 1. Regression coefficient for the four classes of municipalities of the Cantal

The proportion of proximity service jobs offer over professions is assumed to be the same than the one for the whole service sector job offers (which is probably a strong approximation). This allows us to distribute the proximity service jobs in the different jobs in the service sector.

1.2.3 DYNAMICS OF LABOUR STATUS AND JOB CHANGES

A new individual can be generated in a household having the "couple" status with the birth method, or directly from the initialisation of the population, or from the immigration method. A newly born individual is initialised with a student status that she keeps until she enters the labour market with a first profession. Then, she becomes unemployed or employed with the possibility to look for a job. She may also become inactive for a while. When she gets older, she becomes a retiree. We here describe rapidly these dynamics to situate them in the global picture of them model. We describe them in more details, especially the choice of parameters and link to data, in section 3.

Entering on the labour market

The individual stops being a student at the age to enter on the labour market and becomes unemployed. She searches immediately for a job and can get one during the same year. A first profession she looks for has to be defined at the same time the first age of research is determined.

Parameters: probabilistic laws to decide the age a student enters on the labor market and the first profession she is going to look for.

Job searching decision

The decision for searching a job is a two-step process. First, an individual has an activity status indicating if she is susceptible to search for a job or not. She can change her status and then her probability to seek a job. When she decides searching, she has also to decide what type of job to search for. Five different activity statuses define the individual situation regarding the labour market in the model:

1. The **student:** an individual is a student in the first part of its life, until the age she enters on the labour market. We consider the probability of a student to look for a job is 0 since we are only interested in rural municipalities. Students in age working mainly look for a job in the large cities where they study.
2. The **unemployed**: an individual is unemployed when she is considered active (on the labour market) and has no job. For sake of simplicity, we assume an unemployed has a probability 1 to look for a job.
3. The **employed**: she is an individual who has a job. She can decide searching for another job, in the same profession or not. Her probability willing to change job classically depends at least on her age.
4. The **inactive**: she can be inactive for a long time or just stopping to work for one year, having a baby for example. During this period, her probability to search for a job is 0.
5. The **retired**: at the age of retirement, an individual retires. Her probability to look for a job is then assumed to be 0.

We have seen the probability to search for a job (or the law ruling this probability) depends on the activity status. Figure 1 describes the way an individual changes activity status and thereby the probability to search.

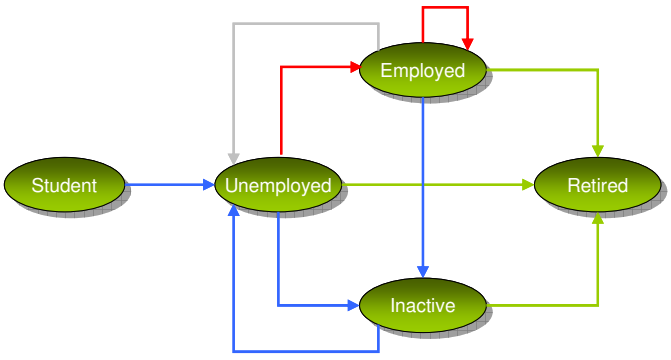


Figure 1 - Transitions of status and their link to the data. Red arrows: change by finding a job; grey arrows: when she is fired; green arrows: at the age of retirement (picked out from a law extracted from data); yellow arrows: due to a probabilistic decision of becoming inactive extracted from the Labor Force Survey data; purple arrows: due to probabilistic decisions extracted from the Labor Force Survey data.

Entering the labour market, the student becomes unemployed and searches for a job with a probability 1. An unemployed, as an employed, can find a job through processes presented in the following sections and become employed. If an unemployed always searches for a job by assumption that is not the case for an already employed individual (her probability to search has to be extracted from data). Employed and unemployed individuals can also become inactive. Then we assume that they stop searching for a job the time they remain

inactive. Every activity states, except student, can be followed by the retirement state in which we assume the individual stops searching for a job. An inactive, if she doesn't retire, either can come back on the labour market adopting an unemployed status to search for a job or can remain inactive.
Most of the laws ruling the activity status changes have to be parameterised. The grey-arrows transitions are much more endogenously defined. That is the employed to unemployed transition which is due to the decreasing availability of job offer implying a sacking. It can also be, for instance a resignation of an individual leaving her municipality to follow her partner to another place of residence.
Knowing an individual searches for a job, we have to compute which profession she looks for. One can notice that an individual only looks for a profession; we neglected to take into account the activity sector in her choice. The activity sector will be defined by the found job among the set of possible job offers for the individual. We expect the job offer to be a sufficient constraint on the activity sector to allow the model exhibiting a statistically correct distribution of occupied jobs by activity sector.
Parameters ruling the job research decision: probability becoming inactive; probability to stop being inactive; probability laws defining what profession to search for; parameters for entering the labour market and to retire

Searching for a job

The question for the individual is now to decide where to search for a job. The challenge consists in preserving the properties of the commuting distance distribution that we assume constant. Both the choice of the place of work and the choice of the place of residence impact on this distance. Thus, these processes have to be designed under this constraint. However, the place of work is not only defined by the strategy of search but also constrained by the job offer, which has to be properly defined.
If the leader of the household has already found a job far (further than the proximity attribute) from the place of residence and the household is trying to move close the leader's place of work, then the other household members, waiting for a change of residence, do not try to change job since they do not know where they will be living. Until the household finds out a new residence place, nobody is going to change jobs.
In the other cases, if the individual is searching for a job, we consider she begins by choosing where she wants to work. Practically, she picks out a distance in the probability law of the "accepted distance to work place".
Then, if the distance is higher than 0, she has to decide whether to work outside the set of municipalities. The decision to work outside is described in detail in 2.3.3. If the individual goes to work outside, she automatically has a job. She is counted as an outside commuter. The job occupation of the outside and its spatial distribution can be used to calibrate the model.
If she doesn't work outside, she goes to see the labour office. The labour office collects every job offer corresponding to the profession she is looking for at the chosen distance. Then the individual chooses one at random. This procedure allows reproducing the effect of the quantity of local offers. It gives to the municipality with a larger job offer a greater probability to be chosen.
If she chooses a job at a distance higher than the proximity distance, she becomes the leader of her household. If the distance is less than the proximity, the next household member, if she exists, will be able to search for a job. The search procedure is repeated x times if the individual has not found a job. The number of times this procedure is repeated is specified in a parameter.
Parameters: probability distribution of accepted distances to cross over to work place; probability to commute outside for an inhabitant of every municipality

Become a retiree

At a given age, the individual becomes a retiree. We assume, for sake of simplicity, that a retiree does not search for a job.
Parameter: probability to decide the individual's retirement age.

1.2.4 DEMOGRAPHIC DYNAMICS

A new household can be created when an individual becomes an adult or when a new household comes to live in the set of municipality (i.e. in-migration). The main reasons for household elimination are out-migration and death. Three main dynamics change the household type (single, couple, with or without children and complex[5]): makeCouple; splitCouple and givingBirth. These processes are now described with more details in the same order they have been presented in this introduction.

[5] A complex household is a household which is not a single, a couple with or without children.

BecomingAnAdult

Becoming an adult means an individual creates her own household. This can lead her to move from parental residence because of a low dwelling satisfaction level, but it's not always the case. An individual loses her child status and becomes an adult when: she finds her first job; or she is chosen by a single adult as a partner; or she remains the only children in a household after her parents leave or die while her age is higher than parameter firstAgeToBeAnAdult.

Parameter: first age to become an adult – 15 is the age considered by the French or other European National Statistical Offices

Household migration and mobility

In changing residence process, we include both residential migration and mobility without making a difference, between short and long distance move, as it is often the case (Coulombel 2010) in the literature. The submodel we propose directly manages both types of moving. However, it turned out easier for us to distinguish two categories of migration: the migration of people coming from outside to live inside the set; the migration of people who already live inside the set.

The immigration into the set is an external forcing. Each year, a number of potential immigrants from outside the set are added to the municipalities of the set. These potential immigrants can really become inhabitants of the set if they find a residence by themselves or by being chosen as a partner by someone already living in the set in case they are single (with or without children). Thus, looking for a place of residence is the only action they execute until they become an inhabitant of the set. Until the potential immigrant becomes a real inhabitant, she cannot search for a job. Indeed, the job occupied by people living outside the municipality set are already taken into account through the scenario and allowing potential immigrants to find a job directly would be redundant. The definition of who are potential immigrants, how numerous they are, and when they are introduced is specified exogenously. Since they are created, the potential immigrants are temporarily places into a municipality from which they can find a residence or being chosen as a partner. They are placed in a municipality following a probability to be chosen, which is computed for each municipality depending on the population size of the municipality and its distance to the frontier of the set. A particular attraction of young people for larger municipalities is also taken into account.

The mobility of people already living inside the set of municipalities is mainly endogenous. Such a mobility can lead the household simply to change residence, municipality or to quit the set of studied municipalities. Overall, a household decides to look for a new residence when:

- a new couple is formed: the couple chooses to live initially in the largest residence among the ones of the partners;
- a couple splits: one of the partners, randomly chosen, has to find out another residence even if she remains for a while in the same residence (creating her own household);
- an adult of the household finds a job away from the current place of residence (beyond the proximity parameter of the MunicipalitySet);
- a student or a retiree decides to move;
- the residence is too small or too large. This can be due to a birth, a new couple or to someone who left the residence for example. The too small or too large characteristic is assessed through a satisfaction function depending on the difference of size between the occupied size and an ideal size for this household, and the average age of the household members. In principle, people tend to move easily when they are younger and/or when the difference of size is high.

The choice of a new household is twofold: first, the household chooses a distance to move; secondly she chooses at random a new residence proposition to examine. The proposition is accepted depending on the level of satisfaction it can give. This satisfaction depends on the difference between the proposed and the ideal size, and the average age of the household members. In principle, with increasing age we assume a decrease in flexibility to accept residences different from their ideal.

A move of a household can result in increased commuting distances for some of its working members, even exceeding the proximity threshold. Such a commuter continues until she becomes the household leader through the job search mechanism and triggers the household to look for a residence closer to her job.

Parameters for immigration: yearly migration rate; number of out of the set migrants in year $t^0 - 1$; probabilities for characteristics of the immigrants (size of the households, age of individuals...); distance to the frontier of the region of each municipality.

Parameters within the set of municipalities and out-migration:

- The level of satisfaction of the size of the current dwelling or the one of a proposed dwelling is a function of the size of the household and of the its age composition; this function requires one parameter called β which has to be calibrated
- distribution of probabilities for an individual to accept moving over a certain distance to get a residence starting for her place of work (see (Huet et al. 2011) for more details)
- Laws for migration of students and retirees and acceptable distance of commuting (see for details on these processes)

Except for β, all these parameters can be extracted from the Mobility data collected in the French Census, directly or after applying some statistical tools.

Death

The death age of the individual is determined when she enters the simulation (through birth, initialisation or immigration). When an individual dies, its household status is updated depending on the number of remaining members and their statuses, parent or children. Households are eliminated when all their members die, when the couple splits up, or when they simply out-migrate.

Parameter: probability to die by a certain age - made available by INED from the various French Census at the national level.

MakeCouple

The method works as follows:

- During each time step, each single individual (with or without children) has a probability to search for a partner;
- If the individual tries to find a partner, she tries a given number of times in every municipality close to her own (her own included) to find someone who is also single and whose age is not too different (given from the average difference of ages in couples and its standard deviation); she can search among the inhabitants or the potential immigrants; the close municipalities are at a maximum distance defined by the threshold parameter "proximity" except for old people who search for a partner only in their own municipality;
- When a couple is formed, the new household chooses the larger residence (the immigrating households always go into residences of their new partners; this move can force one member to commute very far. This situation can change only when she is becoming the leader triggered by the job search method and implying that the household will aim to move closer to her job location.

Parameters: probability to search for a partner; maximum number of trials; average difference of age of couples and its standard deviation.

The last one is given by the INSEE at the national level based on the data from Census. For the two first, they have to be calibrated since they do not correspond to existing data.

SplitCouple

All couples, except the potential immigrants have a probability to split up. When the split takes place, the partner who works further from the residence leaves the household and creates a new household, which implies that she searches for a new residence. When there are children, they are dispatched among the two new households at random.

Parameter: probability to split (no possible data source, has to be calibrated)

Giving birth

To simplify, we made the assumption that only households with a couple can have children, and one of the adults should be in age to procreate. We assumed that couple has a constant probability to have a child over the years. The parameters are the minimum and maximum ages to have a child and the average number of children by couple. From these parameters, we compute for each couple the probability to have a child during that particular year if one randomly chosen individual's age allows reproduction.

Parameters: minimum and maximum age to give birth, number of children an individual can have during her life on average. Usually ages for reproduction ranges from 18 to 45. That is the usual base to compute the total fertility rate corresponding to the number of children divided by the number of women in age to give birth during any given year. From this rate, it is possible to compute the average number of children of any simulated woman, which is about 2 for France. We can start with this value to parameterize the model. But the number of children per couple has to be calibrated since the observed fertility rate of our simulated population can vary

from the value of the parameter. Indeed, the birth can only occur in couples with members having a relevant age. Consequently, the parameter number of children giving the probability of birth does not correspond with the fertility rate (which is a measure in the population, implicitly resulting from different processes leading to a birth).

2 DESIGNING AND PARAMETERISING THE INDIVIDUAL ACTIVITY

This part focuses on the design and the parameterisation of the individual activity. The purpose is to illustrate how to model in a micro simulation approach individuals' behaviour on a labour market utilising existing data. The European project that funded this work did not fund specific interviews or surveys for this purpose. But, even if such funding had been available, it would have been difficult to have a sufficiently large sample to ensure the statistical significance of the obtained attributes and behaviours. Therefore, it seemed better to use existing large database dedicating especially to the labour force, such as the labour force survey, which gives information on the labour force based on a very large sample and the weights for projection at various levels. Moreover these databases, developed by the National Statistical Office, have been built on a data collection model designed by experts. They represent common knowledge, largely shared by every stakeholder since they are used as references in decisions and predictions.

We start from existing databases and the objectives of the modelling to characterise our agents and their attributes and behaviours. That is what we discuss in the following first subsection. The two following subsections give details on the initialisation of the attributes and on the parameterisation of the behaviours. The link between attributes and behaviours is guaranteed as this data is implemented to ensure its compatibility with the agent attribute modalities. Similarly, the projection of attributes and behaviour for the whole virtual population is easy: an innovative generation population algorithm builds directly a robust and significant population of individuals while the link between modalities of attributes and their evolving rules allows an automatic projection at the population level.

2.1 DATA SOURCES AND MAIN MODELLING CHOICES

This is to identify the agent classes and the structure of agent behaviour in each class. The first steps have been:

- to collect all relevant data source regarding the region we want to simulate considering the exact problem (aim of the project) we need to address;
- to make a state-of-the art;

From the literature and the expertise coming mainly from economists, we identify two complementary groups of dynamics to take into account to model the evolution of a local labour market:

- Job offers and corresponding dynamics;
- Job demand and occupation, and corresponding dynamics.

We identify two possible databases to help us conceptualising and parameterizing the model:

- The Census: it gives indication about the situation of individual when being student, retired, or active and also who is occupied and who is not occupied, what occupations individual have aggregated in socio-professional categories and activity sectors; Census data are available at the municipality level for three different dates 1990, 1999 and 2006. We can also benefit from the mobility tables of the Census giving, at least in 1999, an exhaustive description of the commuting flows between municipalities; French Census data are also available for 1982 but not electronically;
- Labour force survey (from 1990) and census data;

From literature and data, we have to define agents:

- corresponding to the local level of offer: the **municipality**
- corresponding to the job demand and occupation: the **individual** is the one who is going to search for a job, deciding if and where she searches taking into account the **household** of which she is a member and her *municipality* of residence.

Then we have a municipality offering jobs, composed from households, themselves composed of individuals who decide, considering their household, if and where they are going to search for a job. A job can be found in a municipality and individuals accept found jobs based on the distance.

Other available data sources include SIRENE and UNEDIC. The SIRENE database includes information on the number of societies by activity sector. The UNEDIC database includes the number of paid employees by activity

sector. But both these data sources describe only a part of our problem and start only in 2000 while the simulation requires longer periods to allow for a proper calibration of the model.
The incompatible coverage also constrains the choice of agents and their attributes. However, given the available datasets we decide to start simulations in 1990. On the one hand, it means some the parameterisation of some attributes is less robust than with shorter calibration periods. A later start would allow us to use the supplementary information given in more recent surveys and not available in older surveys. For example, we use only four modalities of size to describe the size of dwellings because only four are available in 1990 while five and more are recorded in later surveys. On the other hand, the 1990 census data give us the cross distribution socio-professional categories x sector of activities we use to define the jobs while this cross distribution is not available later. Then, we can and have to use IPF to define the job offer after 1990 starting from the 1990 cross distribution.
The definition of a job is directly driven by the available data. Both Censuses and Labour Force Survey (or Employment survey) describe jobs with profession (socio-professional category) and activity sector. Both also contain data on age and situation (student, retired, actives, occupied or not, inactive) allowing us to make a connection between both sources of data. Moreover, when the data sources are "official", it often corresponds to the common knowledge of stakeholders and other decision makers.
Moreover, as a general modelling good practice, it is particularly important to minimise the number of unknown parameters. Indeed, every parameter which is not derived from the data has to be calibrated. The calibration computational cost increases with the number of parameters. Moreover, the more numerous are the parameters to calibrate, the less relevant also is likely to be the model which, given its large number of freedom degrees, can produce almost any trajectory.

2.2 DEFINING THE INITIAL INDIVIDUAL LABOUR ATTRIBUTES

The main source of information to define attributes and their values is Census data. The French Census is available for 1990, 1999 and 2006. The 2006 Census has to be used with caution since it is different from 1990 and 1999. It is now a continuous survey which interviews a part of the population every year. Municipalities having less than 10000 inhabitants are exhaustively surveyed by 1/5 every year. Larger municipalities are sample surveyed every year. In both cases, INSEE, responsible for the Census, give the information allowing the projection at the population level every year. A very good point is that the access to data is easy and free[6].
To compute a population with sufficiently realistic local statistical properties for individuals and households, we propose an algorithm described in (Gargiulo et al. 2010) presenting the generation of households in the Auvergne Region. An improved version has been developed for generating the Cantal population. To summarize our algorithm, we build for each municipality a list of agents with the exact number of individuals being each age and a list of households with the exact number of household members. Then, we try to fill one by one each household with individuals taking into account the probability of households having some particular properties, such as being a couple or having a given number of children. Each time a household is completed, another one is selected to be filled. At the end, we have a virtual population of households following the exact distribution of sizes, having good statistical household properties and composed from individuals following the exact distribution of ages. To built the initial population of Cantal, our algorithm uses for each municipality:

- The distribution of the size of households – available at the municipality level in 1990
- The distribution of ages of individuals – available at the municipality level in 1990
- The distribution of ages of the reference person of households – available at the municipality level in 1990
- The distribution of household types (single, couple, couple with children, single-parent, other) - available at the municipality level in 1990
- The distribution of age differences for couples – only available at the national level in 1990
- The distribution of the probability to be a child (i.e. living at parental home) by age and for each household type – available at the municipality level in 1990

This generation method is different from the nowadays used IPF (Iterative Proportional Fitting) which reweight a measured population under some constraints to obtain a virtual population representing the one the modeller is interested in. However this method can not control the attributes at both levels, the person and the

[6] made available by the Maurice Halbwachs Center of the Quételet Network (http://www.reseau-quetelet.cnrs.fr/spip) for 1990. For 1999 and 2006, they are directly accessible through internet via the website of INSEE http://www.recensement-1999.insee.fr/ and http://www.insee.fr/fr/publics/default.asp?page=communication/ recensement/ particuliers/diffusion_resultats.htm)

household. Some recent work proposed a hierarchical IPF (Müller and K.W. 2011) to control the two levels but they still required an initial sample, which can be reweighted to fit the scale the model is interesting in.

After the virtual population has been built, individuals require a labour market status. That means the following four individual attributes have to be parameterised during the initialisation: Activity status; Profession, approximated by the socio-professional category; Sector of activity to define, with the profession, the occupied job; Place of work.
To characterize the status we distinguish between active and inactive individuals. Active people can be employed or unemployed. For non-active people we distinguish three categories: students, retired and other. No further characterization is required for non-active person. On the contrary, active people, both employed and unemployed require a socio-professional category (SPC) defining their profession. Moreover, employed individuals require a sector of activity defining the occupation (see 1.1.3 and 1.1.5 for details). Once the municipality of employment is determined, the employed individual is successfully parameterized.

Figure 2 | Algorithm for the initialization of the activities for Auvergne case study

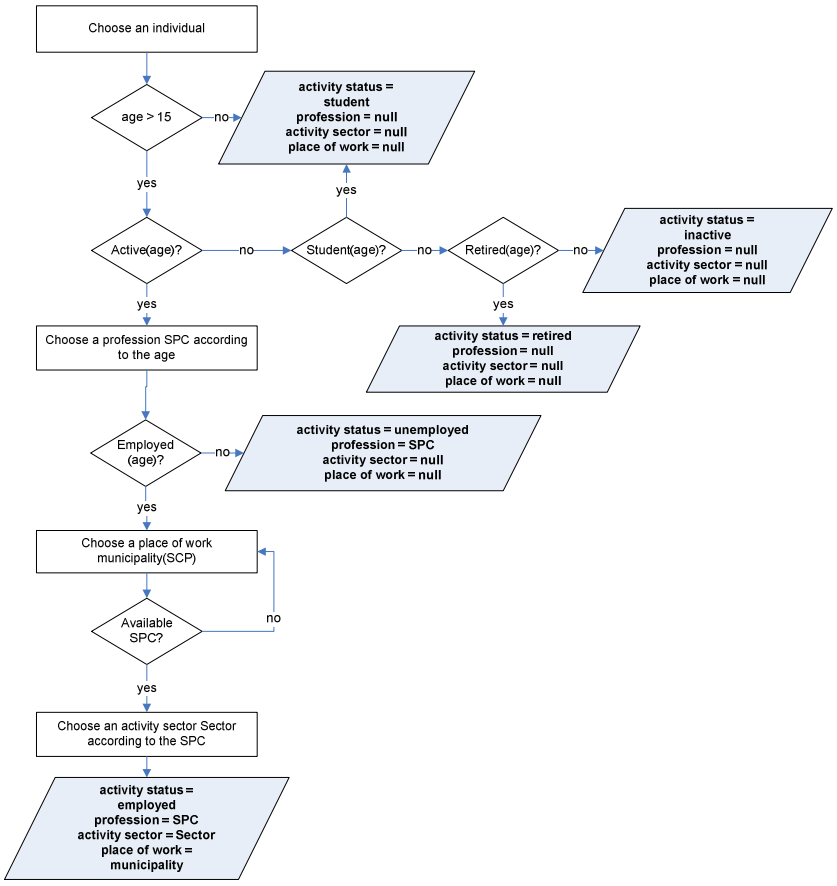


Figure 2 shows the generation algorithm. The initialization of the activities starts from the population of households previously generated for each village: each person is assigned an activity, according to the characterization presented above. All the individuals younger than 15 are automatically considered students. For all the others the first step is the decision about being active or not. This decision depends on the age of the person. If the person is not active then her age determines whether she is retired or a student. If she is neither student nor retired, she will be identified with the status "inactive". If the person is active, the first step is the selection of the socio-professional category (SPC). This choice depends on the age. Secondly it is decided whether the person is employed or unemployed, according to the age. If she is unemployed, no further choices are needed. If she is employed, the municipality of employment is determined. The municipality of employment depends on two questions: first, does she work inside her municipality of residence? If no, find at random a place of work among the possible places of work starting with her own municipality of residence if

employment is available according to the SPC. The possible places of work are defined through a generated virtual network built from the mobility data of the French Census of 1999 (see the generation model proposed in (Gargiulo et al. 2011) and improved in (Lenormand et al. 2011b)). Finding a possible place means the individual can find a free job partly defined by the same SPC as hers. A vector for available jobs is maintained (corresponding to the total number of commuters-in at the beginning of the initialisation) for each municipality and decreases with individuals filling vacancies. If no vacancies remain among the possible places of work while an individual is still looking for employment, the attribution of a place of work among the possible ones is forced. Indeed, this can occur due to the fact the generated virtual network is built under the only constraints related to the job demands and the job offers of each municipality. The virtual network doesn't consider the SPC then it can't ensure a demand with a particular SPC can be satisfied by an offer with this SPC in the set of municipalities it has fixed as possible places of work. Finally, an activity sector is attributed to the employed individual based on the cross distribution SPC. We have to acknowledge that the French Statistical Office, as many Statistical Offices, use two ways to count the jobs: counted on the place of residence – that means corresponding to the job occupation by people living in a municipality wherever they work; and counted on the place of work – that means counted on the municipality where people work wherever they live. The algorithm uses the following data for each municipality of the set:

- Age x activity status counted on the place of residence
- Age x SPC for actives counted on the place of residence
- Distribution of probabilities working inside her place of residence by SPC
- A generated commuting network through (Gargiulo et al. 2011) (Lenormand et al. 2011b) given for each municipality the distribution of commuters out to each of the other municipality
- SPC for actives x activity sector counted on the place of work

2.3 DEFINING THE INDIVIDUAL BEHAVIORAL RULES REGARDING ACTIVITY

This part is dedicated to the parameterisation of events on the labour market. Characterization and parameterization is required for those rules that change the value of the individual's attributes related to its labour activity: Activity status; Profession, approximated by the socio-professional category; Sector of activity to define, with the profession, the occupied job; Place of work.

The main data source to do so is the European Labour Force Survey, and particularly its French declination called in French "Enquête Emploi", meaning "Employment survey". The data are kindly made available for free by the Maurice Halbwachs Center of the Quételet Network[7]. This Employment survey was launched in 1950. It was redesigned in 1968, 1975, 1982, 1990 and 2003. From 1982, the survey became an annual survey. Since the last redesign the survey is implemented continuously to provide quarterly results. The resident population comprises persons living on French metropolitan territory. The household concept used is that of the 'dwelling household': a household means all persons living in the same dwelling. It may consist of a single person, or of two families living in the same dwelling.

As our approach starts the simulation in 1990 the first period is based on annual data while from 2003 on values can be considered in quarterly time steps (Goux 2003) (Givord 2003. The data to select from these two periods vary a bit due to the structural and practical changes in the survey).

Coming back to the description of the whole data, the sample sizes of the data varies from 168883 to 187326 from 1990 to 2002 each year and from 92300 to 95647 each quarter a year for the new Employment survey. The individuals are asked a very comprehensive series of questions from 1990 to 2006, related to their work. In particular, we can follow their situation year by year, and also their wishes to change job and the type of job they are looking for. Table 2 shows the variables we extract from the databases to compute the probabilities we need. However, for the sake of simplicity, we use only data from 1990 to 2002 to explain how to extract the information we need from the data..

Table 2. Data to extract from the various databases of the French labour force Survey to compute the probabilities related to working status of the individual

1990 to 2002	***2003***	***2004***	***2005***	***2006***	***2007***	**Meaning of the variable**
ag	Ag	Ag	ag	Ag	Ag	Age
annee	annee	Annee	annee	annee	annee	Year of interview

[7] http://www.reseau-quetelet.cnrs.fr/spip/

dcse	csepr	Csepr	csepr	csepr	csepr	Socio-professional category
cspp	cspp	Cspp	cspp	cspp	cspp	Socio-professional category of the father
dcsep	cser	Cser	cser	cser	Cser	Socio-professional category one year before
dcsea	cslong	Cslong	cslongr	cslongr	cslong	Socio-professional category which has been occupied for most of the time [for inactive and unemployed people]
tu99	tu99	tu99	tu99	tu99	tu99	Urban area type
fip	eoccua	Eoccua	eoccua	eoccua	eoccua	Occupation one year before
extri	extriA, extriA04	extri99 extri04, extri05,	extri05, extri06,	extri06	extri06	Weights making the interviewed individuals representative (depending on the census done 1999 or of the first result from the last French census (in 2004, 2005, 2006)
rg	reg	Reg	reg	reg	Reg	Region of residence
fi	sp00	sp00	sp00	sp00	sp00	Occupation during the month of interview
-	trim	Trim	trim	trim	trim	For the second period of the survey, the only keep the first quarter of the year.
csrech	csrech					Searched socio-professional category
dre1						Situation in regards to employment (mainly to use dre1=5 meaning people looks for a job (or another job))
	soua ; mrec					Wish another job; Is the individual has searched for a job during the last four weeks?

From the databases, we considered only the population being more than 14 that is not military people of students (FI = 3 and 4).

2.3.1 ENTERING THE LABOUR MARKET

A first step consists of extracting the age from which on the individual is going to look for a job. This will determine the age at which a student status changes to a "on labour market" status. We consider in the period 1990 to 2002 the value FIP=3, which means that the individual was student the year before and the value FI=all the possible values except 3 means that the individual is not a student anymore. Then, for each five-year step we compute the probability to be a given age and having entered on the labour market for every year.

We used the weights to obtain a projection of the data at the Auvergne level. Auvergne is the region containing the Cantal “département” and three others. That is the closer significant and representative level of the Cantal. Then, we assume the probabilities are the same at the regional and the "département" level.

The second step is to allocate a first SPC (proxy used for defining the profession) to the individual allowing us to approximate what she is going to look for. We know that both these variables, the age of entry and the first SPC, are not independent. Moreover, a social determinism rules the choice of the profession by children compared to the profession of their parents. Figure 3 presents such a relation for the Auvergne population. It shows, for example, that almost only farmers’ children become farmers or that executives’ children mainly become executives and/or adopt an intermediary profession.

Thus, starting from this social determinism, we have some indications to set the SPC of children. However, we also have to decide the age of entry in the labor market, and we know that this age is not independent from the level of education, which can be related to the SPC. Consequently, we apply a two-time process which, at first, decides the age at which to enter the labor market using the father’s SPC and then determines the child’s SPC depending on the age of entry.

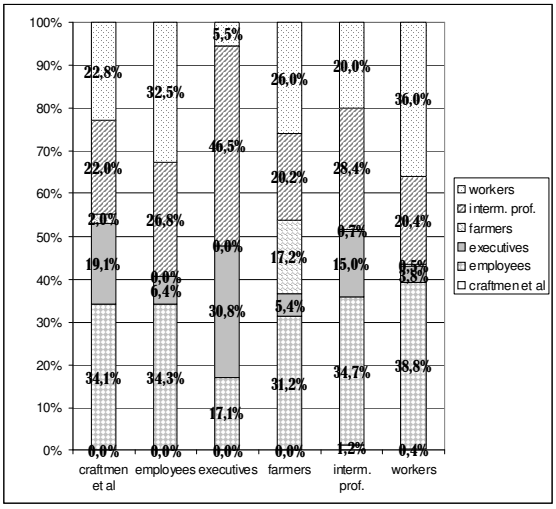


Figure 3. Distribution of SPCs choices by children regarding the father's SPC (in abscissa) for the Auvergne population. Source: French Labour Force Survey, 1990 to 2002 data.

The age of entry on the labour market is determined by the SPC of the father. Since the individual has no gender in our model, the father is randomly chosen between the two parents when there are two.
A criticism can be formulated to this approach since the SPCs of the couple members is not controlled, while we know from the literature that the partner is not chosen at random regarding her SPC (Bozon and Héran 1987). The homogamy can be explained by the constraint associated to the meeting places (Bozon and Héran 1988). It has been identified as a possible next step for modelling.

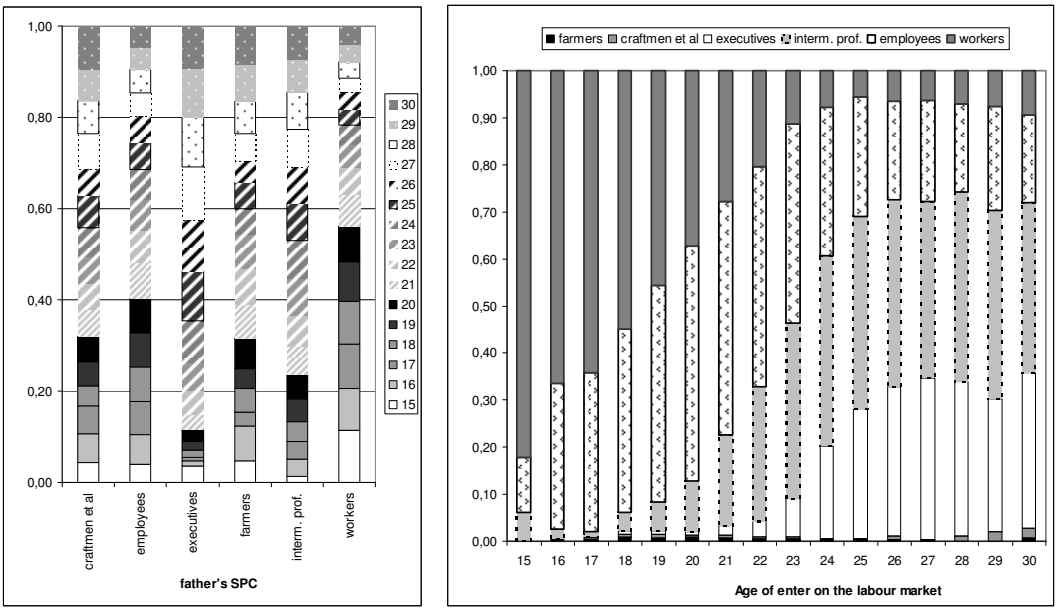


Figure 4. (a on the left) Probability of a "first" SPC depending on the age of entry in the labour market; (b on the right) Distribution of probability to enter the labour market at a given child age for each of the six father's SPC considered – French population. Source: French Labour Force Survey, 1990 to 2002 data.

Figure 4a shows the distributions of probabilities to enter the labour market depending on the various ages of a child for each of the six SPC attributed to the father. We can for example read that if the father is an executive, the probability to enter on the labour market before 20 is only 0.1 while it is more than 0.5 if the father is a worker. Once our individual has an age to enter the labour market, we can determine her first SPC. Figure 4b shows for each age of entry on the labour market (abscissa) the distribution of probabilities over the possible

SPC to provide the individual with a first SPC. For example, one can notice how high the likelihood of looking for a worker position for the individual looking at first for a job at 15 is, while at 30, she will mostly look for intermediary or executive positions. The individual who enters the labour market can decide looking for a job.

2.3.2 INDIVIDUAL JOB SEARCHING DECISION

We assume that the probabilities are stable in time for the Auvergne region. Thus, we mix the data from the years 1990 to 2007 in a single sample. Starting from the variables presented in the table 2, we count the frequencies of transitions between inactive, unemployed, employed, from one year to the following. For each counted transition, we take into account the weight of the related individual in order to have a probability quantified for the Auvergne level.
Finally, we calculate the probability to reach a given situation by dividing the total obtained for a transition starting from the situation *x* by the sum of all the totals related to the transitions starting from this same situation *x*.
We focus on the municipalities of the Auvergne region having less than 50000 inhabitants using the area type "tu99".

From and to the inactive status

The following variables are used to extract the transitions from a starting situation to an arriving situation. They are used for the transitions from and to the inactive status.

- fip = 7 plus 8 or EOCCUA = 6 plus 7 to define the inactive status as starting situation; fi = 7 or SP = 8 to define the inactive status as arriving situation;
- fip = 2 or EOCCUA = 2 to define the unemployed status as starting situation; fi = 2 or sp00 = 4 to define unemployed status as an arriving situation ;
- fi = 1 or EOCCUA = 1 to define employed status as starting situation;
- DCSP or DCSA are used to define to starting SCP for unemployed and employed while DCSE is used to define the arrival SCP (for unemployed).

The table 3 shows the extracted probabilities for the Auvergne region.

Table 3. Probabilities of the transitions "inactive → unemployed", "unemployed → inactive depending on SPC", "employed → inactive depending on SPC"

		Arriving situation	Inactives	Unemployed					
Starting situation	**Starting SCP**	**Arriving SCP**		farmers	craftmen et al	executives	interm. profes.	employees	workers
Inactives				0,00005557	0,00055947	0,00031037	0,00172877	0,00644310	0,00604629
Unemployed	farmers		0,05462738						
	craftmen et al		0,06335331						
	executives		0,11808481						
	interm. profes.		0,06202433						
	employees		0,07066007						
	workers		0,06165634						
Employed	farmers		0,00650018						
	craftmen et al		0,01423226						
	executives		0,01729000						
	interm. profes.		0,01192824						
	employees		0,00930251						
	workers		0,01129013						

Probability to look for a job with a given profession

The probabilities are computed using the same method we used to compute the probabilities of transitions of activity status. The difference is that we use the answers to the questions about the fact that the interviewee looks for another job. For the first period, we select the employed individuals (fi = 1) looking for a job (dre1=5). For the second period of the survey, from 2003 to 2007, we assume people look for a job if they have answered SOUA=1 (want to have another job) and MREC = 1 (have searched for recently) or SOUA=1 and MREC = 2 and NTCH =1 or 2 (have not recently search for because they wait for answer to recent applications or they have been ill for a while).

Deciding looking for a job when unemployed

Unemployed people are assumed to be those who search for a job. Even if, in the labour force survey, only 80% of unemployed people declare searching a job, we assume the probability to search for a job of unemployed

people is one. Indeed, if we consider the whole model, it globally underestimates the job offer and the probability to find a job. This is difficult to correct as, for instance, we cannot consider that in most cases a job offer is proposed before it has been quit while the model time step is not less than one year. Also we assume the job offer equal to the job occupation. Then, the probability to search for a job of unemployed people is one in order to compensate a bit this underestimation and be able to occupy every job offer (which is the state the model has to reach). The data indicates the probability to look for a job for unemployed individuals is quite stable until 54 years of age and dramatically decreases for older individuals. A second step of the modelling work would be to see if this dramatic decrease needs to be considered. We also analyse how different parameters describing the household (the number of unemployed in the household, the number of children, or the type of household) influence the probability to look for a job, and we did not find any clear dependency.

The probability to begin searching (i.e. becoming unemployed) if an individual did not search previously (not because she is employed) corresponds in the model to the transition from inactive to unemployed. As already mentioned, it is the complementary value for each age range of the value to make the transition from inactive to inactive.

Since an individual is unemployed, it is necessary to define which SPC she is going to search for. It varies a lot with the current SPC of the individual. As shown in Table 4 even if there is a tendency to look preferentially for her own SPC, an unemployed individual can prefer changing SPC. That is particularly the case of farmers and craftsmen. Then, we parameterise the process from the computation of the probability distribution to choose a SPC knowing the current SPC.

Table 4. Probability for unemployed people to search for a job with various SPCs knowing the current SPC of the individual

SPC / Looks for	Farmers	craftsmen et al	executives	interm. prof.	employees	Workers
Farmers	0.000	0.000	0.000	0.177	0.376	0.447
craftsmen et al	0.000	0.079	0.012	0.088	0.443	0.377
Executives	0.000	0.037	0.499	0.256	0.171	0.037
interm. prof.	0.000	0.009	0.053	0.591	0.273	0.074
Employees	0.003	0.007	0.006	0.063	0.808	0.113
Workers	0.006	0.010	0.003	0.056	0.251	0.674

Deciding looking for a job when already employed

We consider those respondents being employed who answered that they are looking for another job. We have the age of these people, as well as the type of their current job. The analysis shows that the age is a very significant variable for determining if an employed individual looks for another job (see Figure 6a). Young people are more susceptible to look for another job and this tendency decreases with age.

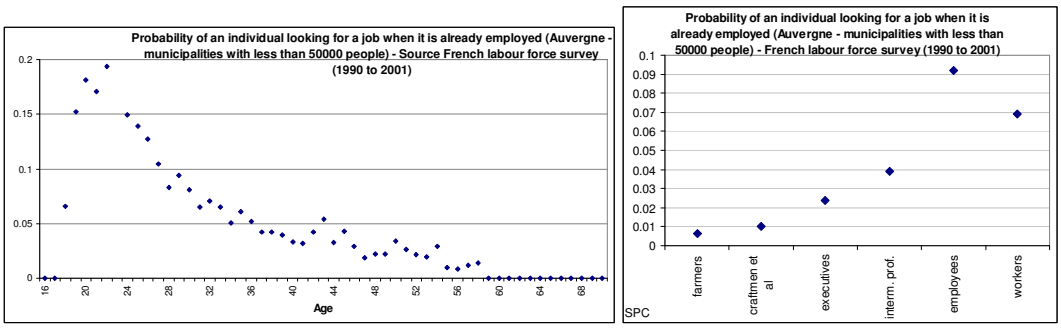


Figure 6 – (a) Probability for an already employed individual to look for another job according to the age (on the left); (b) Probability that an already employed individual looks for another job according to socio-professional category (on the right).

The SPC is also a significant variable to predict the probability to look for a job (see Figure 6b). Some SPC, such as employed farmers or craftsmen are not very susceptible to look for another job. On the contrary, others, such as workers and especially employees have quite a high probability to look for another activity.

Table 5 shows the parameter values for the decision searching for a given profession when the individual is already employed for some age ranges. For employed people, we built a probability containing the both information have decide to search for a job and what she searches for. It is important to point out that the probabilities presented in Table 5 do not add up to one but to the overall probability to search, which is quite low for already employed people.

Table 5. Extract of probabilities for employed people with a given SPC and a given five-year old age to look for a job within a given SPC.

Age Range	Looks for/ Is a	farmers	craftmen et al	executives	interm. prof.	employees	workers
15	Farmers	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	craftmen et al	0.0000	0.0000	0.0000	0.0000	0.0011	0.0014
	executives	0.0000	0.0000	0.0000	0.0000	0.0010	0.0000
	interm. prof.	0.0000	0.0000	0.0000	0.0000	0.0143	0.0040
	employees	0.0000	0.0000	0.0000	0.0000	0.1319	0.0168
	Workers	0.0000	0.0000	0.0000	0.0000	0.0162	0.0498
. . .	Farmers	. . .	. . .	. . .	. . .	. . .	. . .
	craftmen et al	. . .	. . .	. . .	. . .	. . .	. . .
	executives	. . .	. . .	. . .	. . .	. . .	. . .
	interm. prof.	. . .	. . .	. . .	. . .	. . .	. . .
	employees	. . .	. . .	. . .	. . .	. . .	. . .
	Workers	. . .	. . .	. . .	. . .	. . .	. . .
55	Farmers	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	craftmen et al	0.0000	0.0000	0.0000	0.0000	0.0002	0.0002
	executives	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000
	interm. prof.	0.0000	0.0000	0.0000	0.0000	0.0030	0.0005
	employees	0.0000	0.0000	0.0000	0.0000	0.0274	0.0021
	workers	0.0000	0.0000	0.0000	0.0000	0.0034	0.0062

2.3.3 INDIVIDUAL SEARCHES FOR A JOB

Since the individual knows which profession she wants to search for, she has to find a place where to look for a job. Firstly, the individual selects an accepted distance she would want to commute. The next section presents how to the related probabilities. If the chosen distance is higher than zero, the individual has to decide if she is going to work outside her set of municipalities. The law allowing this decision and the way to extract it from data is the subject of what follows in the next section. In case the individual has not found a job, she revises the maximum distance. She revises the distance up to 10 times.

The probability to accept a distance to cross over to work

The distance of search for a job is selected from a probability law giving the probability to accept a certain distance between the residence and the work place. The principle is very simple: the probability to commute at a given distance *i* [*pc*(*i*)] is assumed to be the product of a probability to accept a certain distance *i* [*pa*(*i*)] by the pay offered at *i* [*Oi*] with a renormalisation coefficient *k*: $pc(i) = k\, pa(i) * Oi.$

Then, it is possible to extract the probability to accept a given distance (*pa*) to work place, which will be used in the model. This procedure, coupled to an appropriate job offer, will allow maintaining the statistical properties of the *pc* distribution over the time of the simulation.

We extract from the mobility data of the 1999 Census for every municipality of the Auvergne region data on commuting (*pc*) and data on job occupations, which we assume to be equivalent to job offers (*O*). Evidently, the number of occupied jobs is used as a relevant proxy for the job offer of a municipality. An exhaustive description of the work allowing to build this probability law is given in (Felemou 2011).

Figure 5 shows an example of commuting data probability distribution (DDC = *pc*) and of job offer probability distribution (DOE = *O*) for one randomly chosen municipality.

A classification of acceptable distance distributions shows municipalities can be classified in three different groups, apparently depending on the size of the municipality of residence (see Figure 6 on the right). Thus, we assume for this parameter three probability distributions shown on the left of Figure 6 for three different size-

dependent classes of municipalities (to the right of Figure 6). The data suggests that the larger the municipality, the lower the probability to work in the place of residence and the longer the commuting distance.

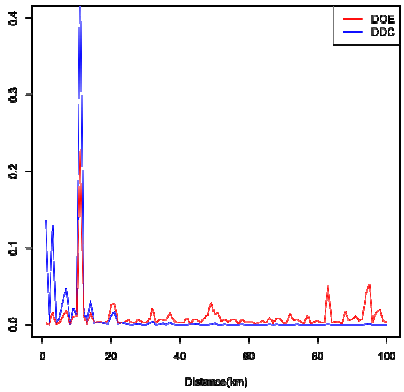


Figure 5 - Example for one municipality of the density distribution of job offers (DOE=*O*) and the one of commuters (DDC=*pc*)

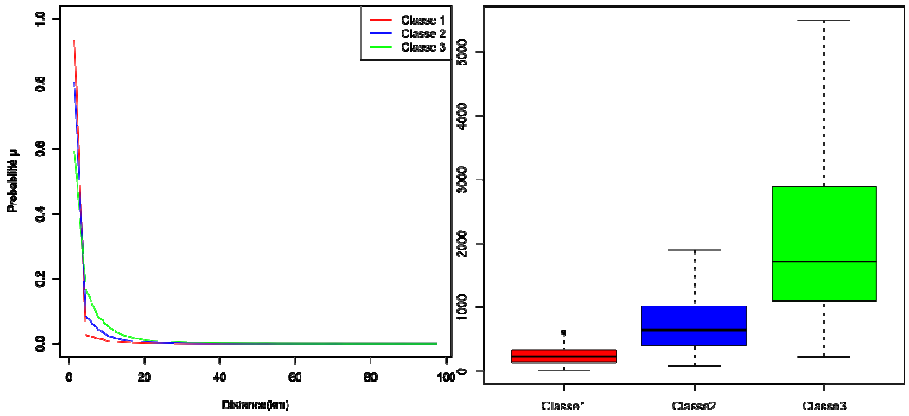


Figure 6 - Probability laws that an individual accept to a certain commuting distance knowing that a job is available for it.(on the left) - Different population sizes for the municipalities of each sub-group (on the right)

It is important to emphasise that only if the selected distance is higher than zero, the individual has to decide if she is going to outside or inside the set.

Going to work outside the set

When the individual is commuting – meaning she has picked out a distance of research higher than 0 – she has to check if she has a chance to commute outside considering her place of residence. Indeed, an individual living close to the border of the set has a higher probability to commute outside the set. Then, the individual chooses at random to work outside depending on the probability associated with her municipality of residence. Each municipality has such a probability which is a function of its distance to the border of the set. This function is extracted from the mobility data from 1999 (Source: INSEE). Figure 7 shows this function for the Cantal department and the whole Auvergne region of which Cantal is a part. Both laws are quite close and it appears relevant to use as a parameter the law extracted for the whole region since it is probably less noisy.

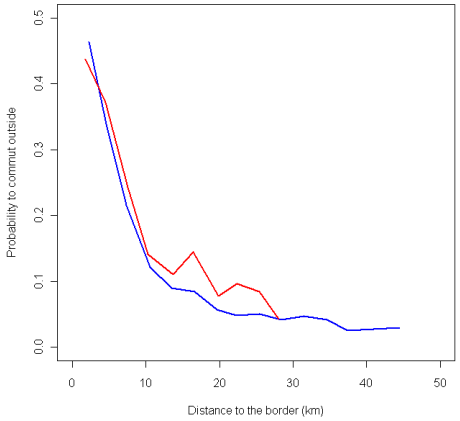


Figure 7 - Probability to commute outside the set (ordinate) depending on the distance of the municipality of residence to the frontier of the set (abscissa in Euclidian kilometers) - Red: Cantal; Blue: Auvergne

We are now describing how to extract the probability law for the final event which is going on retirement.

2.3.4 GOING ON RETIREMENT, AND STOP SEARCHING FOR A JOB

To extract the transition to the retirement, we consider, in the period 1990 to 2002, the value FIP=all except 5 or 6, which means that the individual has not yet retired and the value FI=5 or 6, which means that the individual is now retired. We assume that the retiree does not search for a job anymore since this is generally the case true in France. Figure 8 shows that the speed of transitioning into retirement varies a lot from one SPC to another: we can read for example that at 60, 63 % of workers are retired while only 17 % of farmers are retired. Then, instead of considering a generic retirement law for all the individuals we consider a law for each SPC. Indeed, as these laws influence the job availability at a given moment it is very important to be sufficiently precise.

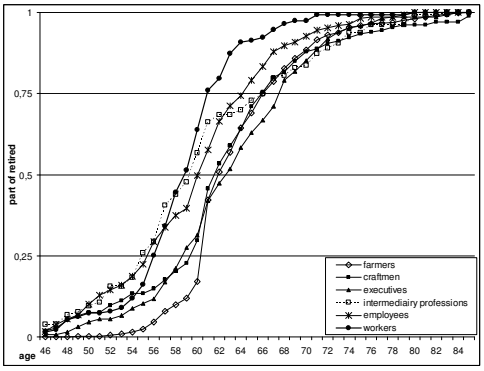


Figure 8. Speed of going into retirement by SPC (source LFS) – France level

3 LESSONS / EXPERIENCE

First, we want to stress the necessity to not only consider the objectives of the model during the design, but from the very beginning exploring existing data sources and studying the implicit model beside the existing databases. The availability of data and the more or less implicit model guiding the collection of data constrain the definition of agents, their attributes and behaviours.

Using large existing databases can appear more relevant, especially the "official" ones from the National Statistical office, than collecting a small sample and reweighting it to obtain a statistically significant artificial population.

For these large databases, the models guiding the collection of data represent the expertise knowledge and generally assume some dynamics, particularly if time series are collecting during the survey. Moreover, if the data sources are collected by the National Statistical Office, they probably represent the commonly used information and knowledge by the stakeholders and policy makers. A model which aims to inform decision making is more useful if it can be easily understood and discussed by the relevant decision makers. This is easier if the model starts with common knowledge.

More generally, the modeller has to identify the rationale behind the considered data sources and use it to build the dynamic model. Indeed, this rationale often makes some implicit assumptions on the dynamics. Let's take the definition of a household as an example. "*In surveys prior to 2005, people were required to share the same main residence to be considered as households. It was not necessary for them to share a common budget. De facto, a household corresponded to a dwelling (main residence*)". Thus, until 2005, the French National Statistical Office (INSEE) assumes the household/family is defined by the place where it lives, which is unique. Indeed, following the INSEE definition, each person in a household may belong to only one family. In this framework, residential mobility is a household/family decision and the number of occupied dwellings in a place corresponds to the number of resident households. That is also what we assume in the model. "*Since 2005, a dwelling can include several households, referred to as "living units". Every household is composed of the people who share the same budget, that is who contribute resources towards the expenses made for the life of the household; and/or who merely benefit from those expenses*." The new definition is based on the fact that related or unrelated individuals can share the same budget and have a habitual residence (the dwelling in which they usually live). This new definition takes into account some cultural evolutions and allows a European homogenization of the way households are defined. However, it modifies the way the dynamic of move can be considered since each individual of the household can have more than one dwelling. This is to point out that the choice between one data source and another corresponds to a representation of the world to which some particular dynamics can be linked. If the first definition of household is more related to the idea that relationships between people can be identified by the concept of family and/or the identical of place of living, the second definition puts the economical constraints (i.e. the sharing budget) much more at the heart of the dynamics of closeness. A modeler, having the choice between a data source containing data built on the first definition and another one based on the second definition, should be aware of the choice to make and communicate about it. Thus, choosing to only use data on the SCP and the activity sector to describe a job while it is possible to use the salary, which is available in some databases, makes having an occupation much more important than the level of salary. It also implies, for example, that an individual can change jobs just to change their working environment. Differently, the classical economic models considering job change start from the salary and assume an individual changes to increase their salary. We simply assume our individual wants to change jobs, without necessarily changing SCP at the same time. However, one can notice our assumption is relevant due to the existence of a minimum salary in France which ensures a minimum amount of money to live with.

The choice of existing databases for facilitating model design and parameterisation needs to consider:

- a longer as possible period of calibration: indeed it is not sufficient to strongly link the model to data if the model is not calibrated or calibrated with poor data compromising the robustness of the trajectory of underlying model dynamics;
- a sufficient number of modalities for each attribute in order to be able to reproduce the diversity of relevant agent types and behaviours. For example, we chose to aggregate in our work jobs in 24 types; at the end this depends on data availability;
- a minimum number of variables to calibrate: too many unknown parameters implies we don't know much about the dynamics and every explanation for observed trajectories can be valuable;
- the possibility to use them simultaneously for initialising agent attributes and defining agent behaviours: that means in particular that they have to have common variables allowing for a link between them. The challenge is to make an easy fit between attributes and behaviours.

Finally, starting from large national databases makes it likely that the model can be easily implemented and parameterized in another country. For instance, the example on the individual dynamics of activities indicated the possibility to apply the model in another European country even if some small adaptations are required. Indeed, Europe tends to harmonise the data bases in order to have common indicators at the European level. Then, large national databases have been designed or redesigned for answering the European demand. For example, the French "Employment survey" is the data source for the French contribution to the European Labour Force Survey. That is why (Baqueiro Espinosa et al. 2011) proposes a way to parameterise our model directly starting from the data of this European survey. For the same reason, national census data in Europe tend to consider more and more comparable or identical variables. That makes it possible to use them to parameterise our model even if a particular attention to the definition of used concepts remains: while to be a retiree in France (at least until a very recent period) means not looking for a job, it is not the case in UK for example.

Taking into account data at an early stage is not an easy task. It is at the same time laborious and confusing since the modeller is confronted with a very large set of information and more or less implicit knowledge. Finding a way to use the data and to choose the object, their attribute and the dynamics in order to remain simple as possible is much more demanding than developing a theoretical model. However, for such complex systems and models as ours that focus on the dynamics of interacting municipalities, the approach allows to properly define and control some sub-dynamics, even if they are not independent from other dynamics in order to test hypothesised system properties. For our concerns, we expect the expertise we developed for the labour market in conjunction with the robust parameterisation of the individual activity dynamics and job offer dynamics, will allow us to better understand how the demography impacts on the population/depopulation phenomena and how these phenomena impact on demography in return.

4 ACKNOWLEDGEMENTS

This work has been funded under the PRIMA (Prototypical policy impacts on multifunctional activities in rural municipalities) collaborative project, EU 7th Framework Programme (ENV 2007-1), contract no. 212345

5 REFERENCES

Aubert, F., Dissart, J.C., and Lépicier, D. (2009), 'Facteurs de localisation de l'emploi résidentiel en France', *XLVIème Colloque de l'Association de Science Régionale de Langue Française (ASRDLF),* 6-8 juillet, Clermont-Ferrand, France, 27.

Ballas, D., Clarke, G.P., and Wiemers, E. (2005), 'Building a Dynamic Spatial Microsimulation Model for Ireland', *Population, Space and Place,* 11, 157-72.

Ballas, D., et al. (2007), 'Using SimBritain to Model the Geographical Impact of National Government Policies', *Geographical Analysis,* (39), 44-77.

Baqueiro Espinosa, O., et al. (2011), 'Two adaptations of a Micro-Simulation Model to Study the Impacts of Policies at the Municipality Level', in PRIMA European project (ed.), *Working paper* (IAMO, Newcastle University, Cemagref), 62.

Berger, T. and Schreinemachers, P. (2006), 'Creating Agents and Landscapes for Multiagent Systems from Random Sample', *Ecology ans Society,* 11 (2), 19.

Birkin, Mark and Clarke, Martin (2011), 'Spatial Microsimulation Models: A Review and a Glimpse into the Future', in John Stillwell and Martin Clarke (eds.), *Population Dynamics and Projection Methods* (Understanding Population Trends and Processes, 4: Springer Netherlands), 193-208.

Birkin, Mark and Wu, B. (2012), 'A review of Microsimulation and Hybrid Agent-Based Approaches', in A.J. Heppenstall, et al. (eds.), *Agent-Based Models of Geographical Systems* (Springer), 51-68.

Blanc, M. and Schmitt, B. (2007), 'Orientation économique et croissance locale de l'emploi dans les bassins de vie des bourgs et petites villes', *Economie et Statistique,* 402, 57-74.

Bousquet, F. and Le Page, C. (2004), 'Multi-Agent Simulations and Ecosystem Management: a review', *Ecological Modelling,* 176, 313-32.

Bozon, M. and Héran, F. (1987), 'La découverte du conjoint : I. Evolution et morphologie des scènes de rencontre', *Population,* 42 (6), 943-85.

Bozon, M. and Héran, F. (1988), 'La découverte du conjoint : II. Les scènes de rencontre dans l'espace social', *Population,* 43 (1), 121-50.

Brown, D.G. and Robinson, D.T. (2006), 'Effects of Heterogeneity in Residential Preferences on an Agent-Based Model of Urban Sprawl ', *Ecology and Society,* 11 (1), 46.
Brown, D.G., Aspinal, R., and D.A., Bennett (2006), 'Landscape Models and Explanation in Landscape Ecology - A Space for Generative Landscape Science', *The Professional Geographer,* 58 (4), 369-82.
Coulombel, N. (2010), 'Residential Choice and Household Behavior: State of the Art', (Working Paper 2.2a: Ecole Normale Supérieure de Cachan), 69.
Davezies, L. (2009), 'L'économie locale "résidentielle"', *Géographie, économie, société,* 11 (1), 47-53.
Deffuant, G. and al., et (2001), 'Rapport final du projet FAIR 3 2092 IMAGES : Modélisation de la diffusion de l'adoption de mesures agri-environnementales par les agriculteurs (1997-2001).'.
Deffuant, G., Huet, S., and Amblard, F. (2005), 'An individual-based model of innovation diffusion mixing social value and individual payoff dynamics', *American Journal of Sociology,* 110 (January) (4), 1041-69.
Deffuant, G., Skerrat, S., and Huet, S. (2008), 'An agent based model of agri-environmental measure diffusion: what for?', in A. Lopez Paredes and C. Hernandez Iglesias (eds.), *Agent Based Modelling in Natural Resource Management* (Universidad de Valladolid: INSISOC), 55 - 73.
Deffuant, G., et al. (2002), 'How can extremism prevail ? A study based on the relative agreement interaction model', *Journal of Artificial Societies and Social Simulation,* 5 (4).
Dubuc, S. (2004), 'Dynamisme rural: l'effet des petites villes', *L'Espace Géographique,* 1, 69-85.
Felemou, Mamourou (2011), 'Analyse de données de flux de navetteurs. Extractions de modèles', (Rapport de stage de 1ère année de Master Statistiques et Traitement de Données; Aubière: Cemagref), 31.
Fernandez, L.E., et al. (2005), 'Characterizing location preferences in an exurban population: implications for agent-based modeling', *Environment and Planning B: Planning and Design,* 32 (6), 21.
Fontaine, Corentin and Rounsevell, Mark (2009), 'An agent-based approach to model future residential pressure on a regional landscape', *Landscape Ecology,* 24 (9), 1237-54.
Gargiulo, F., et al. (2010), 'An Iterative Approach for Generating Statistically Realistic Populations of Households', *PLoS One,* 5 (1), 9.
Goux, D. (2003), 'Une histoire de l'enquête Emploi', *Economie et Statistiques,* 362, 41-57.
Grimm, V., et al. (2010), 'The ODD protocol: A review and first update', *Ecological Modelling,* 221, 2760-68.
Holme, E., et al. (2004), 'The SVERIGE spatial microsimulation model: content, validation and example applications. ', (Technical report: Spatial Modelling Centre, Sweden), 55 p.
Huet, S., et al. (2011), 'Micro-simulation model of municipality network in the Auvergne case study', in PRIMA Working paper (ed.), (Aubière, France: IRSTEA), 63.
INSEE, division "Redistribution et Politiques Sociales" (1999), 'Le modèle de simulation dynamique DESTINIE', *Série des documents de travail de la Direction des Etudes et Synthèses Economiques*, 124.
Lenormand, M., Huet, S., and Deffuant, G. (2011a), 'Deriving the number of jobs in proximity services from the number of inhabitants in French rural municipalities', *Regional Science,* submitted, 13.
Lenormand, M., Huet, S., and Gargiulo, S. (2011b), 'A commuting generation model requiring only aggregated data.', in European Social Simulation Association (ed.), *ESSA 2011* (Montpellier), 16.
Lenormand, M., et al. (2011c), 'Evaluating the number of proximity service jobs per inhabitant in small French municipalities', (PRIMA report: Cemagref, LISC), 12.
Moeckel, R., et al. (2003), 'Microsimulation of land use', *International Journal of Urban Sciences,* 71 (1), 14-31.
Morand, E., et al. (2010), 'Demographic modelling: the state of the art', (FP7-244557 Projet SustainCity; Paris: INED), 39.
Müller, K. and K.W., Axhausen (2011), 'Population synthesis for microsimulation: State of the art', *90th Annual Meeting of the Transportation Research Board* (Washington D.C.).
Orcutt, G.H. (1957), 'A new type of socio economic system', *Review of Economics and Statistics,* 58, 773-97.
Parker, Dawn C., et al. (2003), 'Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review', *Annals of the Association of American Geographers,* 93 (2), 314-37.
Perrier-Cornet, Ph. (2001), 'La dynamique des espaces ruraux dans la société française : un cadre d'analyse', *Territoires 2020,* 3 (Etudes et prospectives - Data), 61-74.
Polhill, J.G., et al. (2008), 'Using the ODD Protocol For Describing Three Agent-Based Social Simulation Models of Land-Use Change', *Journal of Artificial Societies and Social Simulation,* 11 (2 3), 30.
Rindfuss, R., et al. (2004), 'Developing a science of land change: Challenges and methodological issues', *PNAS,* 101 (39), 13976-81.
Soumagne, J. (2003), 'Les services en milieu rural, enjeu d'aménagement territorial', *Revista da Faculdade de Letras - Geografia I série,* XIX.
Turci, L., et al. (2010), 'Provisional demographic outline', (FP7-244557 Projet SustainCity; Paris: INED), 24.
Ullman, E. and Dacey, M. (1960), 'The minimum requirement approach to the urban economic base', *Papers and Proceedings of the Regional Science Association,* 6 (192).

Verburg, P. H., et al. (2006), 'Downscaling of Land Use Change Scenarios to Assess the Dynamics of European Landscapes', *Agriculture, Ecosystems and Environment,* 114, 39-56.
Verburg, P.H., et al. (2004), 'Land Use Modelling: Current Practice and Research Priorities', *GeoJournal,* 61, 309-24.--- (2002), 'Modelling the Spatial Dynamics of Regional Land Use: The CLUE-S Model', *Environmental Management,* 30 (3), 391-405.

APPENDIX B

Commuting Network: Getting the Essentials

Abstract. Human mobility and, in particular, commuting patterns have a fundamental role in understanding socio-economic systems. Analysing and modelling the networks formed by commuters, for example, has become a crucial requirement in studying rural areas dynamics and to help decision-making. This paper presents a simple spatial interaction commuting model with only one parameter. The proposed algorithm considers each individual who wants to commute, starting from their residence to all the possible workplaces. The algorithm decides the location of the workplace following the classical rule inspired from the gravity law consisting of a compromise between the job offers and the distance to the job. The further away the job is, the more important the offer should be to be considered for the decision. Inversely, the quantity of offers is not important for the decision when these offers are close by. The presented model provides a simple, yet powerful approach to simulate realistic distributions of commuters for empirical studies with limited data availability. The paper also presents a comparative analysis of the structure of the commuting networks of the four European regions to which we apply our model. The model is calibrated and validated on these regions. The results from the analysis show that the model is very efficient in reproducing most of the statistical properties of the network given by the data sources.

Manuscript:
Gargiulo, F., Lenormand, M., Huet, S. and Baqueiro Espinosa, O. Commuting Network Model: Getting the Essentials. *Journal of Artificial Societies and Social Simulation* 2012, 15 (2) 6.

Commuting network models: getting the essentials

F. Gargiulo[1,2], M. Lenormand[1], S. Huet[1], O. Baqueiro Espinosa[3]
[1]LISC, Cemagref, BP 50085, 63172 Aubière, France
[2]CAMS-CNRS, 190 Av. de France, 75013, Paris, France
[3]IAMO, Theodor-Lieser-Str.2, D-06120 Halle (Saale), Germany

Abstract

Human mobility and, in particular, commuting patterns have a fundamental role in understanding socio-economic systems. Analysing and modelling the networks formed by commuters, for example, has become a crucial requirement in studying rural areas dynamics and to help decision-making. This paper presents a simple spatial interaction commuting model with only one parameter. The proposed algorithm considers each individual who wants to commute, starting from their residence to all the possible workplaces. The algorithm decides the location of the workplace following the classical rule inspired from the gravity law consisting of a compromise between the job offers and the distance to the job. The further away the job is, the more important the offer should be to be considered for the decision. Inversely, the quantity of offers is not important for the decision when these offers are close by. The presented model provides a simple, yet powerful approach to simulate realistic distributions of commuters for empirical studies with limited data availability. The paper also presents a comparative analysis of the structure of the commuting networks of the four European regions to which we apply our model. The model is calibrated and validated on these regions. The results from the analysis show that the model is very efficient in reproducing most of the statistical properties of the network given by the data sources.

For two decades, not only the number of commuters (i.e. people living in a municipality and working in another) but also, the average distance travelled by workers has increased in most European countries. This makes commuting a fundamental phenomenon in understanding socio-economic macrostructures. The precise description of commuting patterns has a central role in many applied questions: from the studies on traffic and the planning of infrastructures (Ortuzar 2001) to the diffusion of epidemics (Balcan 2009) or large demographic simulations (Huet 2011).

Despite their importance for describing realistic socio-economic frameworks, datasets describing human commuting patterns are rarely provided by statistical offices. Therefore a large effort has been made to find some algorithmic procedures able to reconstruct commuting flows, starting from the aggregate datasets that are usually available. These are models that simulate the morphogenesis of the network, taking into account the constraints given by the available aggregate data and the geographical properties of the networks. Good reviews of these methods can be found in (Ortuzar 2001), in the framework of transport modelling, in (Barthélémy 2011), in the framework of spatial networks modelling, and finally in (Rouwendal 2004) concerning micro-economy. On the other hand the field still has many gaps, mostly due to the difficulties met in calibrating the parameters of the proposed models and in finding good descriptions for zones inhabited by small populations. A discussion on the state of the art is provided in section 1.

Our research takes place in the framework of the European project PRIMA[1]. The microsimulation model developed within the PRIMA project simulates the dynamics of the population living in the European rural (low population density) municipalities. Therefore, one of our main focuses is the commuting structures in the rural areas of our case study regions. These structures had to be analysed and reproduced in the microsimulation model which aims to help decision-making regarding land-use policies. Thus, we needed a simple commuting network algorithm able to

[1] PRototypical policy Impacts on Multifunctional Activities in rural municipalities – EU 7th Framework Research Programme; 2008-2011; https://prima.cemagref.fr/the-project

generate the network of the European regions where the detailed commuting data was not available.

For some of these regions, the only available data at the municipality level consisted of total number of individuals commuting out of the municipality and total number of individuals commuting into the municipality. In these cases, the precise structure of the commuting network was unknown. In other words, the exact flows of individuals going from a municipality where they live to another one where they work was missing. Consequently, these flows had to be recreated on the basis of a set of assumptions. A description of the case studies we analysed is provided in section 2.

This paper describes the method we used to recreate all the commuting flows. Our method generates a commuting network, using a Monte Carlo simulation approach that can also be applied to low density zones. It is based on the individual choices of the commuters. We propose an extremely simplified framework, inspired by the gravity law, which aims to be general enough to be applicable to areas with diverse geographical features and different commuting structures. Despite its simplicity, the proposed approach is capable of faithfully replicating the structure of observed commuting networks.

Our algorithm considers each individual who wants to commute, from their living place to all possible workplaces. Individuals decide where they work following a classical rule consisting of a compromise between the job offers and the distances to the jobs. The further away the job is, the more important the offer should be to be considered in the decision. Inversely, the number of offers is less important for decision-making when these offers are in municipalities nearby. We initialize the algorithm with aggregate data on job seekers (i.e., the number of out-commuters) and job offers (i.e., the number of in-commuters) in each municipality. The algorithm memorizes past choices and after a job is associated to a commuter, the local information for the municipalities involved in the choice is updated. The algorithm is repeated until all the jobs are assigned. The details of the model are explained in section 3.1.

We also provide a method to calibrate the unique parameter of our algorithm, using detailed data from statistical offices. We show that, even if the selected regions are significantly diverse, the parameter does not vary dramatically from one region to another. The calibration method is presented in section 3.2.

Finally, we provide a quantitative framework to compare the network observed by statistical offices with the generated structures of our algorithm (Section 4). In particular, we articulate the validation systems at two levels. In section 4.2 we focus on the global topological properties of the network, such as the probability distributions of important network indicators (e.g., degrees and weights). In section 4.3, we introduce a statistical framework that allows a comparison, at the local level, of the similarity between the flows observed in the real case against those present in the generated network.

An implementation of the algorithm in netLogo, provided as additional material and detailed in the appendix, allows a graphical representation of the generation model.

1. Background

The literature on the construction and use of commuting networks is abundant; both from the point of view of the analysis of the structures, and from the point of view of the models (see the reviews of (Ortuzar 2001; Barthélémy 2011; Rouwendal 2004) in various research domains).

Many recent papers adopted an approach based on network theory. An interesting and complete analysis of the commuting structures from this point of view was introduced in (De Montis 2007; De Montis 2010). In this framework, most importantly concerning the modelling issues, the question about the commuting networks is set in the larger conceptual category of spatially constrained network structures. This kind of analysis concerns not only commuting, but all the situations where the geography has a significant role: from the reconstruction of migrant patterns (Lemercier 2008) to the analysis of the internet at autonomous system level (Pastor-Satorras 2004), to airline network structure (Barrat 2004). A particularly important study in this context is (Barrat 2005) where the concept of "preferential attachment" (Barabasi 1999) is adapted in order to consider not only the strength of a node given by its current in-degree, but also the spatial constraint included in the journey-to-work network.

A more classical approach comes from the micro-economists (Rouwendal 2004). Starting from the monocentric model of residential location proposed by (Alonson 1964), economists and geographers in urban modeling initially did not consider the space as determinant in residence location of the individual, assuming that places of work are all located in the center of a unique city. In the same way, looking at the decision regarding the job, job search theory does not take especially into account the distance of commuting in its first formalization. It assumes a worker's optimal strategy is simply to reject any wage offer lower than a reservation wage, and accept any wage offer higher than this reservation wage. However, commuting time was soon included in new job-search models as in (Van Den Berg 1997). In this model, a job offer consists of a wage and a commuting time pair. To be applied, this approach requires data on wage offers and their locations. When working with models at very local level (e.g., municipalities or villages), wage data is oftentimes difficult to obtain.

However, the most used approach to the modelling of commuting or migration structures is the one based on the so-called gravity law models (Haynes 1988). The term gravity law is a metaphor from classical physics. We can imagine that as it happens in gravitation, the interaction between two municipalities depends proportionally on a parameter: for example, the size of the municipality (equivalent to mass in the gravitational law), and in inverse proportion with some power law of the distance. It is recognized that the concept of "distance" can be formulated as something other than a real geographical or spatial category: it can be a travelling time, a topological distance on a network, but also a "social" distance (e.g. the cases of border cities where different languages are spoken). The classical formalization of probability p_{ij} of a commuter to live in the municipality *i* and to work in the municipality *j* is the following:

$$p_{ij} = \frac{f(M_i)g(N_j)h(d_{ij})}{\sum_{i,j} f(M_i)g(N_j)h(d_{ij})} \qquad \text{Eq. 1}$$

where we consider different proportionality parameters M_i, N_j, respectively for the origin and destination municipalities (this size could refer to the area of the municipalities, its population or the number of working people) and the distance between each pair of municipalities d_{ij}.

Using this probability model, it is possible to determine the traffic between each pair of municipalities with different methods (e.g. IPF, multinomial models, etc.). We notice that the

functions $f(M_i)$, $g(N_i)$ and $h(d_{ij})$ may assume any possible shape. For $h(d_{ij})$, the literature generally agrees that an exponential specification appears to fit better with reality. However, in some applications, a power law decay often seems to be a better fit (De Montis 2007; De Montis 2010; Reggiani 2007). Some studies propose a combined form of the two (Ortuzar 2001), or a different form (de Vries 2009), in order to better fit the empirical data.

The most common applied model of spatial interaction to generate commuting networks is the so called "doubly-constrained" model (Wilson 1998; Choukroun 1975). Based on the gravity law, it predicts the number T_{ij} of journeys-to-work between any pair of origin (*i)* - destination (*j*) zones considering the number of out-commuters of *i* and the number of in-commuters of *j*:

$$T_{ij} = A_i B_j R_i Q_j h(d_{ij}) \qquad \text{Eq. 2}$$

where:

$$A_i = \frac{1}{\sum_j B_j Q_j f(\beta, d_{ij})}$$

$$B_j = \frac{1}{\sum_i A_i R_i f(\beta, d_{ij})}$$

The factors A_i and B_j ensure that the T_{ij} table is consistent with the exogenous rows and columns totals: $\sum_j T_{ij} = R_i$ and $\sum_i T_{ij} = Q_j$. These balancing factors, plus a distance parameter β, implicit in the function $h(d_{ij})$, have to be calibrated. An entropy maximization approach allows calibrating such model considering only one parameter to find (β) since A_i and B_j are automatically solved by this method. This optimization approach consists in associating any particular microstate with a macrostate, which is simply the number of trips from an origin to a destination. A macrostate is feasible if it reproduces known properties referred to as system states (for example, the total number of travelers). Estimating the solution of the model consists in finding the macrostates, maximizing a chosen distance function of the considered macrostate to the observed data among the feasible macrostates (Bernstein 2003).

Several improvements were proposed based on this doubly-constrained model. In (Fotheringham 1981), a competing destination model is introduced to improve the spatial structure of the generated network. (Fik 1990) extend this competing model to measure the accessibility of a destination related to destinations of the same hierarchical order in the system of central places (founded on the Central Place Theory). They also incorporate a measure that relates to the number of intervening opportunities from the living place *i* to the attractive force *j*. These intervening opportunities are the potential destinations within a distance smaller than d_{ij}. To go beyond the gravity law models' weaknesses, some authors developed an approach founded on the network paradigm (Thorsen 1999; Gitlesen 2010). This kind of procedure has the disadvantage of increasing the number of parameters, which is what we wanted to avoid.

Very recently, (Simini 2011) proposed an algorithm free of parameters to generate many different spatial networks. They consider the job demand and the job offer as a part of the population of the origin-destination zones, and compute the probability of a flow between the origin *i* and the destination *j*, considering these parts and the density of people living between *i* and *j*. They apply this principle for the generation of the commuting network of USA at the county level. This model is very interesting, nonetheless we doubt its suitability to reproduce a commuting network at such a low level as the municipalities in our study regions (such as France, where the average size of an

Auvergne municipality is 1024 inhabitants). Though this model addresses similar issues, the authors conclude the lack of an *effective distance* weakens their model fitness.

Our study analyses different regions from various countries. Regions are defined as sets of NUTS3[2] areas for each country; these regions vary in size, population, and other economic and social properties. We are interested in the inter-municipality commuting network. Very few papers deal with this topic on a small scale. Some studies analyse the inter-municipality commuting network (De Montis 2007; De Montis 2010), showing that the Sardinian and the Sicilian inter-municipal commuting networks exhibit a traffic property based on a power law with exponent 2. Others, such as (Thorsen 1998), compare different spatial interaction models by an empirical evaluation of the municipalities of a Norwegian region. One study analyses at the district level (which is higher than the municipality level) the German commuting network (Patuelli 2007), using a comparison of two spatial interaction models. The set of publications shows the interest in such an approach to study the evolution of these types of networks over time.

For the presented model, the individual choice for a job location is probabilistic. Decisions are mainly stochastic, and so is the model. Each time the model is run, we obtain a different network based on the statistical properties used as input. This should be contrasted with the generation of an optimized network making deterministic the flow between the related municipalities. Especially for the latter, a deterministic approach does not appear relevant, since the local commuting choice is influenced by many local decisions which can be seen as random variations. The validation of the model shows that we obtained a good fit of the network given by the observed data. These results are very stable; the stochasticity of the model thus reflects local diversity without perturbing the statistical properties of the network. For the deterrence function, we decided to use a power law; nevertheless, another function could be tested.

2. Regional commuting network structures – Specific differences and global properties.

The first part of our study concerns the analysis of our study regions. The local statistical offices[3] provide all the information necessary to characterize the structure of the commuting network. We consider: two separated NUTS2 regions in France (Auvergne and Bretagne), each composed of four NUTS3 regions; a group of two NUTS3 regions in the UK (Nottinghamshire and Derbyshire); and a group of two NUTS3 regions in Germany (the Altmark region is composed by the districts of Stendal and Salzwedel). Differences between the data availability within regions must be noted, as the data for Auvergne and Bretagne is much more comprehensive, in comparison to the other two case study regions. Indeed, data describing the commuting flows between each pair of municipalities with less than 10 commuters is not available in the German data (Altmark) and the simlar flows smaller than 3 are not available in the English data (Nottinghamshire and Derbyshire).

The selected regions differ on many aspects: the number of municipalities, geographical structure, and socio-economic characteristics. They were chosen by the European project because they are

[2] The Nomenclature of Territorial Units for Statistics. NUTS 2 corresponds to European basic regions for the application of regional policies and NUTS 3 to small regions for specific diagnoses. For more details, see http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

[3] In France: thanks to the Maurice Halbwach Center, which made available the complete French origin-destination tables for commuters in 1999. In Germany: Commuting data was purchased from the German Federal Employment Agency (Bundesagentur für Arbeit) for the year 2000. In the United Kingdom: Origin-destination data was obtained via the Office for National Statistics NOMIS online database (https://www.nomisweb.co.uk/) for the year 2001.

all rural regions with diverse socio-economic characteristics. Table 1. presents some basic characteristics of each case study.

Table 1. Characteristics of selected study regions

Region	Number of municipalities	Average size of a municipality (by number of inhabitants)	Average inter-municipality distance (in km)	Number of commuters living and working in the region	Part of commuters living in and working outside the region	Total area surface (in km^2)
Auvergne (France)	1310	1024	88	261822	7.73%	26,013
Bretagne (France)	1269	2447	99	608587	7.32%	27,208
Altmark (Germany) – subregions	91	2527	50	16770	66.82%	4,715
Nottinghamshire /Derbyshire (UK)	372	5300	44	573022	12.4%	4,839

The objective of this first analysis is to determine the characteristics of the commuting networks composed by the regional commuting flows that are present in each region. For this analysis, we create a commuting matrix from a dataset containing the number of individuals that commute (i.e. reside in one settlement and work in another) within each of the selected regions. A representative section of the matrix used is shown in Table 2. Each row represents the place of residence and each column represents the working place; the cell at the intersection of each row and column contains the number of persons living and working in the corresponding row and column. For our analysis we ignore the cells in the diagonal of the table, as they represent non-commuting individuals (i.e., persons living and working in the same place).

Table 2. Example of commuting data from the Altmark Region

		Municipality of employment						
		81026	81030	81035	81045	81080	81095	…
Municipality of residence	81026	0	0	0	0	0	0	…
	81030	0	0	0	3	0	0	…
	81035	0	0	0	2	0	0	…
	81045	0	2	2	0	2	2	…
	81080	0	0	0	0	0	0	…
	81095	0	2	0	8	0	0	…
	…	…	…	…	…	…	…	0

After analyzing some global properties of the network structure we observe that the presented regions have quite dissimilar structures. The first analyzed property of the networks concerns the distributions of the degrees. The degree is a property of the associated un-weighted network. For the construction of the un-weighted network we consider all the municipalities and add a directed link between the municipality *i* and the municipality *j* if at least one individual commutes from *i* to *j*. The in-degree of a municipality *i* ($k_{in}(i)$) is the number of links entering in *i*, while the out-degree ($k_{out}(i)$) is the number of links starting from *i*.

The probability distributions of the “in and out” degrees are represented in figures 1. As we can observe on these figures, the different case studies are characterized by very different behaviours according to the degree distribution.

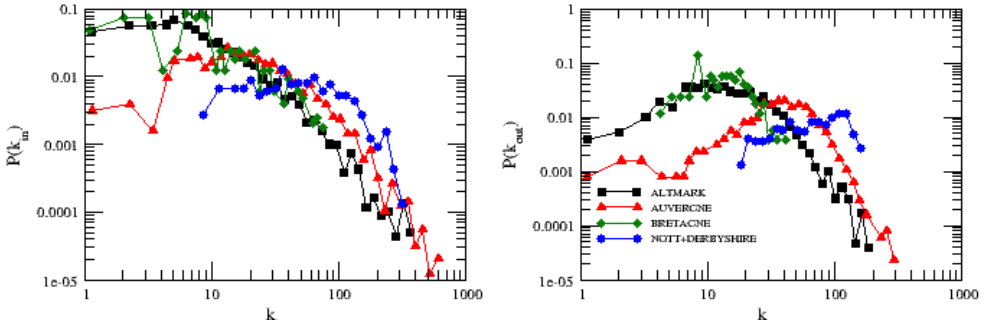


Figure 1. In and Out degree distributions of each case study region

The out-degree distribution shows that the municipalities in the UK region always have a large and uniform degree distribution. This can be explained by the fact that for the UK, the number of commuters is extremely large, and the network is very dense in terms of links. This kind of uniform structure can be connected to the lack of "working hubs" able to attract workers more strongly than the other municipalities. This corresponds with what we observe in the in-degree distribution where we see that few municipalities have a small in-degree while a considerable part has a high in-degree.

The situation in Auvergne and Bretagne, where the in-degree distributions suggest the presence of real "working hubs" in the commuting network (a small but not unimportant part of municipalities reached much more than the others) is totally different.

For the Altmark region, the total number of connections is generally lower, suggesting that this region represents only a part of a larger commuting network. This can be explained considering that the majority of commuters in the studied region work outside this region (66.82% as shown in Table 1).

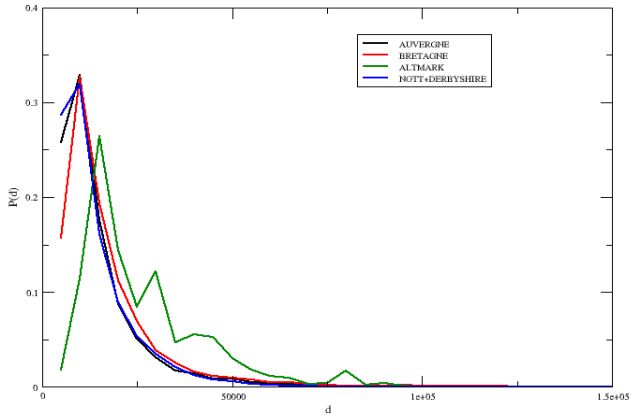


Figure 2. Distribution of the commuting distances (in meters) for the selected case studies

Another important consideration concerns the distribution of distances covered by the commuters. This measure is presented in Figure 2.

The distribution of the distances shows that in the UK regions, smaller distances are favoured. This confirms our intuition that job offers are homogeneously distributed among all the municipalities in the region (thus, there are no working hubs). For this reason people do not need to travel long

distances to find a job. The opposite situation is observed in the Altmark case, where a significant share of the commuters can travel up to 80 km.

In this section we provided a brief description of the selected case studies. We showed the structural differences and global properties of the studied commuting networks. In the following section we present a method to construct a synthetic network, based on the decision of the individual workers. This method is then used to generate commuting networks for regions where the detailed commuting data is not available.

3. A Monte-Carlo simulation approach to generate realistic commuting networks

The usual methods for reconstructing the structure of commuting networks are based on the gravity law. The main hindrance of this approach is that it is not easy to calibrate the gravity law model (Williams 1976). Moreover, it is a deterministic method which appears inappropriate when flows for small municipalities must be predicted, as it is the case for our study regions. We propose a simple network generation model that presents a higher level of universality and which can be applied with a good degree of confidence to all the case study regions.

3.1 The individual-level generation model

The model is based on the individual choices of the commuters, namely people in the active class that do not work in the municipality where they reside.

When looking for an occupation outside of the living place, two factors can influence the choice of the destination: the distance of the potential workplace and its "attractiveness" (defined by the number of jobs it offers). The further away the possible destination is, the more its attractiveness will matter in the decision. If the possible destination is near, the settlement attractiveness becomes less significant for the individual's decision for a workplace.

We start from a typology of data that is usually available, for each municipality, in each case study:

. the total number of out-commuters (R_i), also called the job demand of the municipality *i,*

. the total number of in-commuters (Q_i), also called the job offer (or attractiveness) of the municipality *j,*

. the distances among each couple of municipalities (d_{ij})

In the presented study we use the Euclidean distances in km to describe the distances. Similar results can be obtained using, for example, the road distance or travelled time measures. Some performed test on the results showed that the algorithm is robust to the choice of other distance definitions.

To each commuter residing in each municipality *i*, the algorithm associates a working destination *j* according to the job offers of all the municipalities different from *i* in the region and the distance between the municipality *i* and all the possible destinations. The algorithm for the generation of the network evolves according to the following steps:

For each remaining commuter who has not already found a place to work, we:

- Select a residence municipality *i* at random among the municipalities where there is at least one out-commuter ($R_i>0$)
- Select the working destination *j* randomly following the probability distribution given by:

$$p_{i\to j} = \frac{Q_j d_{ij}^{-\beta}}{\sum_{j\neq i} Q_j d_{ij}^{-\beta}} \qquad \text{Eq.3}$$

- Update the number of out-commuters of i and the number of in-commuters of j: $R_i=R_i-1$, $Q_j=Q_j-1$
- Recalculate the $p_{i\to j}$ distribution

The relation between the offer and the distance is characterized in the model with the parameter β which captures the relative impact of the distance. Using this algorithm we ensure that the generated network respects exactly the incoming and outgoing traffic from each node.

Different values of the parameter β produce different distance and degree distributions for the generated networks. We calibrate the parameter for the case studies where the complete information in the network is known, in order to have the same distance distribution as the one observed for the real network.

Analysing the calibration on the regions where the data is available, we observe that with an appropriate choice of the parameter β we are able to generate a commuting network with statistical properties which are very similar to the real network. The calibration procedure and the analysis of the accuracy of the generation algorithm are presented in the following sections.

3.2 Model calibration

The proposed model depends on the spatial parameter β which represents the relative importance of the distance to the destination when choosing a working place. A typical property that distinguishes commuting networks is the distribution of the travelled distance for each worker. We employ this information to calibrate the parameter β. In fact, each value of β produces a network with a typical distance distribution, as it is displayed in Figure 3 for the Auvergne case study.

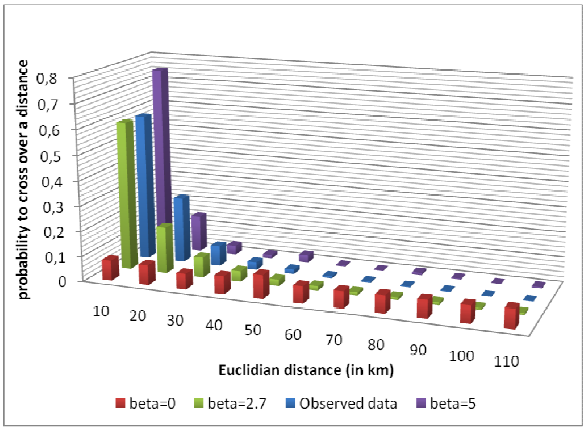


Figure 3. Distance (d in KM) distribution for the real network and three different β values for the Auvergne case study

We observe that, for excessively low values of β, the preference toward distant working places is overestimated, while for excessively high values, the choice of close places is overestimated. We calibrate β in order to minimize the distance between the generated travelled distance distribution and the one obtained from the observed data. The minimized distance is the Kolmogorov-Smirnov distance:

$$D_{KS} = \sup_d \left| P^c{}_o(d) - P^c{}_g(d) \right| \qquad \text{Eq. 4}$$

where $P^c{}_{o/g}(d)$ are the cumulative distance distributions for the observed (*o*) and generated (*g*) networks.

For each case study we calculated this distance for different values of β and chose the minimum of the function $\langle D_{KS} \rangle(\beta)$ as the calibrated parameter value. Indeed, to choose the parameter value, we considered $\langle D_{KS} \rangle$ since the model is stochastic. The value of $\langle D_{KS} \rangle$ is obtained by calculating the average of the D_{KS}, measured on 100 replications of the generated network for each β value. Within these replications, the variation of the measured D_{KS} is very low, at most 1.13% of $\langle D_{KS} \rangle$. The calibration process is described by the Figure 4. Each dot corresponds to a tested value of β (with a step of 0.5 from 0 to 7 for β, and with a step of 0.1 from 2 to 3 for the $\langle D_{KS} \rangle$ values required to identify the minimum). Figure 4 shows that for the analyzed regions, the value of β lays in the range [2, 3].

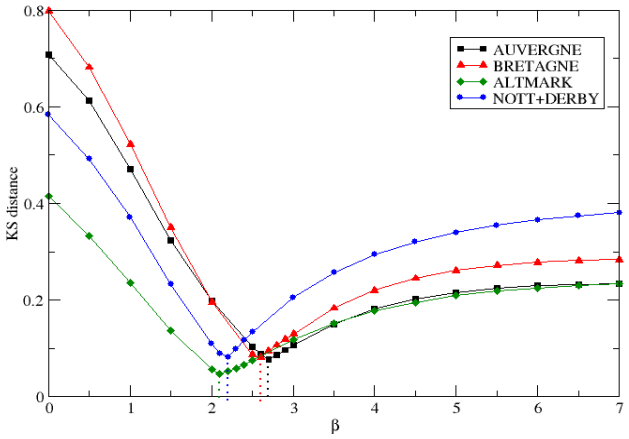


Figure 4. Calibration process results for the four case study regions based on the minimization of the average Kolmogorov-Smirnov distance over 100 replications (each dots represents the result for a tested β value.)

Table 3 lists the optimal values for all the studied regions, where the $\langle D_{KS} \rangle$ distance is minimized.

Table 3. Optimal values of β for the studied regions

Region	β
Auvergne	2.71
Bretagne	2.59
Altmark	2.1
Nottinghamshire and Derbyshire	2.2

In the analyzed regions, notwithstanding the relevant geographic and demographic differences, the coefficient varies slightly in the interval $\beta \in [2,3]$.

Moreover, we can observe that for all the regions in the whole considered interval, the average KS distance $\langle D_{KS} \rangle$ between the observed distribution and generated ones is always small. This suggests a strategy for applying this algorithm to the cases where the calibration datasets are not available. A stochastic procedure where at each replication the β value is randomly extracted in the interval $\beta \in [2,3]$ can reproduce, with a good approximation, the commuting patterns of the region. This last assumption is valid only if the considered region is sufficiently isolated; that is, if the total number of commuters, in and out from a municipality, commute to other municipalities within the same region.

4. Validation

To assess the quality of the generated network, we compare its properties to the properties of the observed network (i.e., data obtained from the regions' corresponding National Statistical Office).

Two different kinds of properties are investigated: a first group is measured on the municipality network where we consider that two municipalities are linked when at least one worker commutes between them, whatever the origin-destination is (i.e., considering an unweighted network); a second one is measured on the weighted network which has direct links weighted by the number of individuals commuting from a given municipality to another one.

For the unweighted network, two different indicators are considered:

1. The ability of the generated data to fit the observed in and out degree distributions of the "municipality" network;
2. The traffic density distribution describing the density of each weight that can be associated to an undirected link. For an arc between two municipalities, this weight is the sum of the individuals going from one municipality to the other in both directions through the arc.

For the weighted network, we compare the number of commuters of both the generated and the observed network.

All these statistics were not used to generate simulated networks. Moreover, we must remember that the number of people looking for a job in a municipality i (R_i) and the job offers in a municipality j (Q_j), are reproduced precisely in all municipalities by the generation algorithm.

4.1 The properties of the municipality network (i.e. the unweighted network)

We consider three variables to describe the topological properties of the network and the characteristic of the commuting flows: the in and out degree distribution ($p(k_{in})$ and $p(k_{out})$) and the traffic distribution (p(T)). These indicators are influenced by the choice of the parameter β. As we can observe in Figure 5 for the Auvergne case study, for β=0 (i.e., when the geography is not important), higher network degrees and lower traffics are observed. As the geography becomes more important (i.e., as β is increased) the maximum network degree decreases and the maximum amount of traffic increases. When distance is not important, people choose their working destination in a wider range of available municipalities. On the contrary, a strong distance constraint forces to choose only between the nearby municipalities. As a consequence of this, traffic on this smaller number of connections will also be globally higher.

For the Auvergne case study, Figure 5 shows the comparison of the generated and the observed data. It can be seen that, the distributions at the calibration point (β=2.7) fit the distributions of the observation network perfectly. This fitness should be observed, considering that none of the three measurements (in-commuting degree, out-commuting degree and distribution of traffic), are used by the model; thus, the fitness of the generated network to the observed one is a positive assessment of the effectiveness of the model.

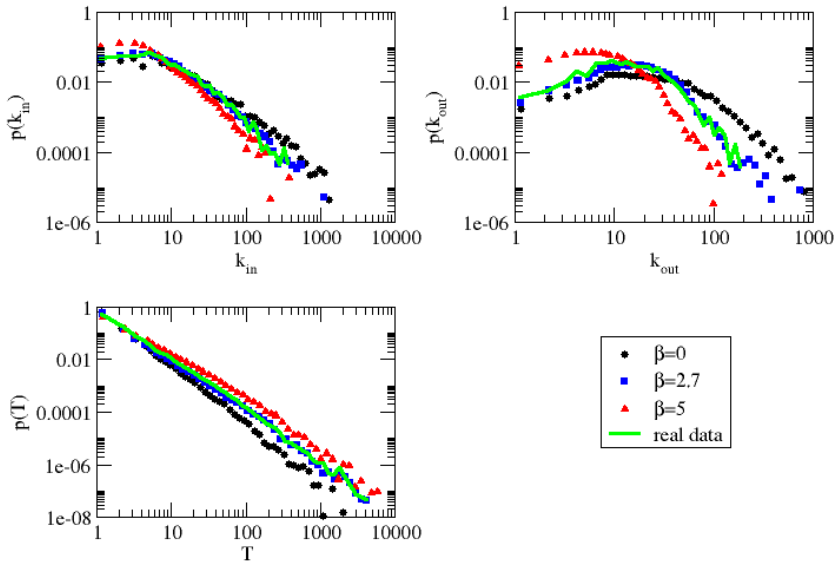


Figure 5. In (k_{in}) and out (k_{out}) degree distributions and traffic (T) distribution for some generated networks with various values of β and for the observed network for the Auvergne case study. The results for the generated networks are averaged on 100 replications of the generation algorithm.

Figures 6 shows the comparison between these measures for the observed network and the generated ones for the other case studies. As we can notice in these figures, the traffic (T) distribution is well reproduced in all the case studies. It is not the case for the degree distributions

in the UK case study where the generation process completely fails in the estimation. We attribute this discrepancy to the quality of the Census data. Indeed, in UK, a small-cell adjustment method (Stillwell and Duke-Williams 2007) is applied to prevent disclosure of personally identifying data. In particular, this method suppresses some commuting data by replacing values of 1 and 2 with 0 or 3. This adjustment makes the definition of a link between two municipalities different in the model beyond the data and in the generated network through our algorithm. According to the census data, two municipalities are linked only if at least three individuals commute among them. In our generated network, they are linked if at least one individual commute among them. A large number of municipality pairs are, in reality, linked by only one or two individuals. Such pairs are underestimated in the real UK data. We believe this is the reason why the model seems to overestimate the connectivity between municipalities.

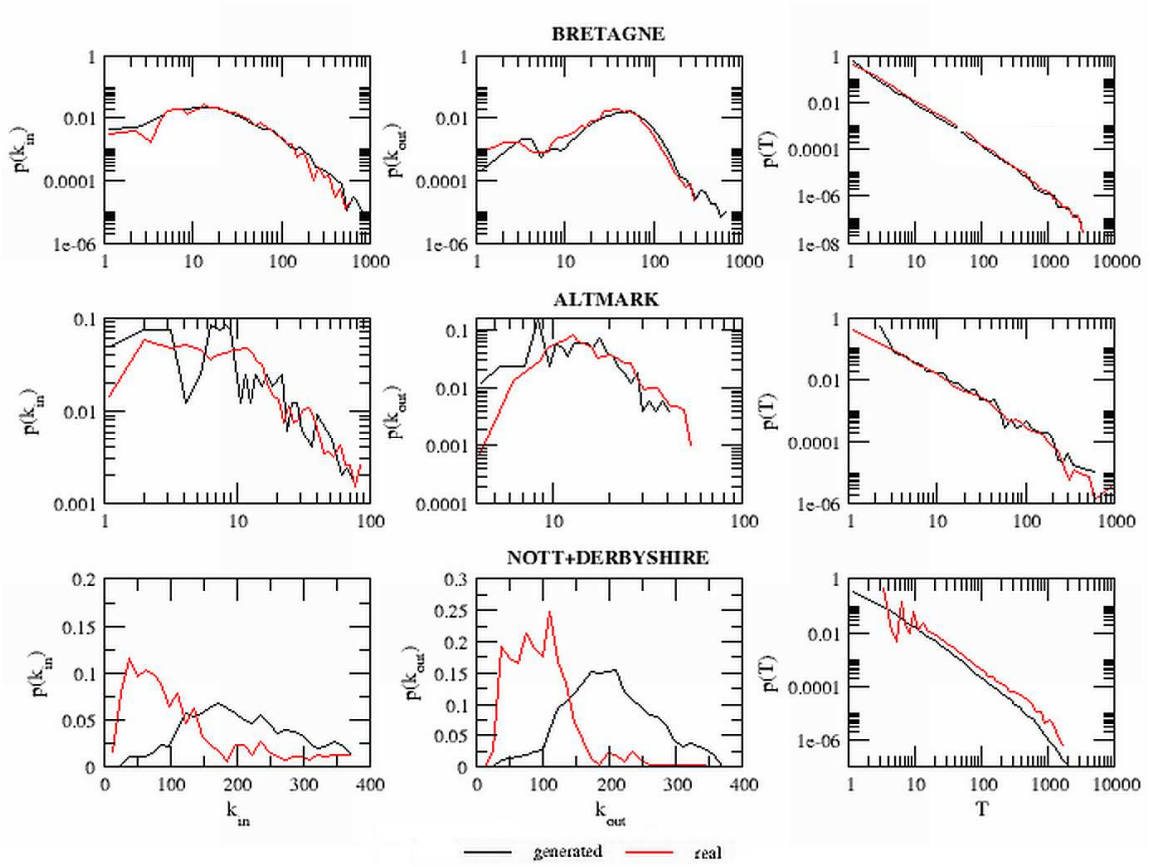


Figure 6. In (k_{in}) and out (k_{out}) degree distributions and traffic (T) distribution for the generated networks at the calibration point and for the real network for the Bretagne, Altmark and UK case studies. The results for the generated networks are averaged on 100 replications of the model.

4.2 The common part of commuters of the weighted network

We now define an indicator to compare the generated commuting network and the observed commuting network. The statistical offices of France, Germany and United Kingdom provided the observed commuting networks. Assuming that $M_n(N)$ is the set of all possible networks for a set of municipalities. Let $O \in M_n(N)$ be one commuting network when O_{ij} is the number of commuters from municipality *i* to municipality *j*. Let $G \in M_n(N)$ be another commuting network between the

same set of municipalities where G_{ij} is the number of commuters from municipality *i* to municipality *j*.

To assess the similarity of flows between the generated and the observed networks, we can compute the common part of commuters (CPC) (Eq. 7) from the number of common commuters (NCC) between *O* and *G* (Eq. 5) and the number of commuters (NC) in *O* (Eq. 6). The CPC appears to be a good indicator of the prediction quality. This indicator may be seen as a simplified variant of the Sørensen index, with the two compared matrices having the same size. The CPC was chosen for its intuitive explanatory power: it is a similarity coefficient which gives the likeness degree between two networks. Its value ranges from 0, when there are no commuters flows in common in the two networks, to a value of 1, when all commuters flows are exactly identical in the two networks.

$$NCC_n(G,O)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\min(G_{ij},O_{ij})\right) \quad \text{Eq. 5}$$

$$NC_n(O)=\sum_{i=1}^{n}\sum_{j=1}^{n}O_{ij} \quad \text{Eq. 6}$$

$$CPC=\frac{NCC_n(G,O)}{NC_n(O)} \quad \text{Eq. 7}$$

This gives us an indicator to directly compare one replication of the generated network with the observed one. We do the same with all the 100 replications for a given β value and compute the average of the obtained 100 CPC to evaluate the quality of the model. Within the 100 replications, the CPC varies at most, by 1.76% of the average; this means that the stochastic model is very stable (i.e., the stochasticity does not have a significant effect on the properties of the network).

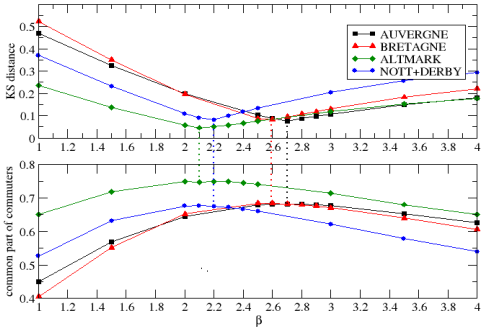


Figure 7. Common part of commuters (at the bottom) for different β values for each case study region (compared to the calibration graph of the Figure 4, presented on the top)

Figure 7 presents the average CPC for each region and for different β values. It is noticeable that the best value of the average CPC function is very close to the one given by the calibration value of β for all the studied regions. This point is stressed out in Figure 7 by the dotted line showing the match between the average CPC value and the minimum of the D_{KS}. The proximity, in terms of β of the minimum of the D_{KS} function with the maximum of the CPC function, is surprising and reinforces the idea the CPC is a good quality indicator. We also notice that the best values for both, the D_{KS} and the common part of commuters, varies when defining the model parameter between β = 2 and β = 3. Also, the results from the CPC indicator reinforce the suggested method

for the generation of a network in the case where the data is not directly available. In fact, for any point in the interval $\beta \in [2,3]$, and for all the considered regions, the CPC value never goes below CPC=0.6, showing that the generation process yields networks that match the observed network with good accuracy.

Table 4 shows the average CPC for each case study region and the optimal β value. Results are encouraging: average CPC values fall between 0.67 and 0.76. On average we obtained about 70% of commuters in common. It means that 70% of the observed network is returned by the model. One may also notice that the optimal β value seems to vary in the same way as the average inter-municipality distances of the region (see table 1), for which the Germany and the UK regions both have a small value, whereas the Auvergne and the Bretagne regions, both show a large value.

Table 4. Average Common Part of Commuters for the four case study regions

Region	β	Average Common Part of Commuters
Auvergne	2.71	0.683
Bretagne	2.59	0.684
Altmark	2.1	0.751
Nottinghamshire and Derbyshire	2.2	0.676

5. Discussion and conclusions

We propose a very simple stochastic individual-based model able to generate a commuting network with good accuracy. This model is based on the doubly-constrained model proposed by Wilson (1998) and has its roots on the so-called gravitational laws (i.e., consider that individuals tend to "gravitate" towards more attractive areas). It is built on the same principles: an individual tends to choose a job location depending on the job offers and the distance to the offer. The effect of the distance decreases as the distance increases, following a function that we have chosen as a power law. Our model has only one parameter which can be easily calibrated. It ensures that the number of out-commuters and in-commuters for each municipality is respected without needing to solve an optimization problem. However, it must be stressed that our proposed model does not try to reconstruct the exact structure of a commuting network. Achieving this would require considering additional local properties, which are very specific for each region. Instead, we aimed to create a model that can generate *realistic* synthetic networks from a limited set of data (number of in-commuters and out-commuters on each municipality), which can be used in cases where the detailed commuting data is unavailable. Moreover, reproducing exactly a network at a very low level, especially for very small municipalities (e.g., around 1000 inhabitants on average in some French regions or less than 200 for the studied German region) makes no sense since very small commuting links between small municipalities can result from stochastic factors that cannot be captured with a real deterministic law. As our algorithm is stochastic, it obtains many possible combinations of generated networks respecting the total local commuting flows. This approach seems more relevant than a deterministic approach for modelling a commuting network at the municipality level.

Our algorithm is validated on four case-study regions situated in France, Germany and the United-Kingdom. We compare the properties of the observed network given by the complete origin-destination table to those of the generated networks. We conclude that the in and out degree distributions of the municipality network, the traffic distribution of the same network are well fitted by the generated networks' distributions. Moreover, the common part of commuters of a

generated network with the observed network (i.e., the complete origin-destination table) appears high for all the case study regions. Incidentally, we have noticed that the optimal parameter value of our algorithm is very close to the parameter value that yields a higher value of common commuters.

The proposed model appears quite relevant for our main problem. Nevertheless, we must remember that aggregated statistics available at the municipality level correspond to all the in-commuters and all the out-commuters of each municipality. This includes commuters that live or work outside the region (i.e., in other municipalities not included in the network). To be sure that our model produces a representative network, it has to be applied on a region where these commuters linked to the outside represent an insignificant part of the total number of commuters. In other words, the region should be what Paelink and Nijkamp (1975) called a "polarized region": "*a connex area in which the internal economic relationships are more intensive than the relationships with respect to regions outside the area*" (Cörvers 2009; Konjar 2010) .

In spite of this limitation, it is apparent from the results of the analysis of the Altmark network (a region where 66.82% of the workers commute outside the region) that the similarity of the generated and real network is good (as shown in the analysis of Figure 7). However, we have to keep in mind that the data regarding the commuting flows smaller than 10 are not available for the Altmark region, and currently we do not know how this limitation impacts on the results. Two issues have affect on the proposed method when the used data includes individuals residing or working outside the region. On the one hand, the model will tend to overestimate the traffic within municipalities, as residents who ought to work outside are distributed within network municipalities. On the other hand, the number of connections may be underestimated as residents occupy jobs which should be taken by individuals living outside the region (thus, leaving municipality with low attractivity without in-commuters).

Such limitations may be addressed with the use of additional data detailing the number of individuals commuting from or to places outside the region. Alternatively, it is possible to conclude through aggregated data at the regional level or expertise, whether a region is sufficiently independent from another regarding the labour market.

The second issue concerns the model calibration. Most of the known power-law networks have an exponent value situated between 2 and 3. Our first case studies seem to show that the exponent of our power-law deterrence function varies in the same range. We notice that the error remains quite low between these two boundaries for β.

A further possible analysis involves testing the quality of an algorithm free of parameters proposed by (Simini 2011), even if it does not take directly into account the number of commuters. Such algorithm should be tested on sparsely populated regions such as the ones we worked on (i.e., at the municipality level). They apply this principle for the generation of the commuting network of USA at the county level. This model is very interesting; albeit we question its quality to reproduce a commuting network for very local regions the ones we studied.

Finally, the model could be improved by the use of other types of distances (such as the commuting time between municipalities). Although our results show that even using a measure such as the Euclidean physical distance (in the case of French regions, or a driving distance (in the case of the Germany region), the model generates networks with similar properties of those observed by the real data. Such refinement is usually limited by the lack of distance data (in this

case, commuting time) for the regions. Furthermore, it may be possible to select a better value of β if additional case study regions with geographical and socio-economic differences are analysed.

Acknowledgments

The work of F.G. is founded by the French ANR project SIMPA.

References

ALONSO, W. (1964), Location and Land Use: Toward a General Theory of Land Rent. Cambridge: Harvard University, 204 pages.

BALCAN, D., et al. (2009), 'Multiscale mobility networks and the large scale spreading of infectious diseases', PNAS, 106 (51), 21485-89.

BARABASI, A.L. and Albert, R. (1999), 'Emergence of Scaling in Random Networks', Science, 286 (5439), 509-12

BARRAT, A., Barthélemy, M., and Vespignani, A. (2005), 'The effects of spatial constraints on the evolution of weighted complex network', Journal of Statistical Mechanics: Theory and Experiment, P05003, 11.

BARRAT, A., et al. (2004), 'The architecture of complex weighted networks', Proc. Natl. Acad. Sci. U. S. A. , 101, 37-47.

BARTHÉLÉMY, M. (2011), 'Spatial Networks'. Physics Reports 499:1-101, arXiv 1010.0302v2.

BERNSTEIN, D. (2003), 'Transportation Planning', in W.F. Chen and R. J.Y. Liew (eds.), The Civil Engineering Handbook (Boca Raton, London, New York, Washington D.C.: CRC Press LLC).

CHOUKROUN, J.M. (1975), 'A General Framework for the Development of Gravity-type Trip Distribution Models', Regional Science and Urban Economics, 5, 177-202.

CÖRVERS, F., Hensen, M., and Bongaerts, D. (2009), 'Delimitation and coherence of functional and administrative regions', Regional Studies 43 (1), 19-31.

DE MONTIS, A., Barthélémy M., Chessa A., Vespignani A. (2007), 'The Structure of Inter-Urban Traffic: A weighted network analysis', Environment and Planning B: Planning and Design., 34 (5), 905-24.

DE MONTIS, A., Chessa A., Campagna M., Caschili S., Deplano G. (2010), 'Modeling commuting systems through a complex network analysis. A study of the Italian island of Sardinia and Sicily', Journal of Transport and Land Use, 2 (3/4), 30-55.

DE VRIES, J. J., Nijkamp, P., and Rietveld, P. (2009), 'Exponential or Power Distance-decay for Commuting? An Alternative Specification', Environment and Planning A, 41 (2), 461-80.

FIK, T.J. and Mulligan, G.F. (1990), 'Spatial flows and competing central places: towards a general theory of hierarchical interaction', Environment and Planning A, 22 (4), 527-49.

FOTHERINGHAM, A. S. (1981), 'Spatial Structure and Distance-Decay Parameters', Annals of the Association of American Geographers, 71 (3), 425-36.

GITLESEN, J.P., et al. (2010), 'An Empirically Based Implementation and Evaluation of a Hierarchical Model for Commuting Flows', Geographical Analysis, 42, 267-87.

HAYNES, K. and Fotheringham, A. (1988), Gravity and spatial interaction models (Sage Beverly Hills).

HUET, S. and Deffuant, G. (2011), 'Common Framework for Micro-Simulation Model in PRIMA Project', (Cemagref Lisc), 12.

KONJAR, M., Lisec, A., and Drobne, S. (2010), 'Method for delineation of functional regions using data on commuters', 13th AGILE International Conference on Geographic Information Science (Guimarães, Portugal), 10.

LEMERCIER, C. and Rosental, P.A. (2008), 'Les migrations dans le Nord de la France au XIXème siècle. Dynamique des structures spatiales et mouvements individuels', Nouvelles approches, nouvelles techniques en analyse des réseaux sociaux, 19.

ORTUZAR, J.D. and Willusem, L.G. (2001), Modelling Transport (3rd (in 2011) edn.; Chichester: John Wiley and Sons Ltd) 439.

PASTOR-SATORRAS, R. and Vespignani, A. (2004), Evolution and structure of the Internet: A statistical physics approach (Cambridge University Press) 270.

PATUELLI, R., et al. (2007), 'Network Analysis of Commuting Flows: A Comparative Static Approach to German Data', Network Spatial Economy, 7, 315-31.

REGGIANI, A. and Vinciguerra, S. (2007), 'Network Connectivity Models: An Overview and Empirical Applications', in T.L. Friesz (ed.), Network science, nonlinear science and infrastructure systems (New York: Springer), 147-61.

ROUWENDAL, J. and Nijkamp P. (2004), 'Living in Two Worlds: A review of Home-to-Work Decisions'. Growth and Change, vol. 35, 3, 287-303.

SIMINI, F., Gonzalez M.C., Martian A., Barabasi A.L., 2011. 'A universal model for mobility and migration patterns'. Arxiv 1111.0586

STILLWELL, J. and Duke-Williams, O. (2007), 'Understanding the 2001 UK census migration and commuting data: the effect of small adjustment and problems of comparison with 1991', J. R. Statist. Soc. A, 170 (Part 2), 425-45.

THORSEN, I. and Gitlesen, J.P. (1998), 'Empirical evaluation of alternative model specifications to predict commuting flows', Journal of Regional Science, 38 (2), 273-92.

THORSEN, I., Uboe, J., and Naevdal, G. (1999), 'A network approach to commuting', Journal of Regional Science, 39 (1), 73-101.

VAN DEN BERG, G.J. and Gorter C. (1997), 'Job Search and Commuting Time'. Journal of Business & Economic Statistics, vol. 15, 2, Structural Estimation in Applied Microeconomics, April, 268-281.

WILENSKY, U. (1999), NetLogo. https://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer Based Modeling, Northwestern University, Evanston, IL.

WILLIAMS, I. (1976), 'A comparison of some calibration techniques for doubly constrained models with an exponential cost function', Transportation Research, 10, 91-104.

WILSON, A.G. (1998), 'Land-use/Transport Interaction Models - Past and Future', Journal of Transport Economics and Policy, 32 (1), 3-26.

Appendix: Implementation of the model in the NetLogo framework

An example implementation of the model is included to illustrate how the model works. The implementation was performed in NetLogo 5.0RC4[4] and may run in previous versions (it was successfully tested in version 4). The implementation provides a way to visualize the generation of a network from two input files containing the in-commuting and out-commuting information for each municipality in a region and the distances between each pair of municipalities.

As mentioned, the model requires two input files to run:

1. The commuters file named *commuters.csv*: Which should contain a list of municipalities (one for each line in the file) and the number of individual who commute-out and commute-in (in that order) for each municipality. Each column must be separated by one blank space.
2. The distances file named *distances.csv*: Which should contain the distance between each pair of municipalities as a three column row containing the *origin* municipality, the *destination* municipality, and the distance between the pair (in that order). Each column should also be separated by a blank space.

The interface of the implementation is shown in Figure 8. Prior to starting a simulation the *beta* parameter must be set in order to define the weight of the distance in the commuting decision. For illustrative purposes the *proportional-sizes* control is provided to present each municipality (depicted as a house in the interface) with a size relative to the initial number of in-commuters (job availability).

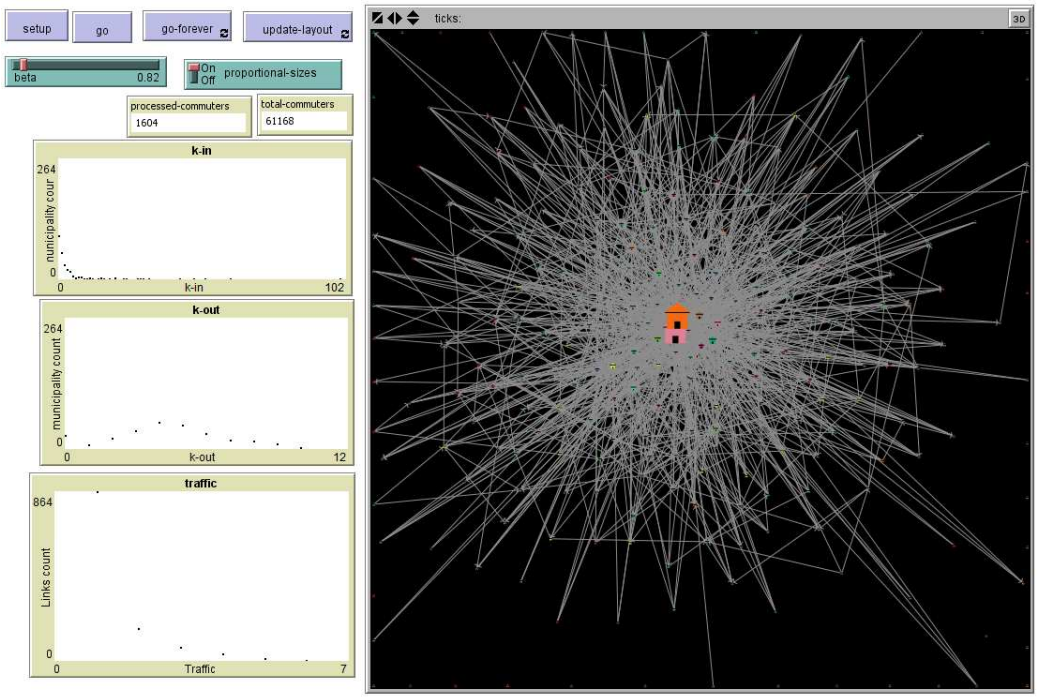


Figure 8: Interface of sample model implementation in NetLogo

[4] http://ccl.northwestern.edu/netlogo/

The buttons *go* and *go-forever* are used to run the simulation for one step or for a continuous loop (until the total number of commuters has been processed). The *update-layout* button runs a network layout procedure until pressed again; this may be used to improve the visual position of the network (it does not have any effect on the simulation results).

Results are reported in the three provided charts, which show the distribution of in-commuters, out-commuters, and traffic (number of links with a number of commuters are present). The number of processed commuters and the total number of commuters read from the input files is also shown.

APPENDIX C

Generating French Virtual Commuting Network at Municipality Level

Contents

Abstract. We aim to generate virtual commuting networks in the rural regions of France in order to study the dynamics of their municipalities. Since it will be necessary to model small commuting flows between municipalities with a few hundred or thousand inhabitants, we have opted for the stochastic model presented by Gargiulo et al. (2012). This model reproduces various possible complete networks using an iterative process, stochastically selecting a workplace in the region for each commuter living in the municipality of a region. The choice is made considering the job offers in each municipality of the region and the distance to all of the possible destinations. This paper will present methods for adapting and implementing this model to generate commuting networks between municipalities for regions in France. We address three different issues: How can we generate a reliable virtual commuting network for a region

that is highly dependent on other regions for the satisfaction of its resident's demands for employment? What about a convenient deterrence function? How to calibrate the model when detailed data is not available? Our solution proposes an extended job search geographical base for commuters living in the municipalities, we compare two different deterrence functions and we show that the parameter is a constant for network linking municipalities in France.

Manuscript:
Lenormand, M., Huet, S. and Gargiulo, F. Generating French Virtual Commuting Network at Municipality Level. *arXiv:1109.6759v2* (Submitted in *Journal of Transport and Land Use*).

C.1 Introduction

The connection between the home and workplace plays a central role in understanding the socio-economic relations in a network of rural municipalities (Clark et al., 2003; Reggiani and Rietveld, 2010). Indeed, new economic theories assume local positive dynamics can be explained by implicit geographical money transfers made by commuters or retired people (see for example Davezies (2009)). Simulation is becoming an increasingly convenient tool to study populations and their interactions over the space. That is particularly the case with the individual-based approaches which allow studying theories at the individual level since they simulate the variations in how individuals interact with each other and with their environment. Recent modeling reviews show the increasing use of such a tool (Parker et al., 2003; Waddell et al., 2003; Bousquet and Le Page, 2004; Verburg et al., 2004; Rindfuss et al., 2004; Birkin and Wu, 2012). However, these approaches require generation models capable of building reliable virtual commuting networks that consider each individual within a population. That is the case in the *SimVillages* dynamic microsimulation model we developed during the PRIMA[1] project. Indeed, in the *SimVillages* model, after generating a synthetic population of individuals (Gargiulo et al., 2010), it is necessary to choose a place of work for each worker within this population because a commuting origin-destination table was unavailable.

The goal of the European PRIMA project was to understand the dynamics of rural municipalities in France. At the most, 95% of them have 3000 inhabitants. This means that most of the commuting flows we want to study are weak, with a spatial distribution largely determined by chance. This is why we opt for the stochastic model recently proposed by Gargiulo et al. (2012). Moreover, we want to consider the commuting network on different dates. Detailed data regarding flows between pairs of municipalities are only available in France for the year 1999. For other dates, the only reliable data is aggregated data for each municipality, which describes how many people work outside of the municipality and how many come from outside of the municipality to work. Such data lacks precision regarding the various places of work and the various municipali-

[1] PRototypical policy Impacts on Multifunctional Activities in rural municipalities - EU 7th Framework Research Programme; 2008-2011; https://prima.cemagref.fr/the-project

ties where citizens reside. Then we also choose the Gargiulo et al. (2012) model for its ability to generate a population of individuals on a commuting network, starting from this data. This model reproduces the complete network using an iterative process that stochastically selects a workplace in the region for each commuter living in the municipality of the region. The choice is made while considering the job offers in each municipality of the region and the distance to all possible destinations. It differs from the classical generation models presented in Ortúzar and Willumsen (2011) since it is a discrete choice model where the individual decision function is inspired by the gravity law model, which is not usually employed on an individual level (Haynes and Fotheringham, 1984; Ortúzar and Willumsen, 2011; Barthélemy, 2011). Moreover, such a model ensures that for every municipality the virtual total numbers of commuters both coming in and going out are the same as the ones supplied by the data. This paper presents a method to adapt and implement this model to generate commuting networks between municipalities for regions in France. This implementation has forced us to address three different issues: How can we generate a reliable virtual commuting network for a region highly dependent of other regions to satisfy the need for job for the people living in the municipalities? What about a convenient deterrence function? How should the model be calibrated when detailed data is not available?

The first problem to solve involves the fact that regions in France are not islands, as presented in the example of De Montis et al. (2007, 2010). Indeed, some of the inhabitants, especially those living close to the borders of the region, are likely to work in municipalities located outside the region of residence. This part, especially if it is significant, causes the generated network to register false if we only consider that people living in the region also work in the region. A method for solving this problem involves generating the commuting network only for people living and working in the region. However, in order to do this it is required that the modeler know the quantity and the place of residence for individuals who work outside but live in the region. Data providing this information is very rare. Therefore, we address this issue by extending the job search geographical base for commuters living in the municipalities to a sufficiently large number of municipalities located outside the region of residence. Then, we compare the model without outside municipalities and the model with outside municipalities in 23 regions in France and come to a conclusion regarding the quality of our solution.

The second problem relates to the form of the deterrence function which governs the impact of distance on choice of the place of work relative to the quantity of job offers. The initial work done by Gargiulo et al. (2012) propose the use of a power law. However, Barthélemy (2011) states that the form of the deterrence function varies greatly, and can sometimes be inspired by an exponential function, such as in Balcan et al. (2009), or by a power law function as in Viboud et al. (2006). To choose the much more convenient deterrence function, we have compared the quality of generated networks for 34 regions in France obtained with both the exponential law and the power law. Better results were obtained with the exponential law.

The final problem was related to calibration. The generation model, as with most of the currently used commuting network generation models, has one parameter to calibrate. This parameter governs the impact of distance on the individual decision regard-

ing the place of work relative to the quantity of job offers. This parameter was calibrated through minimization of the Kolmokorov-Smirnov distance between the observed and simulated commuting distance distribution for individuals of the studied region. When detailed data is not available, it is necessary to find a way to determine this parameter. The only available distance that can be used is the Euclidian distance. While detailed commuting network data was available for the year 1999 and could be used for calibration, it was not available for earlier or more recent years. Though it may be possible to assume the parameter value does not change over time, a transportation network can evolve greatly at the local level to reduce the time distance. Such a change cannot be recorded when using the Euclidian distance. A solution was finally found. Using 34 regions in France, we show that every region can be generated using a constant value for the parameter. Then, we assume that the parameter value is constant over time and space.

C.2 Material and methods

C.2.1 The French regions and data from the French statistical office

A complete description of the regions from which the network was generated is provided in Table C.4. These regions have been randomly chosen for their diversity in terms of number of municipalities, number of commuters and surface areas. Some correspond to a region while others are closer to the county (known as "departements" in French).

The French Statistical Office (INSEE) collects information regarding each individual's residence and place of work. From this collected data, the Maurice Halbwachs Center or the INSEE make the following data available for every researcher:

- in 1999, data regarding the numbers of individuals commuting from location i to location j for every municipality of a region;
- in 1990 and 2006, the total number of commuters, the total job offers and the total number of workers in residence for every municipality. These data allow computations to be made for the number of workers that commute to their office of employment for each municipality.

The Lambert coordinates for each municipality are easy to find on the internet. They allow calculations regarding the Euclidian distance between each pair of municipalities.

Using these data sets, we will begin our implementation of the model presented in the next section.

C.2.2 The Gargiulo et al. (2012) model

Consider a region composed of n municipalities. We can model the observed commuting network starting from matrix $R \in \mathrm{M}_{n\times n}(\mathbb{N})$ where R_{ij} represents the number of commuters from municipality i (in the region) to municipality j (in the region). This matrix represents the light gray origin-destination table presented in Table C.1.

Table C.1: Origin-destination table for the region; The light gray table represents the commuters living (place of residence RP) and working (place of work WP) in the region for each municipality of the region; The dark gray line represents the number of out-commuters from municipality of the region to the region for each municipality of the region (i.e. the row totals of the light gray table); The dark gray column represents the number of in-commuters from the region to a municipality of the region for each municipality of the region (i.e. the column totals of the light gray table).

RP \ WP	M_1	...	M_j	...	M_n	**Total**
M_1	0	...	R_{1j}	...	R_{1n}	O_1
...	...	...	...	...	...	...
M_i	R_{i1}	...	R_{ij}	...	R_{in}	O_i
...	...	...	...	...	...	...
M_n	R_{n1}	...	R_{nj}	...	0	O_n
Total	I_1	...	I_j	...	I_n	

The inputs of the algorithm are:

- $D=(d_{ij})_{1\leq i,j\leq n}$ the Euclidean distance matrix between municipalities.
- I_j the number of in-commuters from the region to municipality j of the region, $1\leq j\leq n$ (i.e. the number of individuals living in the region in municipality i ($i\neq j$) and working in municipality j).
- O_i the number of out-commuters from municipality i of the region to the region, $1\leq i\leq n$ (i.e. the number of individuals working in the region in municipality j ($j\neq i$) and living in municipality i).

I_k and O_k can be respectively assimilated to the job offers for those employed in the region and the job demand of those employed in the region for municipality k, $1\leq k\leq n$. The algorithm starts with:

$$I_j=\sum_{i=1}^{n}R_{ij} \tag{C.1}$$

and

$$O_i=\sum_{j=1}^{n}R_{ij} \tag{C.2}$$

The purpose of the model is to generate the light gray origin-destination sub-table of the region described in Table C.1. To do this it generates matrix $S\in \mathrm{M}_{n\times n}(\mathbb{N})$ where S_{ij} represents the number of commuters from municipality i (in the region) to municipality j (in the region). It's important to note that $S_{ij}=0$ if $i=j$. The algorithm assigns to each individual a place of work with a probability based on the distance from the place of

residence to every possible place of work and their corresponding job offer. The number of in-commuters for municipality j and the number of out-commuters for municipality i decrease each time an individual living in i is assigned municipality j as a workplace. The algorithm is stopped when all out-commuters have a place of work. The algorithm is described in Algorithm C.1 with $m = n$.

Algorithm C.1 Commuting generation model

INPUT: $D \in \mathrm{M}_{n\times m}(\mathbb{R})$, $I \in \mathbb{N}^m$, $O \in \mathbb{N}^n$, $\beta \in \mathbb{R}_+$
OUTPUT: $S \in \mathrm{M}_{n\times m}(\mathbb{N})$
  $S_{ij} \leftarrow 0$
  while $\sum_{i=1}^{n} O_i > 0$ **do**
    Simulate $i \sim \mathscr{U}_A$ where $A = \{k | k \in |[1,n]|, O_k \neq 0\}$
    Simulate j from $|[1,m]|$ with a probability:

$$P_{i\to j} = \frac{I_j f(d_{ij},\beta)}{\sum_{k=1}^{m} I_k f(d_{ik},\beta)}$$

    $S_{ij} \leftarrow S_{ij} + 1$
    $I_j \leftarrow I_j - 1$
    $O_i \leftarrow O_i - 1$
  end while
  return S

Gargiulo et al. (2012) uses deterrence function $f(d_{ij},\beta)$ with a power law shape:

$$f(d_{ij},\beta) = d_{ij}^{-\beta} \quad 1 \leq i,j \leq n \ . \tag{C.3}$$

C.3 Statistical tools

This section presents the tools used to calibrate the model and to compare various implementation choices.

C.3.1 Calibration of the β value.

The same method used in Gargiulo et al. (2012) is used to calibrate the β value. β is calibrated so as to minimize the average Kolmogorov-Smirnov distance between the simulated commuting distance distribution and one building from the observed data. For the basic model we compute the commuting distance distribution with the commuting distance of individuals who are commuting from the region to the region. For the model focused on the outside we compute the commuting distance distribution with the commuting distance of the individuals who are commuting from the region to the region and outside.

As Gargiulo et al. (2012) model is stochastic, the final calibration value we consider is the average β value over ten replications of the generation process.

C.3.2 An indicator to assess the change.

It is necessary to have an indicator to compare the simulated commuting network and the observed commuting network. Let $R \in \mathrm{M}_{n_1 \times n_2}(\mathbb{N})$ represent a commuting network when R_{ij} represents the number of commuters from municipality i to municipality j. Let $S \in \mathrm{M}_{n_1 \times n_2}(\mathbb{N})$ represent another commuting network for the same municipalities. We can calculate the number of common commuters between R and S (Equation C.4) and the number of commuters in R (Equation C.5):

$$NCC_{n_1 \times n_2}(S,R) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \min(S_{ij}, R_{ij}) \tag{C.4}$$

$$NC_{n_1 \times n_2}(R) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} R_{ij} \tag{C.5}$$

From Equation C.4 and Equation C.5 we calculate the Sørensen similarity index (Sørensen, 1948). This index is suitable because it corresponds to the common part of commuters between R and S. Thus it is called the common part of commuters (CPC) (Equation C.6):

$$CPC_{n_1 \times n_2}(S,R) = \frac{2NCC_{n_1 \times n_2}(S,R)}{NC_{n_1 \times n_2}(R) + NC_{n_1 \times n_2}(S)} \tag{C.6}$$

This index has been chosen for its intuitive explanatory power, as it is a similarity coefficient that provides the likeness degree between two networks. The index ranges from a value of zero, for which there are no any commuter flows in common in the two networks, to a value of one, when all commuter flows are identical between the two networks.

C.4 Generating commuting networks for French regions at municipality level

C.4.1 How to cope with regions that are not islands or those that lack detailed data?

A commuting network is defined by an origin-destination table (light gray table in Table C.2). At the regional level, this means that it is necessary to know, for each municipality of residence and for each municipality of employment, the value for the flow of commuters traveling from one to another. This kind of data is not always provided by statistical offices and the datasets are usually aggregated: only the total number of out-commuters and in-commuters for each municipality is available for each (dark gray row and colum in Table C.2). To apply the model and define the commuting network, unless we are on a significantly isolated region[2], we need to find a way to isolate from the total number of in(out)-commuters (dark gray row and colum in Table C.2) the fraction that

[2] An island for example, in this case gray rows and colums in Table C.2 would not exist

relates strictly to the region (light gray table in Table C.2). However, this is not a simple task.

Furthermore, even if these parts can be isolated, a problem remains due to the border effect. Indeed, if we consider only the region, there is the risk of making an error in the reconstruction of the network for municipalities near the region's border. The higher the proportion of individuals working outside of the region, the more significant the error will be.

To go further, we propose to change the inputs for the algorithm. Instead of only considering the regional municipalities as possible places of work, we also consider an outside of the region. The outside represents the surroundings of the studied area. The following section describes a method for considering this outside area practically.

C.4.1.1 A new extended to outside job search base.

We implement the model, while choosing whether or not to take the outside into account, to generate 23 various regions in France. Their outside is composed of the set of municipalities of their neighboring "departments".

We consider the outside of the region to be composed of $m-n$ municipalities, where n represents the number of municipalities in the region. The inputs are the directly available aggregated data at the municipal level:

- $D = (d_{ij})_{\substack{1\leq i\leq n \\ 1\leq j\leq m}}$ the Euclidean distance matrix between municipalities both in the same region and in the outside.
- $(I_j)_{1\leq j\leq m}$ the total number of in-commuters of municipality j of the region and outside of it (i.e. the number of individuals working in municipality j of the region or the outside and living in another municipality).
- $(O_i)_{1\leq i\leq n}$ the total number of out-commuters of municipality i of the region only (i.e. the number of individuals living in municipality i of the region and working in an other municipality).

The purpose of the algorithm that introduces the outside is to generate the origin-destination table (light gray and gray sub-table in Table C.2). To do this the algorithm presented in Algorithm C.1 is used to simulate the Table C.3. From this, through difference the Table C.2 can be obtained with the total number of in-commuters $(I_j)_{1\leq j\leq n}$, the total number of out-commuters $(O_i)_{1\leq i\leq n}$ and the light gray table of the Table C.3.

A matricial representation of the origin-destination table presented in the light gray and gray sub-table in Table C.2, known as the simulated matrix $S \in \mathrm{M}_{(n+1)\times(n+1)}(\mathbb{N})$ is obtained. S_{ij} represents:

- the number of commuters from municipality i (in the region) to municipality j (in the region) if $i, j \neq n+1$;
- the number of commuters from outside to municipality j (in the region) if $i = n+1$ and $j \neq n+1$;
- the number of commuters from municipality i to outside if $i \neq n+1$ and $j = n+1$.

Table C.2: Origin-destination table; The light gray table represents the commuters living and working in the region for each municipality of the region; The gray column represents the out-commuters living in the region and working outside (Out.) for each municipality of the region; The gray line represents the in-commuters working in the region and living outside (Out.) for each municipality of the region; The dark gray line(column) represents the total number of out(in)-commuters for each municipality of the region.

WP / RP	M_1	...	M_j	...	M_n	Out.	Total
M_1	0	...	R_{1j}	...	R_{1n}	$R_{1\text{out}}$	O_1
...	...	...	...	...	...	...	...
M_i	R_{i1}	...	R_{ij}	...	R_{in}	$R_{i\text{out}}$	O_i
...	...	...	...	...	...	...	...
M_n	R_{n1}	...	R_{nj}	...	0	$R_{n\text{out}}$	O_n
Out.	$R_{\text{out}1}$	...	$R_{\text{out}j}$	...	$R_{\text{out}n}$		
Total	I_1	...	I_j	...	I_n		

Table C.3: Origin-destination table from the region to the region and the outside; The light gray table represents the commuters living (place of residence RP) and working (place of work WP) in the region for each municipality of the region; The gray table represents the commuters living (place of residence RP) in the region and working (place of work WP) outside of the region.

WP / RP	M_1	...	M_j	...	M_n	M_{n+1}	...	M_m
M_1	0	...	R_{1j}	...	R_{1n}	R_{1n+1}	...	R_{1m}
...	...	...	...	...	...	...	...	...
M_i	R_{i1}	...	R_{ij}	...	R_{in}	R_{in+1}	...	R_{im}
...	...	...	...	...	...	...	...	...
M_n	R_{n1}	...	R_{nj}	...	0	R_{nn+1}	...	R_{nm}

C.4.1.2 Comparison of the two models: Assessing the impact of the outside.

We assess the impact of the outside through a comparison between the network generations for 23 French regions both with and without the outside. The generation is made on a municipality scale using a power law deterrence function.

Both implementations are compared through their CPC values for each region. We replicate the generation for each region ten times and our indicator on each replicate is calculated. In all the presented figures, the indicator averages ten replications. The variation of the indicator over the replications is very low, averaging 1.02% at most. Consequently, this is not represented on the figures. Figure C.1 presents the common part of commuters $CPC_{n\times n}(S,R)$ between the simulated network S and the observed network R obtained with the regional job search base (square) and obtained with a job

search base comprising the region and its outside (triangle). It's important to note that for the implementation without outside $S \in \mathrm{M}_{n\times n}(\mathbb{N})$ while for the implementation with outside $S \in \mathrm{M}_{(n+1)\times(n+1)}(\mathbb{N})$. In order to compare the two models, the regional network (commuters from the region to the region) must be taken into consideration. Indeed, in the without-outside cases $NC_{n\times n}(S) = NC_{n\times n}(R)$ but this is not necessarily true for the with-outside cases.

Figure C.1 shows that the two job search bases give results which are not different. Thus, introducing the outside solves the problem linked to a lack of detailed data without changing the quality of the resulted simulated network. Indeed, one must keep in mind that the inputs for the with-outside cases do not require detailed data in comparison to the without-outside cases.

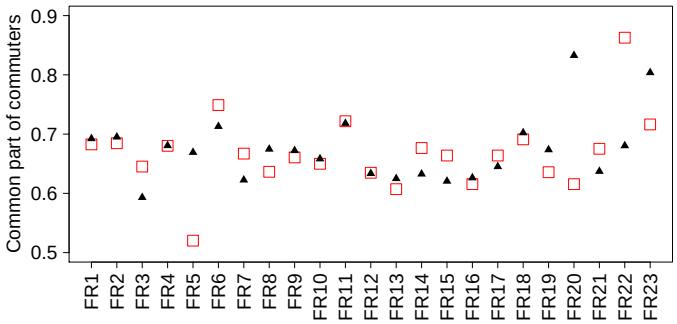


Figure C.1: Average CPC for 23 regions. The squares represent the basic model; The triangles represent the model with outside.

C.4.2 Choosing a shape for the deterrence function

The next problem relates to the form of the deterrence function which rules the impact of distance on the choice of the place of work relative to the quantity of job offers. The initial work done by Gargiulo et al. (2012) proposes to use a power law. However, Barthélemy (2011) states the form of the deterrence function varies significantly, and can sometimes be inspired by an exponential function as in Balcan et al. (2009) or by a power law function as in Viboud et al. (2006). Through choosing the much more convenient deterrence function, we compare the quality of generated networks for 34 French regions obtained with the model with outside using both the exponential law and the power law.

A deterrence function following an exponential law is introduced:

$$f(d_{ij},\beta) = e^{-\beta d_{ij}} \quad 1 \le i \le n \text{ and } 1 \le j \le m \ . \tag{C.7}$$

To compare the two deterrence functions, we have generated the networks of 34 various French regions (see Table C.4 for details) that replicate ten times for each region. The networks were generated with a job search base for the algorithm that considers the outside.

For example, Figure C.2 shows that we obtained a better estimation of the Auvergne commuting distance distribution when using the exponential law.

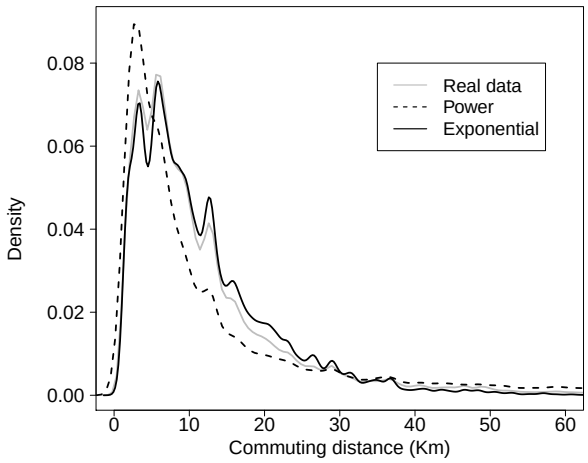


Figure C.2: Density of the Auvergne commuting distance distribution; the solid line represents the observed commuting distance distribution; the dotted line represents the commuting distance distribution obtained with the calibrated model with a job search base comprising the outside and the exponential law; the dashed line represents the commuting distance distribution obtained with a job search base comprising the outside and the power law. The two simulated commuting distance distribution are computed for one replication each.

More systematically, we plot, for the exponential law and power law, the average of the replications for the common part of commuters $CPC_{(n+1)\times(n+1)}(S,R)$ in Figure C.3. This clearly indicates that the average proportion of common commuters is always better when using an exponential law represented by squares.

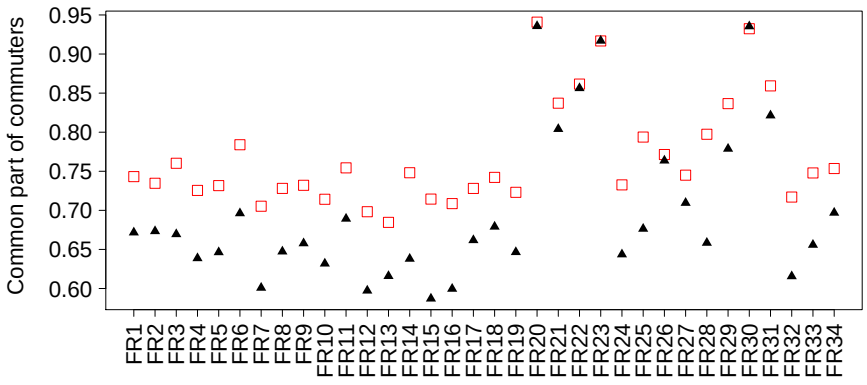


Figure C.3: Average CPC for the power shape (triangle) and the exponential shape (square) for 34 french regions.

C.4.3 Spatial Analysis

To better understand how CPC is spatially distributed at a more granular level we mapped the CPC by municipality for three models and three study areas. In Figure C.4, it can be observed that for all case studies (in rows) the highest values of the CPC were obtained by municipalities using the model with an exponential shape including the outside (third column). It can also be noted that the model without the outside (second column) and the model with the power shape including the outside (first column) give results which are not wholly different.

As we can see in Figure C.4, the CPC values are not uniformly distributed in the municipalities of the three areas. The error seems to increase as distance from the urban areas increases.

We now focus on the third model with an exponential shape including the outside to better understand which types of municipalities compose the three clusters (CPC ≤ 0.5, $0.5 <$ CPC ≤ 0.75 and $0.75 <$ CPC). We identify the number of out-commuters as the most explanatory variable. Indeed, we can observed in Figure C.5 that the distribution of the number of out-commuters in each cluster is significantly different. The higher the average number of out-commuters, the higher the CPC. Having performed analyses of variance (ANOVA) for each case study, we obtained significant differences between the averages for the number of out-commuters in each cluster with a 0.95% level of confidence for each case study.

For the three regions, the CPC value is strongly linked to municipality characteristics. Indeed, the municipalities with $0.75 <$ CPC are urban and suburban municipalities with a high number of out-commuters that are closed to a large urban municipality. In contrast, the municipalities with a low number of out-commuters that are far from large urban municipalities have a CPC lower than 0.5. For this type of municipality, the commuting flows are very small. Thus they are difficult to reproduce with the mechanisms taken into consideration. However, the distance to cities does not appear to be particularly responsible for the error. The timing for the job offer arrival on the job market is probably much more significant in determining the local topology of the network than elsewhere. These flows represent about 4% of the total number of out-commuters for the Auvergne region, 1% for Bretagne and 5% for Aquitaine.

C.4.4 Calibrating the model for French regions

The final problem involves the calibration process, which previously required detailed and accurate data.

Figure C.6 shows the calibrated β values for each of the 34 regions in France. It can be observed that these values display subtle variations from about $1.7 \cdot 10^{-4}$ to $2.4 \cdot 10^{-4}$ with the average β valued ($C = 1.94 \cdot 10^{-4}$) corresponding to the red line.

Then we hypothesize that it is possible to directly calibrate the algorithm to generate the 34 regions in France, by using a constant equal to C. To study the influence of this approximation on the common part of commuters we have computed the CPC with C as the parameter value for the 34 regions. We observe in Figure C.7 that the influence

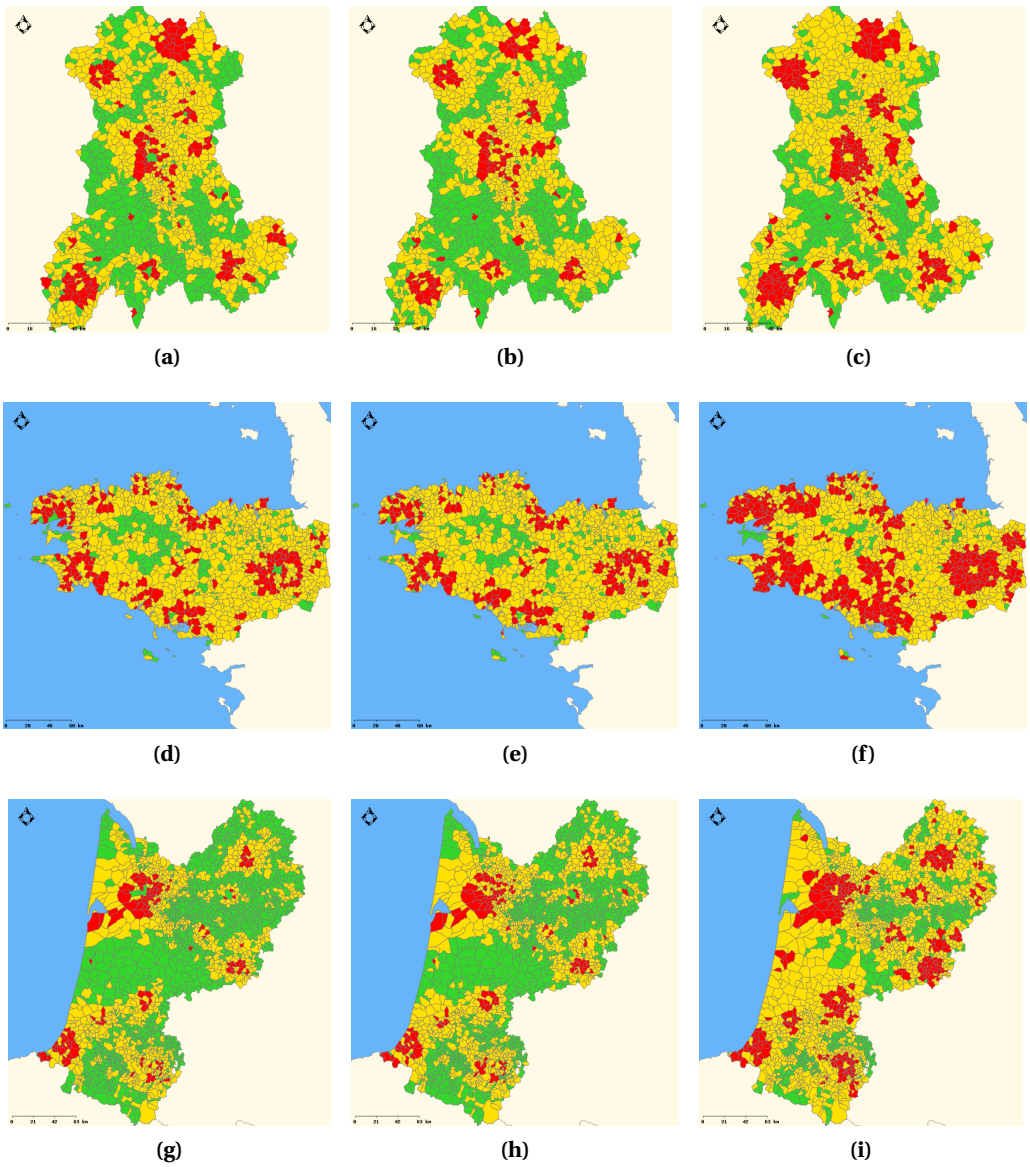


Figure C.4: Maps of the average CPC by municipalities obtained with ten replications. In green CPC ≤ 0.5; In yellow $0.5 <$ CPC ≤ 0.75; In red $0.75 <$ CPC. (a), (d) and (g) Model with the power shape without outside; (b),(e) and (h) Model with the power shape with outside; (c), (f) and (i) Model with the exponential shape with outside. (a)-(c) Auvergne case-study; (d)-(f) Bretagne case-study; (e)-(h) Auquitaine case-study. *Base maps source: Cemagref - DTM - Développement Informatique Système d'Information et Base de Données : F.Bray & A.Torre IGN (Géofla®, 2007).*

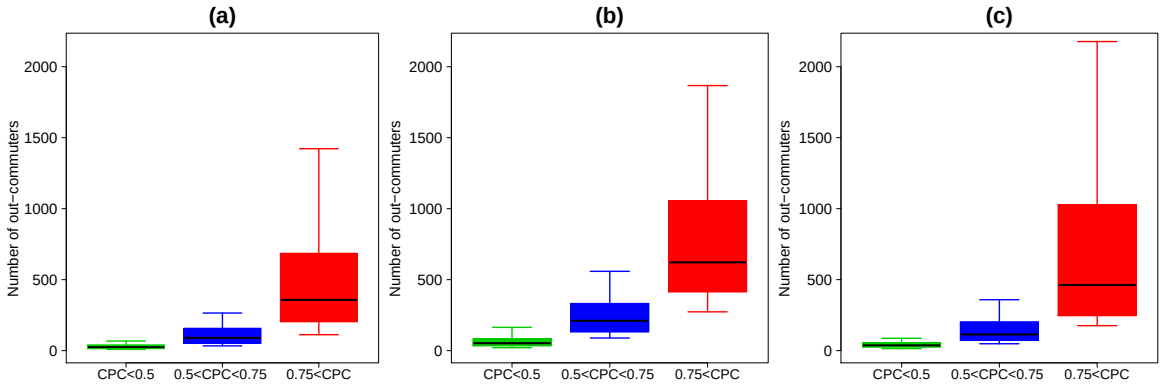


Figure C.5: Boxplots of the number of out-commuters in term of the CPC by municipality for the model with the exponential shape with outside. (a) Auvergne case study; (b) Bretagne case study; (c) Aquitaine case study.

of the β's approximation on the CPC is very weak. It can then be noted that the average CPC obtained with C is, for some regions, higher than the CPC obtained by the β value that is not averaged. It is possible that the common part of commuters is better with another beta value because it is not a calibration criterion.

It is not necessary to study the influence of the β's approximation on the calibration criterion. Indeed, from the studies made by Gargiulo et al. (2012), we know the CPC and the calibration criterion show a significant correlation. The CPC and the calibration criterion follow the same evolution in terms of β. The β value for minimization of the Kolmogorov-Smirnov distance is very close to the one obtained for maximization of the CPC (see the Figure 7 in Gargiulo et al. (2012) which perfectly illustrates this relation). The CPC values remain quasi-identical to β=C or to β valued from the calibration process presented in Sub-section C.3.1, the quality of the approximation of the calibration criterion, i.e. the commuting distance distribution, remains the same.

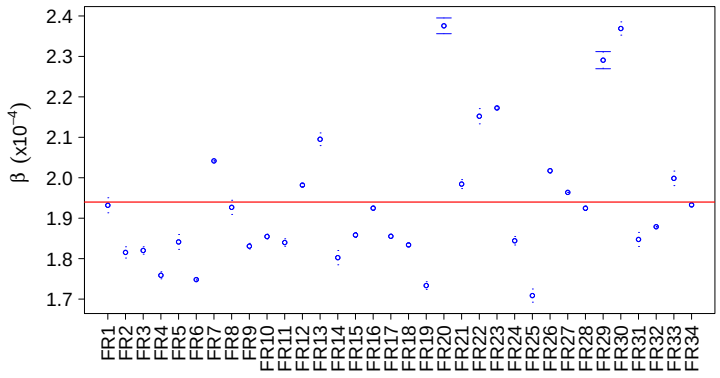


Figure C.6: The circle represents the average calibrated β values for ten replications (The confident interval is composed of the minimum and the maximum) for each regions; the line represents the average β value for the 34 regions.

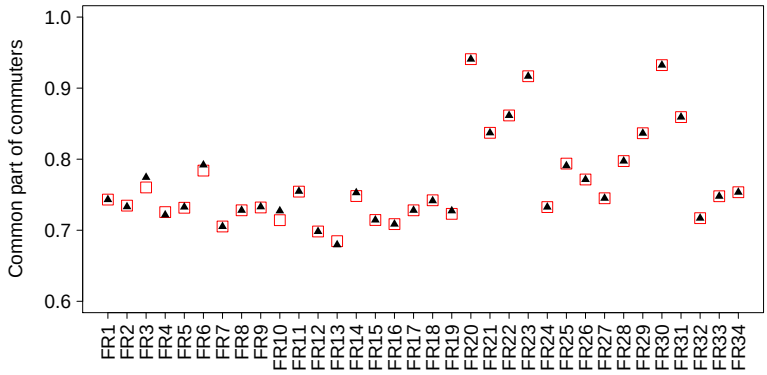


Figure C.7: Common part of commuters for the 34 regions; The squares represents the average CPC (10 replications) obtained with the calibrated β value; The triangles represents the average CPC (10 replications) obtained with the estimated β values (average β value over the 34 calibrated β values).

C.5 Discussion and conclusion

To study the rural area dynamics through microsimulation, we need virtual commuting networks that link individuals living in the municipalities of various French regions. As the studied scale is very low, the flows are low, and we thus decided to opt for a stochastic generation algorithm. The one recently proposed by Gargiulo et al. (2012) is relevant to our problem. Starting from this model, we implement the commuting networks of 34 different French regions. The implementation work leads us to solve three practical problems.

The first problem involves the fact that our French regions are not islands. Indeed, some of the inhabitants, especially those living close to the border of the region, are likely to work in municipalities located outside the region of residence. However, classical approaches to generating commuting networks consider only residents of the region that work in the region. That is also the case for ours. Data providing details, or knowledge, allowing the modeler to evaluate people living in the region but working outside is difficult to obtain. Thus, we address this issue by extending the geographical base of the job search for commuters living in the municipalities to a sufficiently large number of municipalities located outside the region of residence. We compare the model without municipalities located outside and the model with outside municipalities to 23 French regions. We are able to come to a conclusion regarding the relevance of our solution which keeps the value of our quality indicator identical. At the same time, it is not necessary to have information regarding those who do not work in the region, which allows us to generate networks using only the aggregated data.

The Gargiulo et al. (2012) model is based on the gravity law. Then, our second problem relates to the deterrence function, which is more of a power law or an exponential law depending on the study. Moreover, as empirical studies comparing generated networks to "real" data are extremely rare (Barthélemy, 2011), few know which is better.

In order to select the more convenient one for our French regions, we have compared the quality of generated networks for 34 regions obtained with both the exponential law and the power law. Better results were obtained with the exponential law, no matter the region. Indeed, the 34 regions display significant variance in regards to surface area, the number of municipalities, and the number of commuters.

The final problem involved calibration. Applying a model with an extended job search base and an exponential deterrence function, we found a constant equal to $1.94 \cdot 10^{-4}$ to be a perfect parameter value for generating commuting networks for French administrative regions, no matter the region. However, we did not test this result for other countries with different types of administrative regions. The robustness of this result to commuting networks of different scales has been studied in Lenormand et al. (2012). The β value correlated to a scale consistent with the results obtained in this paper.

A spatial analysis of three different case studies has been proposed, and it was shown that the CPC value by municipality strongly correlated with the number of out-commuters for the municipality. Our model is not able to reproduce very small flows which represent between 1 and 5% of the total flows in the region we studied. However, we continue to question if it makes sense to attempt to reproduce them.

Acknowledgement

This publication has been funded by the Prototypical policy impacts on multi-functional activities in rural municipalities collaborative project, European Union 7th Framework Programme (ENV 2007-1), contract no. 212345. The work of the first author has been funded by the Auvergne region.

References

Balcan, D., Colizza, V., Goncalves, B., Hud, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.

Barthélemy, M. (2011). Spatial Networks. *Physics Reports*, 499:1–101.

Birkin, M. and Wu, B. (2012). A Review of Microsimulation and Hybrid Agent-Based Approaches. In Heppenstall, A. J., Crooks, A. T., See, L. M., and Batty, M., editors, *Agent-Based Models of Geographical Systems*, pages 51–68. Springer Netherlands.

Bousquet, F. and Le Page, C. (2004). Multi-agent simulations and ecosystem management: A review. *Ecological Modelling*, 176(3–4):313 – 332.

Clark, W. A. V., Huang, Y., and Withers, S. (2003). Does commuting distance matter?: Commuting tolerance and residential change. *Regional Science and Urban Economics*, 33(2):199 – 221.

Davezies, L. (2009). L'économie locale "résidentielle". *Géographie Economie Société*, 11(1):47–53.

De Montis, A., Barthélemy, M., Chessa, A., and Vespignani, A. (2007). The structure of interurban traffic: A weighted network analysis. *Environment and Planning B: Planning and Design*, 34(5):905–924.

De Montis, A., Chessa, A., Campagna, M., Caschili, S., and Deplano, G. (2010). Modeling commuting systems through a complex network analysis: A study of the Italian islands of Sardinia and Sicily. *The Journal of Transport and Land Use*, 2(3):39–55.

Gargiulo, F., Lenormand, M., Huet, S., and Baqueiro Espinosa, O. (2012). Commuting Network Models: Getting the Essentials. *Journal of Artificial Societies and Social Simulation*, 15(2):6.

Gargiulo, F., Ternes, S., Huet, S., and Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, 5(1).

Haynes, K. E. and Fotheringham, A. S. (1984). *Gravity and Spatial Interaction Models*. Sage Publications, Beverly Hills.

Lenormand, M., Huet, S., Gargiulo, F., and Deffuant, G. (2012). Universal Commuting Network Model. *PLoS ONE*, 7(10):e45985.

Ortúzar, J. and Willumsen, L. (2011). *Modeling Transport*. John Wiley and Sons Ltd, New York.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., and Deadman, P. (2003). Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review. *Annals of the Association of American Geographers*, 93(2):314–337.

Reggiani, A. and Rietveld, P. (2010). Networks, commuting and spatial structures: An introduction. *The Journal of Transport and Land Use*, 2(3):1–4.

Rindfuss, R. R., Walsh, S. J., Turner, B. L., Fox, J., and Mishra, V. (2004). Developing a science of land change: Challenges and methodological issues. *Proceedings of the National Academy of Sciences*, 101(39):13976–13981.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34.

Verburg, P. H., Schot, P. P., Dijst, M. J., and Veldkamp, A. (2004). Land use change modelling: current practice and research priorities. *GeoJournal*, 61(4):309–324.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M., and Ulfarsson, G. (2003). Microsimulation of urban development and location choices: Design and implementation of Urbansim. *Networks and Spatial Economics*, page 2003.

Table C.4: Description of the regions

ID	Region	Number of municip. (region)	Number of municip. (outside)	Region area (km^2)	Average municip. area (km^2)	Number of commuters
FR1	Auvergne	1310	3463	26013	19.86	295776
FR2	Bretagne	1269	1447	27208	21.44	653710
FR3	Ain	419	2809	5762	13.75	162370
FR4	Alsace	903	3081	8280	9.17	440961
FR5	Aquitaine	2296	2835	41309	17.99	700452
FR6	Mayenne	261	3124	5175	19.83	69915
FR7	Lozère	185	1859	5167	27.93	12273
FR8	Poitou-Charente	1464	2467	25810	17.63	375363
FR9	Centre	1842	4718	39151	21.25	624693
FR10	Midi-Pyrénée	3020	3845	45348	15.02	546162
FR11	Limousin	747	3169	16942	22.68	139481
FR12	Franche-Comté	1786	3317	16202	9.07	268399
FR13	Haute-Normandie	1420	3536	12317	8.67	469335
FR14	Haute-Marne	433	3914	6211	14.34	42690
FR15	Vosges	515	3808	5874	11.41	92053
FR16	Lorraine	2339	3067	23547	10.07	547457
FR17	Creuse	260	1814	5565	21.40	23949
FR18	Languedoc-Roussillon	1545	3046	27367	17.71	409116
FR19	Charente-Maritime	1948	1983	25606	13.14	375363
FR20	Haut-de-Seine	36	1245	176	4.89	973173
FR21	Yveline	262	1543	2284	8.72	618741
FR22	Val d'Oise	185	1707	1246	6.74	526600
FR23	Val de Marne	47	1234	245	5.21	642092
FR24	Haut-Rhin	377	2283	3525	9.35	183504
FR25	Tarn et Garonne	195	2338	3718	19.07	41600
FR26	Pyrénée-Atlantique	547	449	4116	7.52	65469
FR27	Alpes-Maritimes	163	353	4299	26.37	163445
FR28	Loire	327	2788	4781	14.62	178828
FR29	Territoire de Belfort	102	2031	609	5.97	45185
FR30	Seine-Saint-Denis	40	783	236	5.90	655200
FR31	Essonne	196	1597	1804	9.20	518321
FR32	Ardennes	463	2588	5229	11.29	59963
FR33	Aube	433	2728	6004	13.87	75561
FR34	Corréze	286	2088	5857	20.48	49815

APPENDIX D

Predicting the Presence and the Number of Jobs in Different Services in Auvergne Municipalities

Contents

D.1 Introduction

This study investigates the possibility to define the endogeneous dynamics of creations and destructions of services, as well as the creation and destruction of jobs, in the PRIMA microsimulation model[1]. The aim is to include in the model some rules that create or destroy services, and to create or destroy jobs in the services according to changes in the population of the municpality.

Hence, the objective of this paper is to elaborate statistical rules that provides the likely availability of services and the number of jobs in Auvergne municipalities except *Clermont-Ferrand* (1309 municipalities), from a set of variables describing the municipality (demographic and geographic descriptors). The demographic and geographic data describing the municipality are provided by the *"Census"*. The numbers of jobs by services come from a survey carried out on one quarter of the population.

[1] see Huet and Deffuant, 2010, a Common Framework for the Microsimulation Model in PRIMA project, PRIMA working Paper

D.2 Predicting the number of jobs in services

D.2.1 Explanatory variables

The explanatory variables used in the model are :

- Density of population (people by meter^2)
- Population
- Population 0-14 years
- Population 15-29 years
- Population 30-59 years
- Population over 60 years
- Number of unemployed
- Number of outgoing commuters
- Number of incoming commuters
- Number of farmers employed
- Percentage of retired people
- Travel time to access the most frequented town (min)

We assume that the number of job in services depends on the municipalities caracteristics and location. The explanatories variables were selected to that effect.

D.2.2 Method

Let J be the variable stating the number of jobs of a given category of services in the different municipalities. Let $Y = \{Y_i\}_{1 \leq i \leq p}$ be the p explanatories variables describing the demography and geography of the municipalities. The purpose of this method is to determine a subset of q decorrelated variables E, $E \subseteq Y$, which provides the best prediction of J through a GLM :

$$J = \beta_0 + \sum_{i=1}^{q} \beta_i Y_{(i)} + \epsilon$$

Where β_i ($j \in |[0,q]|$) are parameters and ϵ is the residual vector.

Algorithm

Initially, we choose the Y variable the most correlated whith J, noted $Y_{(1)}$. We obtain two sets of variables $E_1 = Y_{(1)}$ and $E_2 = Y/Y_{(1)}$. The set E_1 represents the selected variables for the model. In the next steps, the algorithm tries to add relevant variables to E_1:

• We define subset of candidate variables E_{21} selected in E_2, as follows: a variable $Y_{(i)}$ of E_2 is selected if and only if we have $|cor(Y_{(i)}, Y_{(j)})| < r$ for all $Y_{(j)} \in E_1$ (r is a parameter and $0 \leq r \leq 1$). This enables us to insure the independence of candidates variables with the variables of the model. If $E_{21} = \emptyset$ we stop the algorithm.

• Then, we choose in E_{21} the variable which omptimises a criterion C (see below for different type of criteria). We include this variable into E_1 and we suppress this variable of a set E_{21}. If there is no variable which optimises the criterion we stop the algorithm.

• It's possible that because of the addition of a variable in the model, one or several coefficients of variables already in the model become non-significant(see below for the significativity-test). In this case, we suppress these variables from E_1 and we include them into E_{21}.

So we have $E_2 = E_{21} \cup E_{22}$

Criterion

Residual sum of square (RSS)

$$RSS = \sum_{i=1}^{n} (X_i - \hat{X}_i)^2$$

When $\hat{X}_i$ is the approximation of X_i by the GLM, $1 \leq i \leq n$.

Akaike information criterion (AIC)

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2(q+1)$$

Bayesian information criterion (AIC)

$$BIC = n \ln\left(\frac{RSS}{n}\right) + \ln(n)(q+1)$$

Correlation criterion

If Z is the variable candidate the correlation criterion is $|cor(Z,(X-\hat{X}))|$, when $\hat{X}$ is the approximation of X with the GLM without the variable candidate Z.

Significance criterion

For a model :

$$X = \beta_0 + \sum_{i=1}^{p} \beta_i Y_i + \epsilon$$

The significance test for a coefficient β_k $(0 \leq k \leq p)$ is: H_0 : "$\beta_k = 0$" and H_1 : "$\beta_k \neq 0$". Under the null hypothesis, the statistic $T = \frac{\beta_k}{\sqrt{\frac{\|X-\hat{X}\|^2}{n-p-1}({}^t YY)^{-1}_{kk}}}$ follow a law S_{n-p-1} of Student with $n-p-1$ degree of freedom. So the coefficient for a risk α is no-significant if the p-value $p = \mathbb{P}(S_{n-p-1} > |T|) > \alpha$.

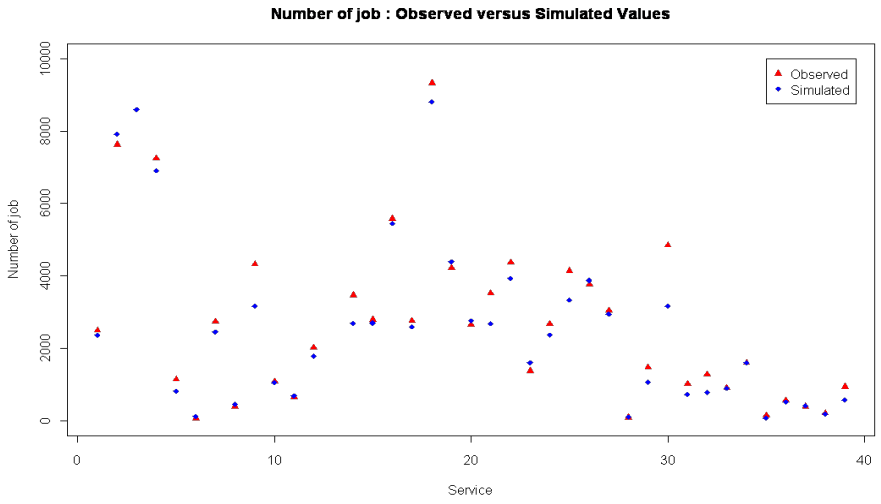


Figure D.1: Number of jobs. Blue diamonds: simulated values; Red triangles: observed values.

D.2.3 Results

We have separated the 1309 munuicipalities in two groups. The first one is used for the construction of the model (66%) and the second one for the validation (33%). For each service, we have used the method described previously, the coefficients associated to the variables are discribed in the Table D.1. We observed that the higher are the population, the population between 15 and 29 years and the percentage of retired people, the higher is the number of jobs in services. The farther is the municipality from the most visited town, the higer is the number of jobs in services.

To validate the model we have constructed two indicators :

- Percentage of good answers. The predictions are rounded to the nearest multiple of 4 because the data come from a survey about one quarter of the population.
- Percentage of good answers with an error inferior or equal to 4.

The results are shown in the Table D.2. On average, we have around 80% of good answers and 13% of errors are inferior or equal to four.

To check the results at the regional level we draw up the graphic below. For each service, we compare the real and the simulated number of jobs in the region. We observed that the results are globally good.

The next step is to continue this study with groups of services (such as health, education, leisure, basic).

Table D.1: Coefficient of the GLM for each kind of service

Service	Intercept	Population	Percentage of retired people	Travel time to access the most visited city
Nursery school	-2,0401	0,0035	0,0293	0
Elementary school	-2,0226	0,0075	0	0,0944
Junior hight school	-9,8059	0,0115	0	0,4470
High school	-21,8209	0,0171	0,4842	0
Nursery	-2,1317	0,0020	0,0411	0
Health Center	-0,0283	0,0002	0	0
Pharmacy	-3,1896	0,0034	0,0430	0,0714
Ambulance	-0,1429	0,0005	0	0
General Praticioner	-5,5929	0,0055	0,1213	0
Dentist	-1,5018	0,0017	0,0312	0
Veterinary	-0,7194	0,0008	0	0,0433
Auxiliary Medical	-1,5499	0,0021	0,0379	0
Hospital	-43,7759	0,0316	1,0268	0
Clinical	-0,8810	0,0060	0	-0,1365
Police Station	-6,6966	0,0057	0,1565	0
Post Office	-4,1422	0,0061	0	0,2216
Automotiv repair	-0,2955	0,0025	0	0
Mason	0,8415	0,0067	0	0
Carpenter	0,9397	0,0026	0	0
Plumber	-1,3877	0,0030	0	0,0697
Hairdressing	-4,2821	0,0043	0,1012	0
Restaurant	-2,6569	0,0047	0,0591	0
Groceries	0,1347	0,0012	0	0
Hypermarket	-5,4726	0,0051	0,1030	0
Supermarket	-3,1587	0,0047	0	0,1091
Bakery	-1,4188	0,0041	0	0,0575
Butcher meat	-2,9586	0,0030	0,0616	0,0660
Fishmonger	-0,0911	0,0002	0	0
Bookshop	-2,1375	0,0020	0,0482	0
Clothing store	-7,7926	0,0063	0,1806	0
Shoe store	-1,8320	0,0013	0,0341	0,0223
Appliance Store	-1,1066	0,0013	0,0237	0
Furniture store	-0,7909	0,0017	0	0
Drugstore	-0,7319	0,0022	0	0
Cinema	-0,2602	0,0002	0,0060	0
Fuel station	-0,1735	0,0006	0	0
Coffee Tobacco	-0,1558	0,0004	0	0,0121
Tobacco	-0,0808	0,0003	0	0
Laundry	-0,6510	0,0012	0	0

Table D.2: Percentage of good answer for each kind of service

Service	Percentage of good answers	Percentage of good answers with an error ≤ 4
Nursery school	82,56	94,42
Elementary school	53,02	83,26
Junior hight school	61,16	76,74
High school	73,02	82,09
Nursery	88,84	96,98
Health Center	99,07	100,00
Pharmacy	80,70	94,88
Ambulance	94,19	98,14
General Praticioner	80,23	92,33
Dentist	90,70	98,14
Veterinary	91,16	97,44
Auxiliary Medical	85,35	95,81
Hospital	70,23	76,51
Clinical	76,74	88,84
Police Station	80,00	92,56
Post Office	60,23	83,49
Automotiv repair	76,74	92,09
Mason	32,09	81,40
Carpenter	58,84	89,30
Plumber	74,19	95,12
Hairdressing	80,23	95,35
Restaurant	70,23	90,23
Groceries	79,07	92,79
Hypermarket	83,72	90,70
Supermarket	75,12	90,47
Bakery	72,79	89,07
Butcher meat	78,14	92,56
Fishmonger	98,60	100,00
Bookshop	90,23	96,74
Clothing store	78,60	91,16
Shoe store	91,86	97,21
Appliance Store	90,93	97,91
Furniture store	91,16	96,05
Drugstore	86,51	95,12
Cinema	99,30	99,77
Fuel station	93,95	98,37
Coffee Tobacco	94,65	99,07
Tobacco	97,44	99,07
Laundry	93,02	98,60
Mean	**81,19**	**92,97**

D.3 Predicting the presence and absence of services

In this section, we try to predict the presence or the absence of services in the different municipalities. Indeed, it can be important to include this dynamics in the model because it can have an impact on the probability that people settle in the municipality or not.

D.3.1 Explanatory variables

The explanatory variables used in the model are :

- Density of population (people by meter^2)
- Population
- Population 0-14 years
- Population 15-29 years
- Population 30-59 years
- Population over 60 years
- Number of unemployed
- Number of outgoing commuters
- Number of incoming commuters
- Number of farmers employed
- Number of jobs
- Percentage of retired people
- Travel time to access the most frequented town (min)
- Distance to the nearest urban pole (meter^2)
- Main town of a "canton" (binary)

We assume that the presence of a service depends on the municipalities caracteristics and location. The explanatories variables were selected to that effect.

D.3.2 Method

We have a variable to be explained X. We have a set of variables $Y = \{Y_i\}_{1 \leq i \leq p}$ composed by p explanatory variables. We use a *logit model*:

$$X = \begin{cases} 1 \text{ if there is the service} \\ 0 \text{ otherwise} \end{cases}$$

$$\mathbb{P}(X = 1 | Y_1, ..., Y_m) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i Y_i)}}$$

When β_i ($j \in |[0, p]|$) are parameters. We used a stepwise algorithm with *Bayesian information criterion* to select the variables the most explanatory for each service.

D.3.3 Results

For each services, we have used the method described previously to construct a GLM with the data of 1988 in Auvergne. The coefficients associated to the variables are discribing in the Table D.3. We observed that the variable used by the major part of services are *Population over 60 years* and *Main town of a "canton"*. The probability of a presence of a service increases if it's a main town of a canton.

To validate the model we have constructed tree indicators :

- Percentage of good answers.
- Percentage of 0 "catch" by the model.
- Percentage of 1 "catch" by the model.

The results are discribing in the Table D.4 for the Auvergne in 1988 and 2007. To validate the model we observed the results for 2007, about 62% of services have more than 70% of 1 "catch" by the model and about 90% of services have more than 70% of 0 "catch" by the model. The method is running well for allmost all the service.

The results at the national level are discribing in the Table D.5. We observed that about 50% of services have more than 70% of 1 "catch" by the model and about 89% of services have more than 70% of 0 "catch" by the model.

These results are reasonably satisfactory. However, the approach should now be adapted when considering broader groups of services (basic, health, education, leisure).

Table D.3: Coefficient of the GLM for each kind of service

Service	Intercept	Population	Population over 60 years	Percentage of retired people	Travel time to access the town's frequentest(min)	Main town of a canton	Distance to the nearest urban pole
Nursery school	-3,1525	0,0045	0	0	0	1,8316	0
Elementary school	-1,8003	0,0175	0	-0,0285	0	0	0
Junior hight school	-5,9126	0	0,0085	-0,0827	0,0644	2,7427	0,0001
Health Center	-4,6287	-0,0009	0,0063	0	0,0646	0	-0,0001
Pharmacy	-7,9636	0,0029	0,0156	0	0	5,0488	0,0001
Ambulance	-3,8480	-0,0011	0,0125	0	0	1,4751	0
General Praticioner	-6,7065	0,0035	0,0122	0	0	4,3575	0,0001
Nurse	-3,3989	0,0013	0,0090	0	0	2,1669	0
Physiotherapist	-3,1578	0	0,0130	-0,1016	0,0405	1,4224	0
Dentist	-6,5629	0,0036	0	0	0	2,7314	0,0001
Treasury	-6,0121	-0,0010	0,0063	0	0,0777	5,1709	0
Job center	-3,2001	0,0002	0	-0,0610	0,0546	3,2949	0
Police station	-6,4060	0	0,0033	0	0,0787	5,0247	0,0001
Post office	-3,4737	-0,0023	0,0337	-0,0410	0	16,6582	0,0001
Bank	-6,3183	0	0,0129	-0,0898	0,0731	3,3406	0,0001
Automotiv repair	-1,7686	0	0,0225	-0,0300	0	15,1279	0
Mason	-1,3064	0,0025	0,0074	0	0	0	0
Plasterer painter	-2,3924	0,0040	0	0	0	1,3873	0
Carpenter	-1,3039	0	0,0158	0	0	0	0
Plumber	-1,1131	0	0,0163	-0,0490	0	2,2831	0
Electrician	-1,7380	0	0,0155	-0,0354	0	1,6171	0
Hairdressing	-4,3541	0	0,0241	-0,0764	0	2,9036	0,0001
Veterinary	-4,3797	-0,0005	0,0050	0	0,0472	2,0757	0
Restaurant	0,2584	-0,0019	0,0232	-0,0252	0	0	0
Seniors: Accommodation	-4,4917	-0,0008	0,0082	0	0,0462	2,1472	0
Groceries	-3,0197	0	0,0251	0	0	14,4594	0,0001
Hypermarket	-2,9347	0,0001	0	0	0	0	-0,0080
Supermarket	-3,8019	0	0,0068	-0,1311	0,0900	0	0
Large surface craft	-4,6882	0	0,0018	0	0,0675	1,3724	-0,0001
Bakery	-3,1932	0	0,0234	0	0	15,6587	0
Butcher meat	-4,7440	0	0,0223	0	0	3,7717	0,0001
Fishmonger	-1,0961	0	0,0025	-0,0271	0,0250	0	0,0001
Bookshop Stationery	-5,8711	0	0,0105	0	0,0409	2,8087	0,0001
Clothing store	-5,3591	0	0,0085	0	0,0408	2,0967	0,0001
Shoe store	-5,3088	0	0,0066	0	0,0353	1,8684	0,0001
Appliance Store	-4,2433	0	0,0091	0	0	1,8648	0,0000
Furniture store	-3,4620	0	0,0038	0	0	1,6209	0
Drugstore	-3,5742	-0,0011	0,0155	-0,0484	0	2,3776	0,0001
Cinema	-5,1465	0	0,0016	0	0,0631	1,5374	0

Table D.4: Percentage of good answer for each kind of service in 1988 and 2007

Service	Percentage of good answers 1988	Percentage of 0 catch 1988	Percentage of 1 catch 1988	Percentage of good answers 2007	Percentage of 0 catch 2007	Percentage of 1 catch 2007
Nursery school	84,57	93,99	67,17	79,98	79,63	82,22
Elementary school	88,92	64,38	94,24	81,13	51,46	98,19
Junior hight school	95,95	98,47	73,88	95,42	98,48	66,40
Health Center	97,63	99,68	41,30	97,40	98,44	53,33
Pharmacy	94,81	97,85	83,80	94,42	97,45	83,85
Ambulance	91,37	97,76	54,40	92,21	94,57	70,99
General Praticioner	93,43	97,67	80,43	93,51	96,37	84,59
Nurse	87,24	96,39	63,76	89,92	96,27	72,09
Physiotherapist	92,67	97,55	66,99	90,30	99,52	53,44
Dentist	94,73	98,10	76,59	94,81	98,44	76,61
Treasury	97,17	98,40	85,60	96,03	96,36	91,92
Job center	94,04	98,01	48,08	97,48	97,91	64,71
Police station	96,49	98,70	79,74	97,10	98,46	85,92
Post office	83,19	92,15	70,64	88,77	90,25	85,11
Deposite and savings banks	96,41	98,51	82,04	92,21	98,90	59,64
Automotiv repair	77,31	83,55	71,60	79,91	83,92	73,80
Mason	75,02	69,18	78,51	71,28	62,80	80,00
Plasterer painter	79,60	91,71	62,84	76,09	86,65	60,57
Carpenter	73,26	67,26	77,02	70,44	64,90	76,08
Plumber	77,31	88,73	63,77	74,26	91,68	55,13
Electrician	79,37	92,12	60,15	78,53	92,20	57,39
Hairdressing	91,83	97,24	75,83	83,88	98,26	56,15
Veterinary	93,12	98,22	46,09	94,12	98,55	54,89
Restaurant	78,15	11,45	97,73	71,96	25,30	93,93
Seniors: Accommodation	92,90	97,74	57,86	92,82	98,39	60,00
Groceries	78,69	79,16	78,35	75,71	68,86	89,74
Hypermarket	99,31	99,92	20,00	98,85	100,00	11,76
Supermarket	96,72	99,02	65,56	94,65	99,50	44,35
Large surface craft	97,56	99,45	44,44	96,72	99,52	39,34
Bakery	82,05	92,23	69,04	85,18	91,73	75,33
Butcher meat	86,63	95,02	69,25	86,78	89,09	79,07
Fishmonger	64,63	82,16	43,00	82,89	82,84	85,71
Bookshop Stationery	94,65	98,56	73,00	93,74	96,37	77,09
Clothing store	93,28	98,20	65,31	92,97	94,37	78,99
Shoe store	93,28	97,99	59,88	93,35	94,18	78,87
Appliance Store	90,68	97,77	58,30	91,14	92,03	80,77
Furniture store	91,37	98,37	34,27	95,11	97,56	58,54
Drugstore	91,29	97,83	63,45	90,22	92,36	70,99
Cinema	96,03	99,36	29,03	97,25	98,67	42,42

Table D.5: Percentage of good answers for each kind of service in 2007 in France

Service	Percentage of good answers 2007 in France	Percentage of 0 catch 2007 in France	Percentage of 1 catch 2007 in France
Nursery school	78,28	78,12	78,82
Elementary school	79,67	48,56	97,96
Junior hight school	94,96	98,36	64,66
Health Center	96,10	98,11	44,53
Pharmacy	92,55	94,41	86,29
Ambulance	92,18	94,93	69,20
General Praticioner	90,79	94,48	81,25
Nurse	88,07	93,18	74,26
Physiotherapist	90,97	99,04	61,84
Dentist	93,82	97,17	78,70
Treasury	92,20	97,54	26,35
Job center	98,53	99,36	50,33
Police station	90,32	96,39	34,23
Post office	85,88	86,94	83,11
Deposite and savings banks	91,06	98,64	56,10
Automotiv repair	78,49	79,30	77,30
Mason	71,85	63,41	79,90
Plasterer painter	77,13	82,77	69,42
Carpenter	73,42	70,29	76,65
Plumber	74,67	87,36	60,83
Electrician	77,34	88,82	60,13
Hairdressing	84,46	98,12	60,38
Veterinary	93,27	99,37	42,06
Restaurant	62,06	26,27	95,72
Seniors: Accommodation	90,91	98,88	46,80
Groceries	73,25	66,78	91,19
Hypermarket	97,56	99,81	20,78
Supermarket	93,39	99,21	50,54
Large surface craft	94,23	99,26	28,36
Bakery	83,94	86,64	79,56
Butcher meat	84,50	85,76	80,37
Fishmonger	83,01	82,91	85,36
Bookshop Stationery	92,58	96,03	71,28
Clothing store	93,03	96,14	70,57
Shoe store	94,49	95,64	78,46
Appliance Store	90,85	92,59	75,21
Furniture store	93,59	97,54	51,95
Drugstore	89,63	92,13	67,97

APPENDIX E

List of Publications

International Journals

Works already published

- **Gargiulo, F., Lenormand, M., Huet, S. and Baqueiro Espinosa, O.** Commuting Network Model: Getting the Essentials. *Journal of Artificial Societies and Social Simulation* 2012, 15 (2) 6.

- **Lenormand, M., Huet, S. and Deffuant, G.** Deriving the Number of Jobs in Proximity Services from the Number of Inhabitants in French Rural Municipalities. *PLoS ONE* 2012, 7(7): e40001.

- **Lenormand, M., Huet, S., Gargiulo, F. and Deffuant, G.** A Universal Model of Commuting Networks. *PLoS ONE* 2012, 7(10): e45985.

Works under review

- **Lenormand, M., Huet, S. and Gargiulo, F.** Generating French Virtual Commuting Network at Municipality Level. *Journal of Transport and Land Use (arXiv:1109.6759v2).*

- **Lenormand, M., Jabot, F. and Deffuant, G.** Adaptive Approximate Bayesian Computation for Complex Models. *Computational Statistics (arXiv:1111.1308v3).*

- **Lenormand, M. and Deffuant, G.** Generating a Synthetic Population of Individuals in Households: Sample-Free vs Sample-Based Methods. *Journal of Artificial Societies and Social Simulation (arXiv:1208.6403v1).*

Book Chapters

- **Huet, S., Lenormand, M., Deffuant, G. and Gargiulo, F.** Parameterisation of individual working dynamics. *Accepted in Empirical Agent-Based Modeling: Parametrization Techniques in Social Simulations*, 2012, Chapter ??, 22 pages, A. Smagl and O. Barreteau eds, Springer.

International Conferences

- **Lenormand, M., Jabot, F. and Deffuant, G.** Adaptive approximate Bayesian computation for complex models. *The 8th World Congress in Probability and Statistics*, 2012, Istanbul, Turkey.
- **Lenormand, M., Jabot, F. and Deffuant, G.** Adaptive approximate Bayesian computation for complex models. *The 11th World Meeting of The International Society for Bayesian Analysis (ISBA2012)*, 2012, Kyoto, Japan.
- **Huet, S., Lenormand, M. and Deffuant, G.** Modelling a network of rural municipalities. *In Proceeding of the 7th Conference of The European Social Simulation Association (ESSA 2011)*, 2011, Montpellier, France.
- **Lenormand, M., Huet, S. and Gargiulo, F.** A commuting generation model requiring only aggregated data. *In Proceeding of the 7th Conference of The European Social Simulation Association (ESSA 2011)*, 2011, Montpellier, France.
- **Lenormand, M., Deffuant, G. and Huet, S.** Calibrating a complex social model. *European Conference on Complex Systems (ECCS'11)*, 2011, Vienna, Austria.
- **Lenormand, M., Gargiulo, F. and Huet, S.** From the Auvergne commuting network to every commuting network. *In Proceeding of the European Conference on Complex Systems (ECCS'10)*, 2010, Lisbon, Portugal.

Other Communications

- **Lenormand, M.** Génération et validation d'une population synthétique : Application à la région Auvergne. *Microsimulation sociale : théorie, applications, défis. Journée de travail LISC-naXys.* Université de Namur.
- **Lenormand, M.** Estimation de paramètres avec le calcul bayésien approché (ABC). *Journée Réseau Mexico, 22 et 23 novembre 2012*, Nantes Ifremer.

Technical Reports

- **Lenormand, M.** Predicting the presence and the number of jobs in different services in Auvergne municipalities. 2010, Technical report, IRSTEA.
- **Huet, S., Dumoulin, N., Deffuant, G., Gargiulo, F., Lenormand, M., Baqueiro Espinosa, O., and Ternès, S.** Micro-simulation model of municipality network in the Auvergne case study. 2010, Technical report, PRIMA Project, IRSTEA, LISC.