Quality based approach for updating geographic authoritative datasets from crowdsourced GPS traces

Stefan Ivanovic

UNIVERSITÉ PARIS-EST

IGN INSTITUT NATIONAL DE L'INFORMATION GÉOGRAPHIQUE ET FORESTIÈRE

COGIT

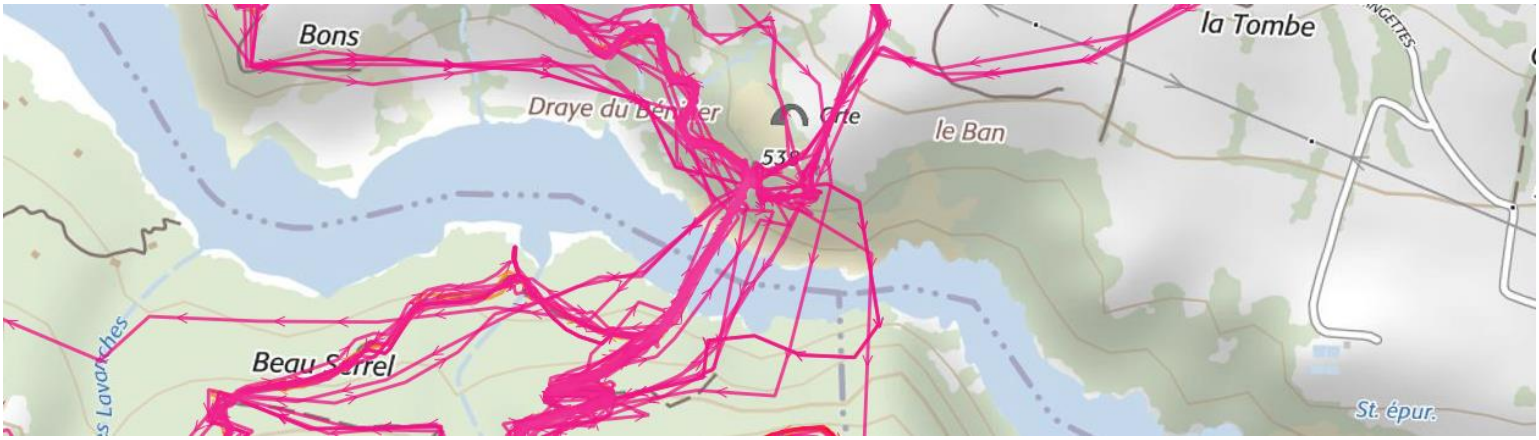
PhD thesis

submitted for the Degree of Doctor of Université Paris-Est

Geographic Information Science

Quality based approach for updating geographic authoritative datasets from crowdsourced GPS traces

Stefan S. Ivanović



Thesis defended publicly: 19.01.2018

Jury

Dr. Aldo Napoli	Chargé de recherche, MINES ParisTech	**Rapporteur**
Pr. Didier Josselin	Professeur, Université d'Avignon	**Rapporteur**
Pr. Alexis Comber	Professor, Leeds University	**Examinateur**
Pr. Karine Zeitouni	Professeur, Université de Versailles	**Examinateur**
Pr. Thomas Devogele	Professeur, Université de Tours	**Directeur de thèse**
Dr. Sébastien Mustière	HDR, IGN-ENSG	**Directeur de thèse**
Dr. Ana-Maria Olteanu Raimond	Chargée de recherche, IGN	**Encadrante**

mom ocu Slavku

to my farther Slavko

Acknowledgement

In the first place I would like to thank the directors of my PhD thesis, Sébastien Mustière who was also a director of COGIT laboratory, the laboratory where I conducted my PhD research, and Thomas Devogele, Professor of Université de Tours. Thank you very much both for giving me an opportunity to work on this very interesting and innovative subject under your supervision, as well as for all your guidelines, remarks and support you provided me, from the very beginning, till the very end of my PhD research.

I would like to thank members of my PhD defense jury for accepting our invitations and for very interesting and useful questions and remarks: the president and examiner Professor Karine Bennis Zeitouni, examiner Professor Lex Comber and especially to two reviewers Dr. Aldo Napoli and Professor Didier Josselin, who revised my manuscript and authorized me to defend this thesis.

Also, I would like to say a big MERCI, to Dr. Ana-Maria Olteanu Raimond, my supervisor, for not only supervising my work in a very professional way, but also for being here for me in all difficult moments I had during my PhD thesis, in my work as well as in my private life. Thank you also for making my manuscript much better than it was after a first version, and for the time spent on working in GéOxygene.

In addition, I would like to express my sincere gratitude to all my colleagues from IGN and COGIT for receiving me very nicely as a new team member in November 2014, and for all unforgettable moments that we shared together. Thank you all for giving me brief courses of French during my entire stay at COGIT. I would especially point out all nice moments spent with Imran, Bertrand, Abdel and Michael. Thank you Imran for listening my complaints on French food and bed Parisian weather all the time, and thank you Michael for debugging my GéOxygene codes many times!

Regarding this manuscript, a special thank is going to Bertrand Duminieu and Alexander Edwards for correcting the abstract and the introduction.

Finally, a very special gratitude is devoted to my parents Slavko and Zorica for making the fulfillment of this goal possible!

Abstract

Nowadays, the need for very up to date authoritative spatial data has significantly increased. Thus, to fulfil this need, a continuous update of authoritative spatial datasets is a necessity. This task has become highly demanding in both its technical and financial aspects. In terms of road network, there are three types of roads in particular which are particularly challenging for continuous update: footpath, tractor and bicycle road. They are challenging due to their intermittent nature (e.g. they appear and disappear very often) and various landscapes (e.g. forest, high mountains, seashore, etc.).

Simultaneously, GPS data voluntarily collected by the crowd is widely available in a large quantity. The number of people recording GPS data, such as GPS traces, has been steadily increasing, especially during sport and spare time activities. The traces are made openly available and popularized on social networks, blogs, sport and touristic associations' websites. However, their current use is limited to very basic metric analysis like total time of a trace, average speed, average elevation, etc. The main reasons for that are a high variation of spatial quality from a point to a point composing a trace as well as lack of protocols and metadata (e.g. precision of GPS device used).

The global context of our work is the use of GPS hiking and mountain bike traces collected by volunteers (VGI traces), to detect potential updates of footpaths, tractor and bicycle roads in authoritative datasets. Particular attention is paid on roads that exist in reality but are not represented in authoritative datasets (missing roads). The approach we propose consists of three phases.

The first phase consists of evaluation and improvement of VGI traces quality. The quality of traces was improved by filtering outlying points (machine learning based approach) and points that are a result of secondary human behaviour (activities out of main itinerary). Remained points are then evaluated in terms of their accuracy by classifying into low or high accurate (accuracy) points using rule based machine learning classification.

The second phase deals with detection of potential updates. For that purpose, a growing buffer data matching solution is proposed. The size of buffers is adapted to the results of GPS point's accuracy classification in order to handle the huge variations in VGI traces accuracy. As a result, parts of traces unmatched to authoritative road network are obtained and considered as candidates for missing roads.

Finally, in the third phase we propose a decision method where the "missing road" candidates should be accepted as updates or not. This decision method was made in multi-criteria process where potential missing roads are qualified according to their degree of confidence.

The approach was tested on multi-sourced VGI GPS traces from Vosges area. Missing roads in IGN authoritative database BDTopo® were successfully detected and proposed as potential updates.

Résumé

Ces dernières années, le besoin de données géographiques de référence a significativement augmenté. Pour y répondre, il est nécessaire de mettre jour continuellement les données de référence existantes. Cette tâche est coûteuse tant financièrement que techniquement. Pour ce qui concerne les réseaux routiers, trois types de voies sont particulièrement complexes à mettre à jour en continu : les chemins piétonniers, les chemins agricoles et les pistes cyclables. Cette complexité est due à leur nature intermittente (elles disparaissent et reparaissent régulièrement) et à l'hétérogénéité des terrains sur lesquels elles se situent (forêts, haute montagne, littoral, etc.).

En parallèle, le volume de données GPS produites par crowdsourcing et disponibles librement augmente fortement. Le nombre de citoyens enregistrant leurs positions, notamment leurs traces GPS, est en augmentation, particulièrement dans le contexte d'activités sportives. Ces traces sont rendues accessibles sur les réseaux sociaux, les blogs ou les sites d'associations touristiques. Cependant, leur usage actuel est limité à des mesures et analyses simples telles que la durée totale d'une trace, la vitesse ou l'élévation moyenne. Les raisons principales de ceci sont la forte variabilité de la précision planimétrique des points GPS ainsi que le manque de protocoles de collecte et de métadonnées (par ex. la précision du récepteur GPS).

Le contexte de ce travail est l'utilisation de traces GPS de randonnées pédestres ou à vélo, collectées par des volontaires, pour détecter des mises à jours potentielles de chemins piétonniers, de voies agricoles et de pistes cyclables dans des données de référence. Une attention particulière est portée aux voies existantes dans le monde réel mais absentes du référentiel. L'approche proposée se compose de trois étapes :

La première consiste à évaluer et améliorer la qualité des traces GPS acquises par la communauté. Cette qualité a été augmentée en filtrant (1) les points GPS aberrants à l'aide d'une approche basée sur l'apprentissage automatique et (2) les points GPS qui résultent d'une activité humaine secondaire (i.e. des micro-déplacements en dehors de l'itinéraire principal). Les points restants sont ensuite évalués en terme de précision planimétrique en utilisant la même méthode proposée pour la détection des points aberrants. .

La seconde étape permet de détecter de potentielles mises à jour. Pour cela, nous proposons une solution d'appariement par distance tampon croissante. Cette distance est adaptée à la précision planimétrique des points GPS classifiés pour prendre en compte l'hétérogénéité de la précision des traces GPS. Nous obtenons ainsi les parties des traces n'ayant pas été appariées au réseaux de voies des données de référence. Ces parties sont alors considérées comme de potentielles voies manquantes dans les données de référence.

Enfin, nous proposons dans la troisième étape une méthode de décision multicritère visant à accepter ou rejeter ces mises à jours possibles. Cette méthode attribue un degré de confiance à chaque voie manquante potentielle.

L'approche proposée dans ce travail a été évaluée sur un ensemble de trace GPS multi-sources acquises par crowdsourcing dans le massif des Vosges. Les voies manquantes dans les données de références IGN BDTOPO® ont été détectées avec succès et proposées comme mises à jour potentielles.

Table of content

Introduction

Introduction

A need for very up-to-date formal spatial data such as authoritative (e.g. National Mapping Agencies' data) or commercial data (e.g. Google maps) has increased significantly with the development of navigation and Web 2.0 technology. That has enabled citizens to be made use of as sensors. Wide availability of GNSS devices and navigation software in the last decade has enabled one to employ a significant part of the population using GNSS navigation, not only while traveling but also in their everyday life (e.g. cell phone GNSS navigation). At the same time analog maps have still remained a navigation tool for not a negligible part of the population. For both GNSS and analog map navigation, a complete reference road network is used as a basis, thus its up-to-datedness is essential.

Normally, authoritative data is updated permanently, mainly through stereo-restitution (3D mapping from areal and satellite images) and field surveys. This is a very expensive and technically complex process. Thus, a continuous update of authoritative spatial databases becomes a highly demanding task in both of its aspects, technical and financial. The road network is an especially challenging theme to keep updated due to frequent changes. Three types of roads are particularly challenging: footpath, tractor and bicycle road. Firstly, they are not of the highest priority for mapping agencies like motorways, thus fewer resources are invested in their update. However, they are still very important for navigational purposes, especially in mountainous areas, to keep connectivity of the road network, for touristic maps, defense purposes, etc. Secondly, they have an intermittent nature, in that they appear and disappear very frequently. Sometimes even during one season (especially during autumn and spring). Due to their coverage (generally unpaved), they can easily appear and be destroyed. In addition, these are the road types with the most frequent changes in geometry. Thirdly, it is difficult to collect and classify them in stereo restitution since they are narrower than the other road types and less visible because they are unpaved. Using the stereo restitution technique is even more challenging in mountainous area where dense vegetation limits their visibility, which can lead to complete failure in the identification of roads altogether. As a result, footpath, tractor and bicycle roads are the least up-to-date road types in authoritative datasets, and are in need of an alternative update solution.

Concurrently, with the improvement of data collection technologies and the development of Web 2.0, the amount of data collected has increased rapidly in recent times as well as the number of citizens involved in data collection activities. Besides formal data generated by professionals within institutions and companies, the data generated by the 'crowd' appeared. Various terms have been used in defining this data, Crowdsourcing (Howe, 2006), Neo-Geography (Turner, 2006), Volunteer Geographic Information – VGI (Goodchild, 2007) or user generated spatial content (Antoniou et al., 2009). The type of content generated by the crowd is diverse from spatial (Open Street Map project-OSM) to descriptive data (Wikipedia). Social networks with their growing popularity and use are particularly promoting VGI data collection. They became a kind of platform for sharing VGI. In particular, intensive sports activities of professionals and amateurs are very frequently recorded using GNSS devices. The amount of collected data has increased gradually in recent years. The traces obtained are then made openly available to the community through social networks (e.g. Facebook), blogs, sport and touristic communities' web-sites. Traces are recorded using a wide range of GNSS devices, from very low to high class receivers. However, most of them are collected by medium-low class receivers integrated in smartphones as many Android and IOS applications exist for GNSS traces

recording like Endomondo[1], Runkeeper[2], StravaGPS[3], etc. A vast majority of applications use GPS data besides other GNSS systems that are less supported, e.g. Glonass or Galileo, thus in this work we will consider GNSS traces as GPS traces.

One the one hand, GPS traces collected voluntary (VGI traces) are collected without any protocol, with low and heterogeneous frequency sampling. In addition, they are made available with few or inexistent metadata. Moreover, various errors are introduced by different external factors such as topography, canopy, etc. These errors can cause a significant bias and may limit the usability of the traces in different analyses. Currently, are mostly used for visualization purposes and some basic metric data analysis (e.g. total time of the trace, distance, speed) (see Figure 1).

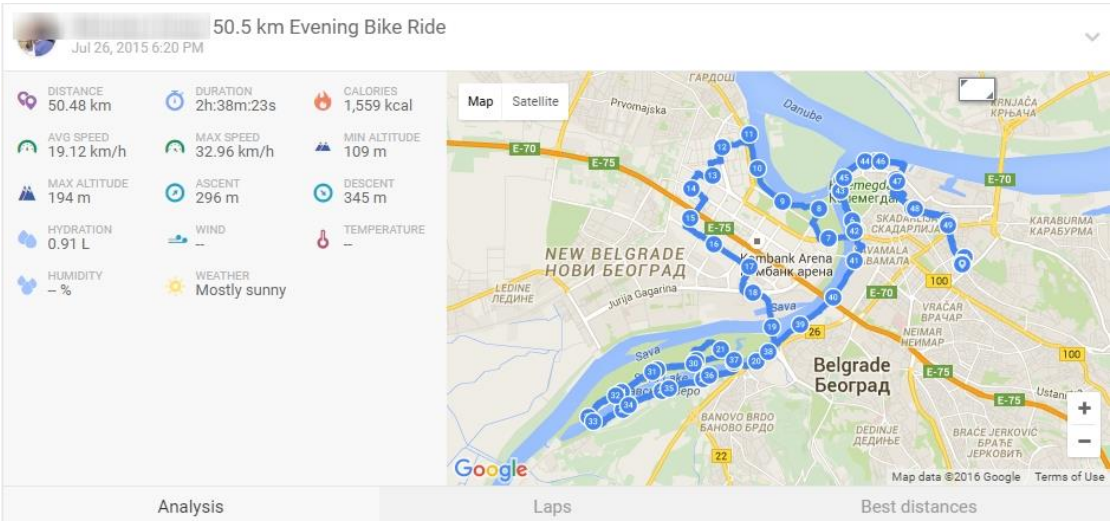


Figure 1 : VGI bicycle trace indicating basic matric data

On the other hand, VGI traces have a potential to be used for more advanced purposes such as bicycle routing (Bergman and Oksanen, 2016). The potential is based on three important properties of VGI traces. Firstly, they are available in a huge amount that has been increasing gradually. Secondly, their data is continuously collected, which puts their actuality at high rate. Thirdly, even if sometimes imprecise, they still provide indications of activities in the areas they are collected.

The main focus of this work is exploring and evaluating possibilities of using VGI traces for detection of updates in authoritative datasets. More specifically in this work we present an approach for updating road types which are challenging to update using standard methods (footpath, tractor and bicycle road). Figure 2a illustrates a typical situation where a road (here, a small footpath) is missing in the authoritative dataset, whereas it exists in reality, illustrated in Figure 2b and c.

[1] www.endomondo.com
[2] www.runkeeper.com
[3] www.strava.com

a) b) c)

Figure 2 : An example of a missing road in an authoritative dataset: a) A VGI trace collected along the missing road; b) The path A in the real world; c) The path B in the real world

As it can be seen, a VGI trace collected by the author of this thesis in Bois de Vincennes (France) using two paths (A and B) exists in the reality but not in the authoritative road network of the French Mapping Agency (IGN France). Due to the photos taken on the spot (Figure 2b and c, it is evident that paths A and B exist, thus could be an indication of missing roads in the authoritative dataset. The path B is more visible in the aerial photo (Figure 2a thus can be collected by stereo-restitution, however the path A is not visible due to the very dense canopy cover. Here, VGI traces are particularly useful since they can overcome the limits of stereo-restitution. When considering the path B as a potential update of the authoritative road network, the question of complying with authoritative data specification arose. As it can be seen in the Figure 2 b, the path is very narrow and not very persistent, thus the decision if it should be introduced in the authoritative road network has to be made according to the data specification. By visual checking of the aerial photo, we can observe that the quality of the VGI trace is mainly satisfying with no major deviations from the path followed while collecting the trace. Good quality traces provide a more reliable indication of missing roads, this means particular care must be addressed to considering the quality of traces.

In our work we are relying on the growing amount of VGI traces shared by 'crowd' in the context of their sports activities, like running, hiking or biking. Those sport activities are usually done using road footpaths, bicycle and tractor roads, thus such VGI traces seem to be suitable data source updating those road types. Those traces may highlight missing roads in authoritative data, or may confirm that some roads are still in use. However, this is a complex issue. In related goals, like updating major road networks, the limited quality of VGI may be counterbalanced by the large amount of available data. In our context, the data is rich enough to detect clues of updates, but only few VGI traces exist for the same path. Knowing and improving the quality of each trace is then important.

Therefore, in order to examine the possibilities of using VGI traces in detecting updates in authoritative road network, the research questions we will address are: How to evaluate and improve the quality of VGI traces having missing metadata? How to match VGI traces having heterogeneous quality to authoritative road network in order to identify differences between datasets (potential updates)? Finally, how to make a final decision on which updates should be

proposed to integrate into an authoritative dataset? In order to address presented research questions, we propose a global approach in three steps (see Figure 3).

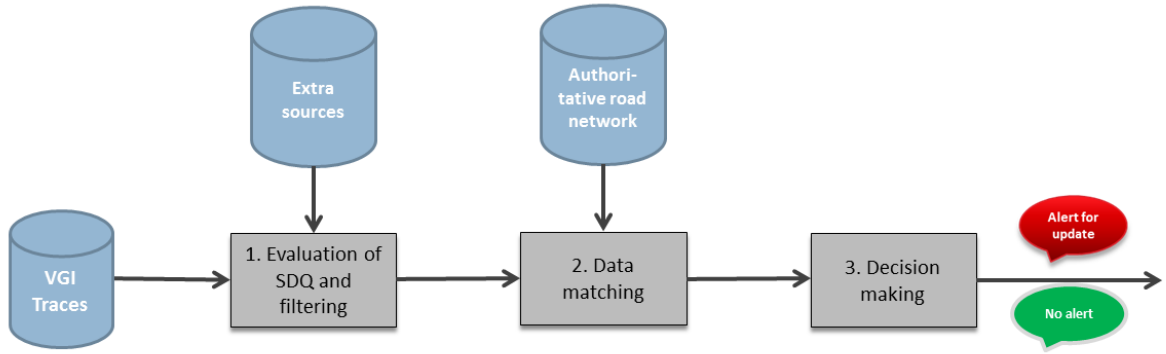


Figure 3 : Global approach for detection of updates in an authoritative dataset

As it can be seen, the approach starts from VGI traces which first have to be evaluated and improved in terms of their quality (spatial data quality - SDQ). Further, data matching between evaluated VGI traces and an authoritative road network is done so that differences (potential updates) between those two datasets are found. Finally, each potential update is examined and a decision is made if it is a real update (alert for update) or not (no alert).

Organization of the thesis

Following the three steps of the approach illustrated in the Figure 3, the thesis is organized in three chapters as follows.

Chapter 1 deals with SDQ. It reviews the background of evaluation of spatial data quality and highlights its importance. Special attention is paid to GPS data quality issues (sources of errors and performance in different environments). The chapter then addresses the problem of quality assessment of heterogeneous VGI traces with significant lack of metadata and redundancy. It fully describes our proposed approach for evaluation of spatial data quality of VGI traces composed by three methods, two intrinsic: detection of outlying GPS points, and secondary human behaviour and one extrinsic: detection of low accurate GPS points. Finally, it presents the results of VGI traces quality improvement and evaluation.

Chapter 2 is devoted to data matching, focusing on its use in detection of updates. First, a relevant state of the art of data matching measures and methods is presented. It investigates not only the relevant research works, but also the possibilities of applying some of the existing matching approaches in matching VGI traces and authoritative road network so that the potential updates can be found. The chapter describes then our proposed approach for detection of updates of authoritative road network by using quality evaluated VGI traces. The results and analysis of application of the proposed method in the test are illustrated and described.

Chapter 3 concentrates on final decision making if a detected update is a real one or a false one and to which extent we can trust the proposed updates. First a measure for qualifying the relevance of

updates is proposed. Then, relevant criteria for calculation of the measure are defined and combined. Finally, each update is qualified according to its relevance, and the decision regarding which updates will be proposed is made. The updates are then analysed and discussed.

Finally, we conclude by summarizing the achievements of the thesis and recalling the main results. In addition, the conclusion presents the perspectives and possible future work, especially new ideas which could lead to more accurate results.

Chapter 1

Evaluation of spatial data quality

1 Evaluation of spatial data quality

1.1 Introduction

With the development of Web 2.0 techniques, citizens are able to act like sensors and produce spatial data. Bruns (2008) use the term of “producer” for describing citizens producing spatial data. In the Geographic Information Science domain many terms are used to define this type of activity: neogeography (Turner, 2006), volunteered geographic information (Goodchild, 2007), user generated content (Krumm, 2008), user generated spatial content (Antoniu, 2009), crowdsourcing (Howe, 2006), etc. A detailed review on terms on the current terminologies is made in See et al., (2015).

Citizen contributed data is generally collected by volunteers. In some cases, there are no protocols and guidelines, such as in mountain activity. In some other cases, when the volunteered activity is carried out inside a VGI project such as OpenStreetMap, WikiMapia, Geo-Wiki or Degree confluence Project protocols exist but they are not always strict and detailed or not always followed by volunteers. Volunteers are not always skilled especially for collecting spatial data and have usually different backgrounds. That is why VGI data is very heterogeneous regarding many aspects e.g. granularity or values of attributes. As a result, four main issues of this data can be distinguished: high heterogeneity, incompleteness, low accuracy, no metadata regarding the process of collection. Data quality issues are the main barrier in using VGI data for advanced purposes.

In addition, in case of collection of spatial data, the equipment used is usually low budget and subsequently low class equipment e.g. GPS devices integrated in mobile phones. This introduces a significant uncertainty in data quality. Furthermore, a huge variety of equipment and methods used in data collection causes a very random distribution of quality.

VGI data quality assessment has been studied in the last decade. In this chapter we are focusing on spatial data quality (SDQ) because it is an important aspect of our work. Since the quality of analysis depends on methods applied and the quality of input data, using VGI data especially in complex analyses requires information about its quality a priori. However, VGI data quality assessment is a challenging task from many reasons. First of all, VGI is relatively a recent research field, coined by Goodchild in 2007 as VGI and by Howe in 2008 as Crowdsourcing, thus many aspects of it are still in defining such as quality aspect.

For assessing VGI data quality, few general frameworks exist up to now (Barron et al., 2014; Leibovici et al., 2015; Bordogna et al., 2016; Ballatore and Zipf 2015). Partial methods were developed mainly for evaluating OSM data (Haklay et al., 2010; Girres and Touya, 2010; Fan et al., 2014), whereas some VGI data sources have not been particularly analysed for the quality, such as VGI traces coming from sports activities. Since they are crucial input data in our research, special attention will be dedicated to the assessment of VGI GPS data quality in this chapter.

The goal of this chapter is to describe methods and issues of spatial data quality assessment. First, relevant state of the art on assessing spatial data quality is presented in Section 1.2. A practical study of VGI traces completeness described in Section 1.3. Finally, Section 1.4 describes the approach we proposed for assessing spatial data quality of VGI traces, and the obtained results.

1.2 State of the art of spatial VGI data quality assessment

1.2.1 Elements of spatial data quality

Among several standards for spatial data quality, ISO 19157 (latest version 2013) is most widely used and accepted one. The standards are mainly used for documenting the extrinsic evaluation of data quality, i.e. when data quality elements are evaluated by comparing the data to 'ground truth', that is, to referential dataset. Quality elements proposed by ISO 19157 are listed below:

- Positional accuracy represents the accuracy feature's position compared to its true position. It actually represents how close or far a feature is from its true location. Positioning measurements have always a deviation from a true position. The question is to which extend the measured position deviates from the true one.

- Thematic accuracy represents the accuracy of attributes that are associated to geographical features. It is for example used in evaluation of Land Cover data by verifying if the class associated to a Land Cover feature is in accordance with ground truth (e.g. check if a type of forest declared for a Land Cover polygon is correct, compared to the reality).

- Completeness is a measure of presence or absence of features and their associated attributes. According to the ISO 19157:2013 elements of completeness are commission and omission. First considers the presence of extra data and its attributes in a dataset, whereas second considers an absence of data or its attributes in the dataset.

- Temporal quality is related to the quality of data temporal attributes and temporal relations between features. For spatial data, temporal quality is mainly related to the evaluation of its up-to-dateness. In this case, the more recent the data is, the better its temporal quality is.

- Logical consistency refers to the evaluation of consistency of relationships between features in a dataset, as defined by constraints on the schema or data. Logical contradictions within a dataset are in focus of evaluation of logical consistency.

- Usability addresses the quality of a dataset from final user's point of view – fitness for use. In other words, to which extend a dataset fulfills the needs of final users.

In the work of (Antoniou and Skopeliti, 2015) quality elements evaluated by extrinsic approach are recognized as Measures for VGI data quality, while elements evaluated by intrinsic approach are recognized as Indicators for VGI data quality.

The ISO quality standards related to spatial data were mostly defined to describe the evaluation of official spatial datasets. However, many researches have been made in the last decades when ISO data quality elements were used to assess the quality of VGI data by comparing it to an authoritative data. Most of the works are focused on studying OpenStreetMap (OSM) data quality by measuring positional accuracy (Haklay et al., 2010; Fan et al., 2014; Girres and Touya, 2010), thematic accuracy (Girres and Touya, 2010; Jokar Arsanjani et al., 2015), completeness (Girres and Touya, 2010; Neis et al., 2011), temporality and evolution (Antoniou et al., 2015) or logical consistency (Touya and Brando, 2013; Hashemi and Rahim, 2015). For a complete review of VGI data quality assessment

methods, see (Jokar Arsanjani et al., 2015 Antoniou and Skopeliti 2015; Senaratne et al., 2015; Fonte et al., 2015).

However, knowing the specificity of VGI data, these indicators are not sufficient for assessing VGI data quality. Thus, the quality elements specifically dedicated to VGI have been proposed in particular. Some of them are: local knowledge defined as familiarity of the user to an area (van Exel et al., 2010), reliability (Comber et al., 2013), expertise (Flanagin and Metzger, 2008), etc.

As far as we know, there have not been studies particularly dealing with evaluation of VGI traces quality. In a close context, evaluations have been done on GPS traces recorded in animals tracking (Janeau et al., 2004) as well as on GPS traces as a part of OSM data (Girres and Touya, 2010). However, quality evaluation of OSM data have been conducted by globally applying same approaches on the OSM data regardless of data source, e.g. same approaches have been applied on roads obtained by vectorization of satellite images and on those obtained by GPS. Thus, the specificity of GPS data has not been particular taken into account.

1.2.2 Extrinsic approaches

Evaluation of data quality up to now follows most of the time an extrinsic approach: data are compared to referential data considered as 'ground truth'. The uncertainty of the evaluation is in then a combination of the uncertainty of referential data, which is usually small, and the uncertainty of computational operations done for the comparison. For example, using Hausdorff distance (Hausdorff, 1914) to measure deviation of an examined line segment to referential one instead of using Euclidian distance can make differences in results.

There exist many works for assessing data quality through extrinsic approaches, more and more works dealing with VGI data quality (Girres and Touya, 2010; Al-Bakri and Fairbairn, 2010, Haklay, 2010). An important milestone for those works was a start and rapid development of OSM project. OSM has been providing a wide range of spatial data in the last decade worldwide. Usages are various, from navigation (OSMTracker), crises management (National Geographic[4], 2012) to mapping remote and already not mapped areas in developing countries (Dar Ramani Huria[5]). Since OSM is a rapidly growing project, massively used nowadays, the question of its quality arose. One of the very first works that tackled this problem was described in Haklay (2010). The positional accuracy of OSM dataset for Great Britain (road network and urban areas) was examined by comparing to Ordnance Survey dataset. Linear features were examined by a method proposed by Goodchild and Hunter (1997) that uses buffers to determine the percentage of a line from the dataset which has to be evaluated that is within a certain distance of the same feature in dataset of higher accuracy. The method is illustrated in Figure 4:

[4] "How Crisis Mapping Saved Lives in Haiti", retrieved April 2017, from www.nationalgeographic.com,
[5] retrieved April 2017, from www.ramanihuria.org

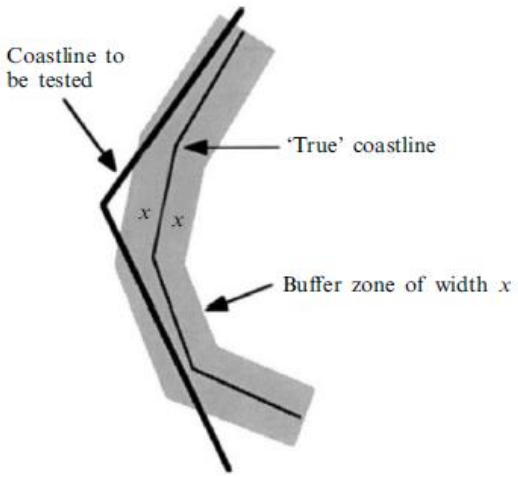


Figure 4 : Buffer zone method principle (by Goodchild and Hunter 1997)

Using buffers to measure deviation of linear features from referential data is also present in the work of Al-Bakri and Fairbairn, (2010).

Similarly to this work, Girres and Touya (2010) tested the quality of French OSM data comparing OSM data to French referential database BDTopo® where Hausdorff and average distances have been used to evaluate the positional accuracy of linear primitives in OSM. OSM road features were first matched with corresponding BDTopo® roads by using the matching algorithm proposed by Mustière and Devogele (2008).

Evaluation of road network positional accuracy was also done by comparing junctions (Helbich et al., 2012). The accuracy was assessed by calculating discrepancies between coordinates of OSM roads junctions and coordinates of referential roads junctions as illustrated in Figure 5.

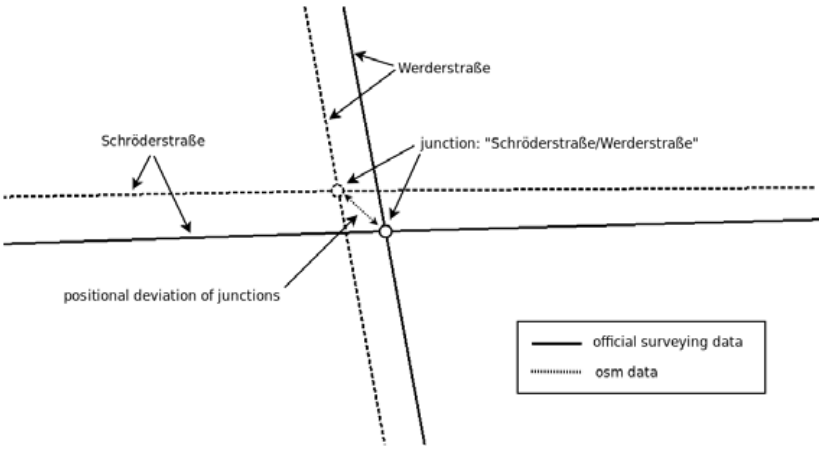


Figure 5 : Comparison of OSM and referencial juctions in road network (by Helbich et al., 2012)

Comparison of OSM spatial completeness was conducted by Neis et al., (2013). The study is done over 12 different urban areas in the world. Features from OSM urban areas were compared to the urban areas extracted from Bing satellite data (Bing Maps). Significant differences in completeness were spotted (i.e. different amount of contributions). Urban areas in European cities proved to have higher level of OSM data completeness compared to the other urban areas in the world (e.g. Cairo, Sydney, Los Angeles, Buenos Aires, etc.). Surpassingly, a correlation between the density population and number of contributors was not found strong.

Among other works, completeness of linear VGI data was evaluated by means of referential data in works of Haklay (2010), Ciepluch et al., (2011), Koukoletsos et al., (2012) and temporal quality was

assessed by Girres and Touya (2010) by calculating the number of objects uploaded in OSM in specified time intervals.

Overall, most of presented works concerned evaluation of spatial accuracy and completeness that are also in our focus. The closest application to our needs is an application of OSM road network data that can come from different sources, not only from GPS. Thus, the presented methods are not completely adapted to our data, since there was no study carried out specifically on GPS traces. Furthermore, absence of referential data in case of detection of updates in referential dataset which is the main goal of our work, makes evaluation of positional accuracy comparing to the 'ground truth' impossible.

1.2.3 Intrinsic approaches

It is not always possible to use referential data in evaluation of quality that is applying extrinsic approach due to the lack of available referential data in some areas. In those conditions, approaches relying only on data itself and its metadata need to be developed. Compared to extrinsic, intrinsic data quality evaluation is limited with respect to quality aspects that can be examined as well as to the reliability of obtained results. Insufficient amount of data to compare to itself or a lack of its metadata can sometimes significantly limit full application of intrinsic approach. In addition, results of extrinsic quality assessment are more reliable since the data are compared to something which is considered ground truth, whereas in intrinsic assessment data is evaluated only by itself. Works on this topic are not numerous, such as (Bishr and Kuhn, 2007, Keßler et al., 2011, Neis et al., 2011, Mooney and Corcoran, 2012, Barron et al., 2014). However, they address a fundamental part of intrinsic data quality assessment.

For example, Keßler et al., (2011) deal with intrinsic assessment of linage of OSM data, especially paying attention on evaluation of contributor's profile and activities while contributing to OSM. For that purpose, an OSM provenance vocabulary that makes implicit features linage information explicit was introduced. Some of the proposed classes of the vocabulary are: Edit, FeatureState, User, Changesets etc. The classes are connected so that "Edit is the central class that links changes on a specific FeatureState (i.e. a specific version of a node or way) to the User who made the change and to the corresponding timestamp."

Mooney and Corcoran (2012) did a comprehensive analysis of annotation process in OSM. They particularly focused on "heavily edited objects" i.e. objects with more than 15 edits. These cases were examined for exploring the uncertainty between contributors tagging road names, (e.g. tagging a highway as a double carriage way). The experiment was conducted for 4 road datasets from UK, Germany, Austria and Ireland. Global results showed very small percentage of roads with only one single or two different tags for all countries, except for Germany where 35.01% of roads were tagged with only one or two different tags. Most of the roads in all studied countries have between 3 and 6 different tags like primary, tertiary, residential etc. The annotation uncertainty was particular evaluated in case of highways showing that a non-negligible percentage of highways (23%) have different tags for the attribute 'name' during edit version. This uncertainty is important since the type of road such as highway is not difficult to differentiate among other road types, however it is also possible that a majority of errors were coming from spelling mistakes.

An intrinsic approach for the evaluation of precision of OSM road data was proposed by Barron et al. (2014). The approach is based on method of Helbich et al., (2012) but adapted to use data itself. Again, the road junctions were used in comparison, but not those coming from referential dataset. The global idea is to compare the position of a current junction to its previous position – obtained in previous update. The process of evaluation includes the analysis of angle and distance between two road junctions that are different updates of the same feature, within polar scatter plot.

Besides the evaluation of particular quality aspects of OSM e.g. positional accuracy, one of the first comprehensive frameworks for intrinsic VGI quality assessment was proposed by Barron et al. (2014). The framework named iOSMAnalyzer proposed more than 25 methods and quality indicators designed to assess data fitness for purpose. These indicators are grouped by purposes into seven groups as illustrated in the Figure 6.

Figure 6 : iOSMAnalyzer's intrinsic Quality Indicators (Barron et al., 2014)

Based on presented literature review, it can be concluded that the comprehensive evaluation of VGI datasets intrinsically is at the beginning of its development. Solutions proposed have addressed quality indicators that are only partly in our focus: completeness, temporal quality, spatial accuracy. Most intrinsic approaches rely on historical data and compare current data to its previous updates. They are efficient and convenient in evaluation of OSM data as it was presented, however it would be difficult to evaluate our data in this way. OSM is a wide community supplied every day with a dozens of updates especially in road network, thus most of the roads has a sufficient number of updates for comparison. At the opposite, our dataset has significantly less data, which limits the application of solutions proposed for OSM. Also, evaluation of accuracy by comparing features to their previous updates can be biased for two reasons. First the data are compared to their previous updates, not to the ground truth. Second, data in OSM is coming from few very different sources like: satellite, aerial images, GPS etc. Comparing data with different and unknown accuracy (road recorded by unknown GPS device and same road vectored from unknown satellite image) can give biased and not reliable results. Two VGI roads can be very close to each other even if their real (ground truth) accuracy is very low. In this case evaluated accuracy would be falsely high.

Comparing current valid features to their previous updates through the time for the purposes of assessing positional accuracy as presented by Barron et al., (2014) seem to be problematic. Positional accuracy is strictly related to the ground truth values. Comparing different updates of the same

feature in OSM data theoretically can only provide some conclusions regarding precision, but not an accuracy since the ground true position of the feature is unknown.

Knowing the fact that we have a lack of history of our data (updates of VGI traces for our test area through the time) for application of presented solutions, we are supposed to propose a new approach for its quality evaluation. Therefore, the first goal of this research is to propose a method for VGI traces quality evaluation, which will mostly rely on evaluation of accuracy of the traces and external factors that influence GPS measurements in specific environments. The traces are supposed to be evaluated independently, to avoid discussed inconsistences of results in case of mixing traces having various accuracy.

1.2.4 Elements of spatial data quality of VGI traces

In the previous section, without being exhaustive, some methods proposed to assess the quality of both formal and VGI spatial data. This section is focused on quality assessment for VGI data collected by GPS devices. Before describing types of errors of VGI traces we present an evaluation of different influences on the quality of GPS measurements. What should be stressed is that tests performed and described in the existing scientific papers had been done under a variety of different conditions. Thus, comparing the influence of each factor is not straightforward.

1.2.4.1 Sources of errors of GPS traces

Like any sensor measurement, GPS measurements are also burdened with different errors and inconsistencies. There is no unique and widely accepted nomenclature of GPS sources of errors. Nevertheless, few slightly different nomenclatures exist. There are two main nomenclatures: those provided by leading manufacturers such as Trimble and Leica[6], and those provided by academic society and institutions e.g. University of Newcastle-Geodesy group[7], from www.ncl.ac.uk/gps. Basically, all of them identify following sources of errors in common:

- Ionospheres and tropospheric effects
- Errors of ephemerides
- Errors of satellite and receiver clocks
- Geometry of satellites (dilatation of precision - DOP)
- Reflection of signal (multipath)

Apart from these sources of GPS errors that are mutual for all GPS devices, there are a plenty of additional factors that affect GPS measurements in specific environments or specific data capture conditions. From our point of view, the factors can be organized and grouped in the following way:

Topography
Many authors identify topography as a significant factor which affects GPS measurements (Klimànek, 2010; Tucek and Ligos, 2002; Cain et al., 2005; Lewis et al., 2007; DeCesare et al., 2005). Some of them consider it in general, whereas others consider most important features of topography independently, mainly inclination and orientation of relief slopes. Generally speaking, topography influences GPS measurements by reducing the area of visible sky, as well as making signal reflections, which both reduce positional accuracy and precision of GPS measurements considerably.

[6] retrieved October 2016, from www.trimble.com and www.leica-geosystems.com
[7] retrieved October 2016

Canopy cover

Canopy cover was recognized by various authors as the most influencing factor on GPS working process (Janeau et al., 2004; DeCesare et al., 2005; Lewis et al., 2007). Depending on characteristics of canopy cover, number of visible satellites and quality of signal vary (signal is completely blocked or its strength is reduced). Because of that, as well as its complexity, effects of canopy cover could be more fragmented according some characteristics of the cover such as: percentage of coverage, type of trees (species), leaf OFF/leaf ON period and height of trees.

Time intervals

Time interval is another element identified and studied in the in the literature (Cain et al., 2005; Janeau et al., 2004; Klimànek, 2010) and is related to GPS data collection. Two factors were identified: frequency and measurement time. Frequency is related to time between the acquisitions of two consecutive positions, whereas measurement time represents total time of data acquisition during one session.

Frequency can potentially affect GPS measurement by reducing the number of acquired points. The sensibility of success rate of acquired points is very high in case of low frequency sampling done in environments unfavorable for GPS signal such as deep forest. On the other hand, measurement time can affect the quality of measurements due to prospective decrease of device performance in case of longer period of constant exploitation.

Number of visible satellites

Number of visible satellites during measurement session proved to be highly important for the quality of GPS measurements (number of positions acquired within scheduled time intervals - fix rates and the precision and accuracy of). As we know from the basics of GPS working principles, the more the satellites are visible the better the positioning is. Additionally, the number of satellites determinates if the position is collected in 2D or 3D. That is to say three visible satellites ensure 2D positioning, whereas for 3D positioning 4 satellites are required. This was also confirmed in work of DeCesare et al., (2005).

Type of GPS device

This parameter refers to construction characteristic of GPS device. One of the most important is if the GPS device is created as non-differential (i.e. having no position corrections from base station) or differential (i.e. having position corrections from base station). According to product specifications differential GPS have far more advanced performances than non-differential. Those advantages provide both, higher spatial quality and number of positions acquired especially in demanding environmental conditions (Janeau et al., 2004). However, most GPS devices used by non-professional are non-differential.

Smartphone GPS is widely used nowadays. The amount of data recorded by such type of GPS has been increasing gradually. A majority of data in our study are coming from smartphone GPS devices. Thus, when speaking about different type of GPS devices, it is important to point out main differences between professional GPS device and smartphone one. Here also, most important differences are caused by different constructions. First, the quality of antenna in professional device is better and its size and position are better determined than in case of smartphone GPS. Especially important fact here is that professional GPS even low quality has an antenna with far better multipath suppression than smartphone GPS. Second, GNSS chipsets integrated in professional device is higher quality than those integrated in smartphones. All of this has enabled professional device to reach positional accuracy on the level of centimeters, whereas smartphone devices are still on the meter's level.

Position of receiver

We recognized receiver positon in the space during signal acquisition as prospective influencing parameter on receiver's performance. We identified two sub-factors: how the GPS is hold and how it is oriented. First sub-factor, holding GPS, represents the position of receiver in relation to receiver's holder such as in the pocket, on the arms –collar GPS, in the hands, (Blunck et al., 2011). Second, the orientation of GPS devices refers to how the device is oriented toward cardinal directions, in the moment of data acquisition.

Multipath (signal reflections)

Multipath effect is provoked by the noise of signal due to multi reflection that happened on its way from satellite to receiver. The obstructions make signal reflect and as a result cause a delay on signal arrival time, which finally induce significant errors of GPS coordinates (Tucek and Ligos, 2002). Multipath effect is especially strong in urban canyons where high-rise buildings reduce significantly the part of visible sky needed for good GPS signal reception. Reflections of signal are very numerous, thus the accuracy of GPS positioning is subsequently very poor and randomly distributed in the space (Ben-Moshe et al., 2011; Xie and Petovello, 2014). Effects of urban canyons on GPS positioning can be observed in Figure 7 that shows traces collected by GPS in red, while the real roads are presented in blue. Huge discrepancies between them can be very easily observed.

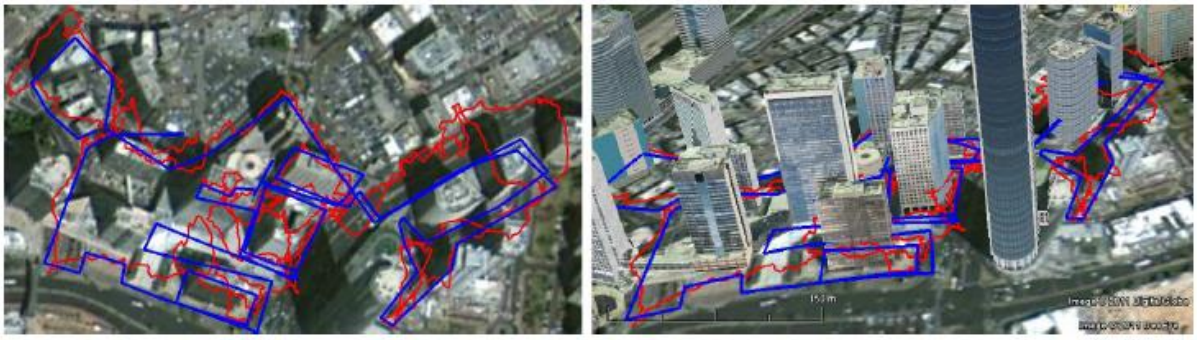


Figure 7 : Urban canyon effect (by Moche et al., 2011)

Conversion

Since GPS coordinates are calculated in WGS 84 ellipsoid reference system, further conversion for their use together with data expressed in another reference system is required. This process brings some degradation of coordinate's accuracy depending on how it is conducted.

1.2.4.2 Types of errors of GPS traces

In this section we are discussing the types of errors caused by sources of errors identified in the previous sections.

Positional error (error of X and Y coordinates) in the literature mostly refers to errors of position in 2D space (X, Y). Due to that fact, we have decided to use this term in the same way.

Error of altitude refers to error of Z coordinate. Since positional error is related to 2D position, and especially since method of calculation of Z coordinate in GPS measurements is different than method for planar coordinates, we also decided to consider error of third coordinate separately.

Missing position and missing 3D position are also important errors in GPS data. Depending on number of visible satellites, GPS positions acquired can be 2D (by means of 3 visible. Hence we differentiate two types of measurement errors: missing 2D position and missing 3D position.

Based on software and/or application used for GPS data collection, timestamps are usually collected for each point. However, during the observation of GPS traces coming from different VGI platforms, we noticed that timestamps are missing very frequently, sometimes completely (for all points of one trace), sometimes just for particular points.

It should be stressed that most of listed errors are significantly more frequent at low cost devices (mostly used in VGI data collection), than at professional GPS devices. Differences between professional and low cost GPS receivers (e.g. smartphone GPS) are very sharp, not only in the construction characteristics and price, but also in the performance and subsequently in the quality of data produced. Owing to the better construction of the antenna, professional receiver has better signal reception, thus the missing positions are less frequent than in case of low cost one. Most, importantly, professional receiver is more dominant in positional accuracy. Test done by very well-known GPS manufacturer Trimble, gave an insight into this difference. Test was carried out on Garmin Map 76 recreational receiver and Trimble GeoXT in the same conditions and during 80min. Figure 8 shows the deviation from ground truth positions for both receivers.

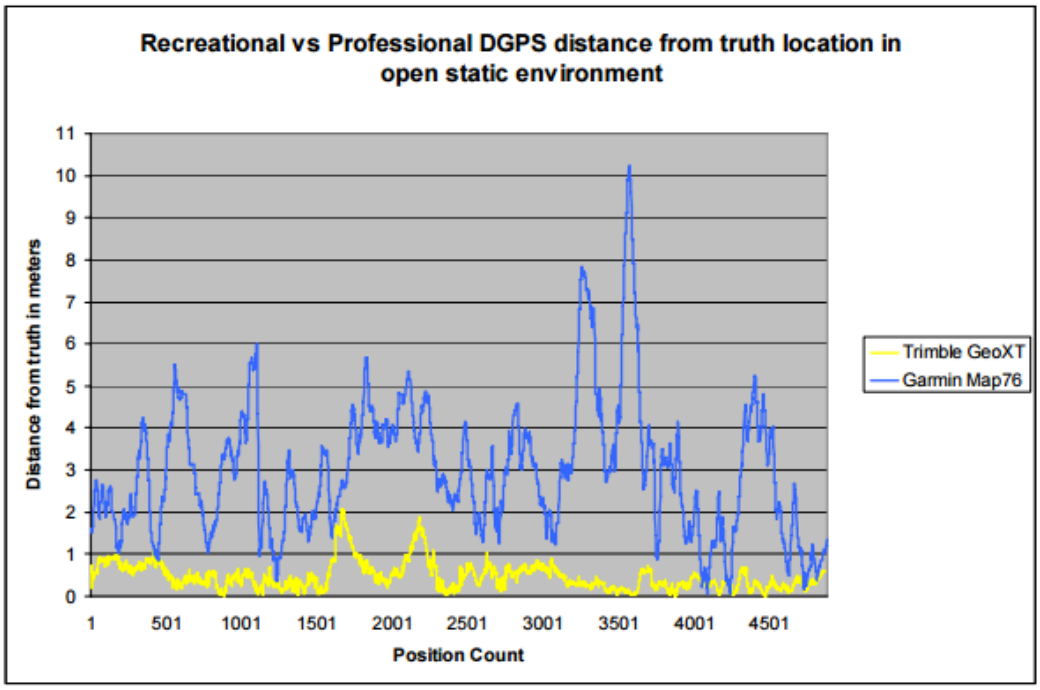


Figure 8 : Low cost versus professional GPS receiver accuracy (Trimble White paper, 2006)

Here, it is obvious that professional receiver outperformed the low cost one, and that the difference in positional accuracy is considerable. Therefore, in terms of the quality, VGI GPS data should be taken with significantly more reservation than data collected by professional GPS receivers.

1.2.5 How much the sources of errors influence the quality of VGI traces?

In this section, we aim to evaluate the quality of VGI traces through existing studies. First, the methodology used in considered studies is described. After, a detailed overview of the influence of each source of errors on GPS measurement results will be presented and discussed.

1.2.5.1 Methodologies

Our study is based on papers which dealt with factors that influence GPS performances in different environment. Some related papers are principally devoted to analysis of animal tracking and

behavior by means of GPS data (Janeau et al., 2004; Frair, 2010; Allen et al., 2014). However, they also possess tests and analysis of dependence of GPS performances on various environmental conditions. Other papers, which have environmental influences on GPS performances in main focus, mostly take into account influences of canopy cover (Lewis et al., 2007; Tucek and Ligos, 2002; DeCesare et al., 2005), topography (terrain) (Lewis et al., 2007; Tucek and Ligos, 2002; Cain et al., 2005), and measurement time (Klimànek, 2010; Cain et al., 2005), whereas effects of stand age, receiver construction, specific tree species, presence of leaf and snow etc., are less treated.

When it comes to technical details of the studies, they were carried out using different models of GPS devices in static or dynamic mode. Frequency of devices ranged from 2s (DeCesare et al., 2005) to 180min (Janeau et al., 2004), whereas measurement time varied from 1min (Klimànek, 2010) to 24h (Lewis et al., 2007).

Environmental conditions were also various, generally and in each particular study. Significant difference in altitude was spotted in all studies. For example, (Janeau et al., 2004) carried out their tests in an area with altitude ranging from 800 to 1400m, whereas the test area of (Cain et al., 2005) varied between 200 and 900m. In most of them, a huge diversity of canopy cover was presented. Thus, environmental conditions were very heterogeneous.
Tests were set and performed in different ways, however, since they had similar goals, the results and conclusions have been comparable. In order to test effects of canopy cover, some authors prepared various canopy classifications, such as classification of canopy cover according to its density (DeCesare et al., 2005), or classification according to type of forest and forest stand height (Janeau et al., 2004). Some other had not applied the same strategy stating they performed tests in different stand and canopy features (Klimànek, 2010; Tucek and Ligos, 2002), whereas (Lewis et al., 2007) decided to measure influence of canopy cover together with topography creating sixteen categories of different canopy and topography characteristics. Effects of topography were also evaluated in work of (Cain et al., 2005) by means of “map of available sky - AS”.

In terms of data processing, most studies are based on comparisons between referential and collected data, either in static mode (Lewis et al., 2007; Tucek and Ligos, 2002; Klimànek, 2010; Cain et al., 2005) or in dynamic mode (DeCesare et al., 2005; Janeau et al., 2004). First approach (static mode) considers that data were collected on already defined locations with known coordinates (referential points) placed in different environmental conditions and then compared to referential points in order to calculate positional error. The second approach considers that data were collected in dynamic mode following routes which had to pass throughout already defined classes of habitat (environmental conditions) in order to be used in making comparison of GPS performances within different habitat classes. Overall, the objective is the same – to establish relations between GPS performance and different environmental factors and their combinations. For that purpose, ANOVA test (Analysis of variance) or MANOVA test (Multi-factor analysis of variance) are currently used. The ANOVA test includes only one dependent variable while the MANOVA includes multiple dependent variables. Also, the main aim of ANOVA is to determine the differences in samples’ means, whereas MANOVA determines if the dependent variables get significantly influenced by changes in the independent variables. For example, (Klimànek, 2010; DeCesare et al., 2005) used ANOVA test to distinguish between effects of stand characteristics, dilutation of precision - PDOP and measurement time, and latter to distinguish between effects of different canopy categories. To distinguish between canopy categories, (DeCesare et al., 2005) applied an ANOVA test followed by post hoc analysis by means of Tukey test (Tukey, 1953).

Significant results of ANOVA test only prove that the aggregate difference among the means of the several samples is significantly greater than zero. However, it does not provide information if means of particular samples significantly differ between each other. For that purpose, it is necessary to apply multi comparison test as a posteriori. On the other hand, (Tucek and Ligos, 2002) evaluated

influences of their factors of interest (receiver type, stand age, tree species composition and terrain configuration) relying on MANOVA test.

Multi comparison can also be conducted by using Bonferroni method. For example, Cain et al., (2005) implemented Bonferroni method to compare effects of various AS classes and fix rates. Further, in order to evaluate continuous variables canopy cover, satellite view and their interaction, (Lewis et al., 2007) relied on multiple linear regression. Moreover, differences in GPS performance depending on different factors can also be statistically evaluated (identify) using testing of hypothesis for wanted level of significance.

1.2.5.2 Influence of topography

Studies studying the influences of topography (relief) consider either topography in general or influences of the orientation of slope in particular.

Topography in general
Conclusions of studies related to the influences of topography are very sharply divided. Authors such as Klimànek (2010) and Tucek and Ligos (2002) reported no statistically significant differences in the positional error among points collected in locations with different topographic characteristics, the former not giving information about level of significance, the latter using level of significance $t_{0.025}$. On the other hand, some research works confirmed a relation between positional error and topographic conditions (Lewis et al., 2007; Cain et al., 2005; DeCesare et al., 2005).

In terms of missing 2D position, Cain et al., (2005) found that fix rates varied significantly between designed classes of available sky (based on topography), whereas Lewis et al., (2007) found that fix rates were very high in all habitat types, thus not related to different topographical characteristics.

As we described previously, collected locations could be 2D and 3D. Since it depends on the number of available satellites, 3D fixes (Z coordinate) can be missing frequently. Hence, it dependents on topography. Results provided by Lewis et al., (2007) and Cain et al., (2005) strongly confirmed negative effects of topography on success of collecting 3D positions. In his research, Lewis et al., (2007) demonstrated that the proportion of 3D positions plunged from 97.7% to 58.9% between locations with low canopy cover and low terrain obstruction and locations with high canopy cover and high terrain obstruction. Similarly, Cain et al., (2005) stated “The proportion of 3D fixes was significantly different between available sky classes”.

Orientation of slope
As far as we know, no empirical study shows a relation between orientation of slopes and positional or altitude error. Anyway, (Lewis et al., 2007) claimed that “satellite view accounted for the fact that satellite coverage is greater on south-facing slopes than on north-facing slopes at northern latitudes.” Similar attitude was made by D’Eon and Delparte (2005) who expressed their reservation concerning the results of their GPS orientation tests due to the fact they were performed on the northern hemisphere.

1.2.5.3 Influence of canopy cover

Four different aspects of canopy cover were distinguished as most influencing on GPS measurement process: percentage of canopy coverage, type of trees, leaf off or leaf on period, and height of trees. Following findings from the state of the art will be presented in that order.

Percentage of coverage
All authors fully agree that the canopy cover and its density influence positional error considerably. Lewis et al., (2007) concluded “As canopy cover increased and satellite view decreased, location error

increased and PDOP values tended to increase". Moreover, they compared extreme values of Circular positioning error (CEP) between 19,9m for locations with high canopy cover and terrain obstructions and 284m for locations with opposite characteristics. (DeCesare et al., 2005) calculated that "GPS error added an average of 27.5% additional length to tracks recorded under high canopy, while adding only 8.5% to open canopy tracks real length."

As we stated before, general conclusions for effects on positional errors may be also applied on errors of altitude. However, apart from mentioned conclusions, no other papers dealing explicitly with relation between canopy and error of Z coordinate have been found.

Regarding the habitat types, Janeau et al., (2004) tested fix rates in different habitat types, mainly differing according to canopy characteristics (e.g. small deciduous, large deciduous, medium coniferous, small mixed coniferous, large mixed coniferous, very large mixed coniferous, open field). Percentage of failed 2D positions rose gradually as canopy features became more obstructive and finally reached its highest level in very large mixed coniferous forest (39.7%). In contrary, tests done by Lewis et al., (2007) resulted in very high fix rates (mean 99.5%, range 97.9–100%) in all habitat types, thus no related to different canopy characteristics.

Dependence of 3D fixe rates and canopy cover has been already mentioned in the part dedicated to topography effects on location error, since (Lewis et al., 2007) tested effects of an interaction between canopy and topography on GPS performances. Once again, they found out that the proportion of 3D positions plunged from 97.7% to 58.9% between locations with low canopy cover and low terrain obstruction and locations with high canopy cover and high terrain obstruction. Additionally, a proportion of 3D positions fluctuated between habitat models proposed by (Janeau et al., 2004). Differences are considerable between Open area and Large mixed coniferous forest – almost 70% less 3D positions recorded in later. The drop of 3D positions between Open area and Very Large mixed coniferous forest is smaller – 59.9 % which seems to be a bit strange. However it can be treated as exception, since all other differences between forests sizes are expected – less 3D positions acquired in larger forest, as it can be observed in the Figure 9.

Habitat types	6-channel GPS collars			
	n	% Failed	% 2D	% 3D
Open field	145	1.4	16.5	82.1
Small deciduous	145	2.8	27.6	69.6
Large deciduous	140	7.9	41.4	50.7
Medium coniferous	107	7.5	55.1	37.4
Small mixed coniferous	96	1.0	51.1	47.9
Large mixed coniferous	117	32.5	53.8	13.7
Very large mixed coniferous	126	39.7	38.1	22.2

Figure 9 : 3D positions acquired in different canopy models (Janeau et al., 2004)

Type of trees (species)

Regarding the influences of tree species on measurement accuracy no statistically significant effect of types of trees has been found (Klimànek, 2010; Tucek and Ligos, 2002). A statistical significant result was only spotted between broadleaved and coniferous forest, whereas there were no significant results found between other types of spices tested. By testing stand age effects on positional error, both authors discovered that measurement error increased in different stand age compositions for all tree species. Nevertheless, another study carried out by Janeau et al., (2004) demonstrated significant decrease of GPS performance between boreal and mix coniferous forest,

which indicates high correlation between positional error and some tree species. Influences of tree species in combination with size of forest very influencing on 3D positions success rate are shown in Figure 9.

One way ANOVA testing distances' error of GPS tracks recorded in study conducted by (DeCesare et al., 2005) showed insignificant difference ($F_{2t92} = 0.696$, $P = 0.501$) across canopy cover for available satellites. However, they also stated that the reason for that was almost equal constellation of satellites across all canopy types tested. Even if we know that by default number of satellites in highly related to location success, especially 3D locations, no information about that has been found in reviewed papers.

Leaf OFF/Leaf ON period

The only study in reviewed papers mentioning the leaf on / leaf off period is the one of (Janeau et al., 2004) that did not report any influence, as they report: "During the movement tests, we found no negative effect of leaf presence on 6-channel GPS collars. Therefore, the increased location success recorded on free-ranging red deer during leaf-off season cannot be explained by leaf absence, but rather by modifications in spatial behavior of individuals."

Height of trees

This factor was not taken into account separately, but some relevant conclusions were made in work of (Janeau et al., 2004): "By conducting movement tests in our temperate forest mountain conditions, we found that the performance of GPS collars decreased under mixed coniferous forests with tall trees."

1.2.5.4 Measurement time

Correlation between measurement time and positional error was very meticulously evaluated in the work of (Klimànek 2010) using differential GPS that can relies on post processing corrections of measured GPS vectors. Their experiments show that mean square error (MSExy) varied among different measurement intervals (e.g. 1min, 2min, 5min, 10min). After post processing, values of MSExy went down gradually for approximately 1m except between 5 and 10min intervals when it went up from 2.7 to 3.3m. Same trend was spotted in case without applying post processing corrections.

In the same paper, correlation between measurement time and 3D positions was particularly examined, and shows the same trend for Z errors than for 2D (X,Y) positional errors. Differences between Z and (X,Y) errors are slightly smaller in case without post processing, but almost twofold smaller in case with post processing corrections.

Successful positions or fix rates were also proved to be sensitive to measurement time intervals. Cain et al., (2005) demonstrated that percentage of acquired locations decreased gradually as measurement time rose. Hence, measurement intervals of 0.25, 0.5, 1, 4, 6 and 13 hours resulted in 99, 98, 96, 94, 93, 92% of fix success rate respectively. However, statistically significant differences for confidence interval of 95% were only detected between the longest and the shortest intervals. A similar conclusion was made by Janeau et al., (2004) "Location success increased with a shorter location interval (10 min vs 180 min)."

Situation with 3D fixes were separately examined in studies done in (Cain et al., 2005), but significant differences related to measurement time have not been found.

1.2.5.5 Snow

Test done by Janeau et al., (2004) showed big influence of snow accumulated on branches of trees in mixed coniferous habitats on both, successful 2D and 3D locations. The fix rates were "drastically reduced." Their hypothesis was that it could be explained by reflection of GPS signal caused by snow coverage, especially connected to percentage of water in the snow.

1.2.5.6 Type of GPS (differential or not)

Technical characteristics of a GPS device may improve its performance. Studying that aspect, Janeau et al., (2004) concluded: "Overall, 8-channel GPS collars performed slightly better than 6-channel GPS collars ($\chi 2\ 2 = 18.4$, $P < 0.001$), with more 3D locations and fewer failed attempts, except under large and very large mixed coniferous; In mountain environments, multipath can also have an important effect on GPS performance. These drawbacks will probably be reduced in the future by the use of 12-channel GPS receivers, with more time allowed for signal acquisition. "

1.2.5.7 Receiver position

Some studies paid attention on effects of different position of the device, i.e. how it is held, while collecting data. For example, Janeau et al., (2004) stated: "Wearing the collar on the shoulder was preferred because the presence of the body improves the ground plane underneath the GPS antenna, thus enhancing GPS antenna characteristics and reducing multipath effects, as when mounted on an animal."

Jiang et al., (2008) proved in their tests that device orientation consistently influences both fix rate and location precision. Additionally, collar orientation alone tends to have a negligible effect on GPS errors in open areas, but can reduce fix rates up to 80 per cent and location precision as much as 17m under dense canopy cover. Similar conclusions were obtained by Heard et al., (2008).

Blunck et al., (2011) did a comprehensive analysis of receiver position on positional accuracy. They tested GPS receivers integrated in two cell phones Google Nexus One and Nokia N97. Measurements were collected by a person walking a 4.85 km twice through both, open sky and urban areas. Six cell phones were carried in six different positions: Upper compartment of bag pack, Datastyle 5 fingers, Datastyle 3 fingers, Left jacket pocket, Trouser front pocket, Trouser back pocket. Figure 10 illustrates two hand grip styles: Datastyle 3 fingers and Datastyle 5 fingers.

Figure 10 : Illustrations of different hand grip styles (Blunck et al., 2011)

Collected traces were compared to ground truth data recorded by high accuracy u-blox LEA-5H receiver. Figure 11 illustrates cumulative distributions of the individual positioning errors of four collected traces. The biggest differences in positional accuracy can be spotted between Upper compartment of bag pack and Datastyle 5 fingers in case of Google Nexus One and between Trouser back pocket and Datastyle 3 fingers in case of Nokia N97. Considering the median, in case of Google

Nexus One the rise of positional error is from five to ten meters, whereas in case of Nokia N97 is from ten to twenty meters. Hence, it is obvious based on this test that the positional accuracy is considerably affected by position of GPS receiver while the measurements are conducted.

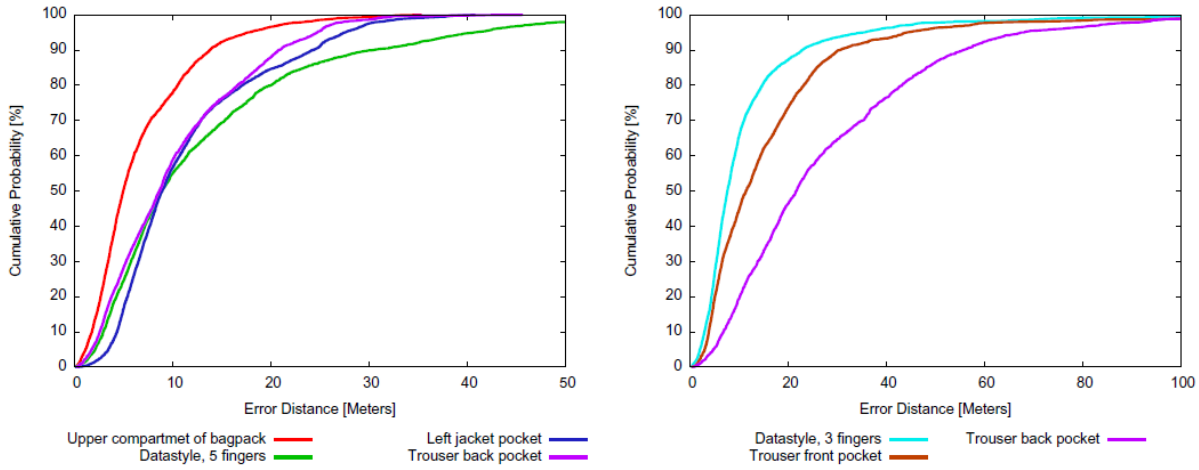


Figure 11 : Drops in positioning accuracy with different hand grip styles and body placements: Google Nexus One (left) and Nokia N97 (right) (Blunck et al., 2011)

1.2.5.8 Multipath effect

Tucek and Ligos (2002) found an increase of measurement error due to multipath effect caused by obstacles for sheer signal reception made by trunks and branches.

1.2.5.9 Conversion

Conversion from WG84 to a local coordinate referential system is necessary for compering GPS measurements to reference data. However, we found only one paper that recognized this procedure and its potential effects on the exactness of GPS measurements. Based on tests performed (Tucek and Ligos, 2002) claimed: "The conversion accuracy was tested earlier on an experimental polygon and did not exceed the error tolerance applied at the tested points." Thus, the effect of conversion is negligible.

1.2.5.10 Overview

Based on review of all identified factor influences on GPS measurements, a favorable environment for GPS measurements could be distinguished. This is the environment that implies flat areas, or less complex relief areas with low or without canopy cover, preferably composed by broadleaved species without snow coverage, lower frequency and not so long measurement time, followed by higher number of visible satellites and conducted by differential GPS devices paying attention on position of device and its orientation.

Table 1 shows an overview of all identified factor influences on GPS measurements. Influences of each factor on specific aspects of GPS measurements such as positional error, error of attitude, missing position, missing 3D position and missing timestamp are labeled by a star (*) in case the impact is medium, or with two stars (**) in case the impact was found high. Empty filed means that the impact was not found or tested.

It has to be stressed that most of those studies were done on professional and collar GPS devices.

Table 1 : Overview of state of the art results

FACTORS		TYPE OF ERROR			
		(X, Y) - positional error	Z – error of altitude	Missing position	Missing 3D position
TOPOGRAPHY	In general	**	**	**	**
	Orientation	*	*	*	
CANOPY COVER	% of coverage	**	**	**	**
	Type of trees (specis)	**	**	**	**
	Leaf off/on	*	*	*	
	H of the tree	**	**	**	**
TIME INTERVALS	Measurement time	*	*	*	
TYPE OF GPS	Differential/ Non-differential	**	**	**	**
POSITION	Holding GPS	**	**		
	Orientation of GPS	*	*	*	*
MULTIPATH	Obstructions	**		*	

1.2.6 Detection of outliers in VGI traces

GPS data is affected by a plenty of influences during the measurement process as presented in subsection 1.2.5. Owing to that, GPS data may be significantly degraded by presence of outliers. Thus, it is important to conduct a filtering before proceeding in further analyses. On the one hand, it gives a global overview of data quality: a dataset with high percentage of outliers has a lower global quality than a dataset with a small percentage of outliers. On the other hand, filtering improves the quality of the examined data. In general, two outlier detection approaches in GPS data may be distinguished:

- Considering GPS measurement errors as outliers
- Considering geometric anomalies of resulting VGI traces as outliers

In the first approach, the assumption is that outlying observations (pseudo and code distances), cause geometric incorrectness and anomalies in GPS data. Following this approach, Knight and Wang (2009) proposed an outlier test that examines measures of pseudo ranges as more successful way in detecting outliers in GPS measures. Similarly, relying only on GPS observations, Ordonez et al., (2011) tested their approach to identify outliers. GPS speed and time measures was used by Duran and Earleywine (2012) to detect outliers caused by effects of GPS data logging errors such as sudden signal loss, data spiking, signal white noise, and zero speed drift. On human mobility using mobile phone data, Iovan et al., (2013) proposed a method based on speed and direction between consecutive points to filter erroneous points (i.e., locations produced by ping-pong phenomenon, a very well-known phenomenon in telecom domain).

Some authors proposed the application of different filters such as Kalman, not exactly for detection of outliers but for minimizing their effects on GPS trace geometry. The main idea is refining traces geometry especially on the outlying points. Kalman filter was successfully used by Eliasson (2014) and Gomez-Gil et al., (2013) to reduce positional error for pedestrian applications in areas with bad GPS signal and to improve precision in low cost positioning of tractors respectively. Main idea and result of applying Kalman filter in the work of Eliasson are presented in the figure bellow.

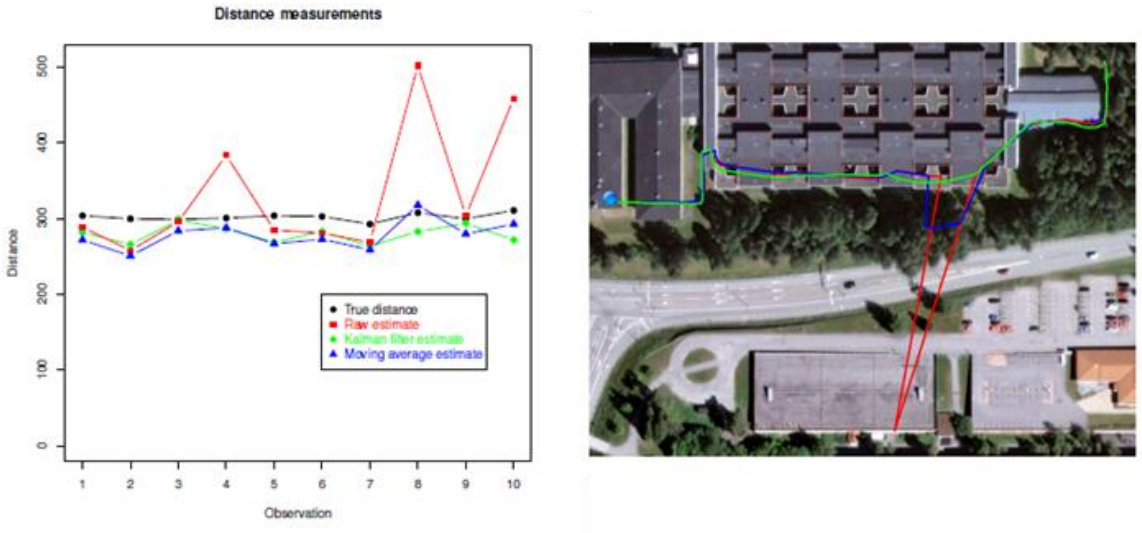


Figure 12 : Kalman filter application (Eliasson, 2014)

In Figure 12, on the left observations of distance measurements between consecutive are presented whereas on the right the geometry of the trace is illustrated. Red line labels raw values on the left and similarly row geometry of trace on the right. After applying Kalman filter on raw observations the outlying geometry is drastically refined on the outlying point, presented in green color on the right.

The second approach has been less followed for the detection of outliers (Etienne 2013 and Gil de la Vega et al., 2015). According to Gil de la Vega et al., (2015) an outlier is a point, segment or track which differs from the general shape of other tracks following the same itinerary, as illustrated in the Figure 13, where outlying elements (i.e. point segment, multi-segments and traces) are labelled by red colors, whereas the mean trajectory of VGI traces is illustrated in blue.

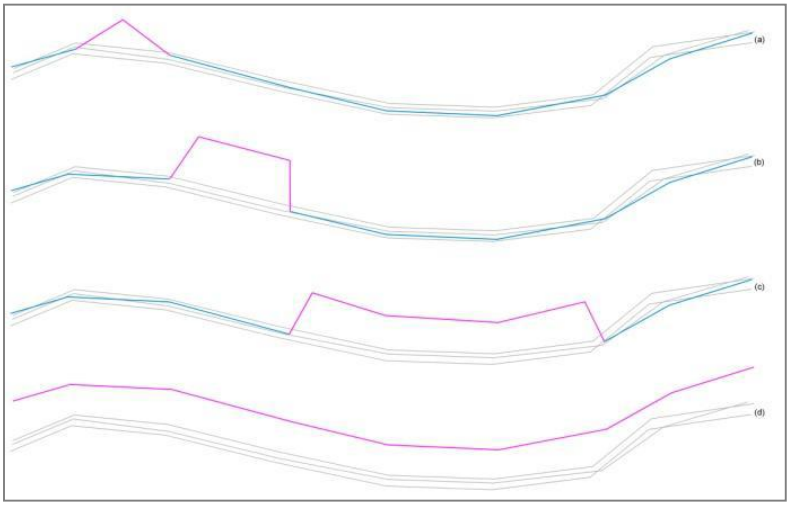


Figure 13 : Outliers according to Gil de la Vega (2015)

Outliers are detected using mean 3D axis calculated based on intersections of traces and perpendicular planes along the path. A GPS point is an outlier if the distance to the mean axis exceeds a given threshold. Outlying points are subsequently eliminated and the traces are reconstructed. The process is iterative and is done as long as there remain outliers to be eliminated. Results are illustrated in the Figure 14:

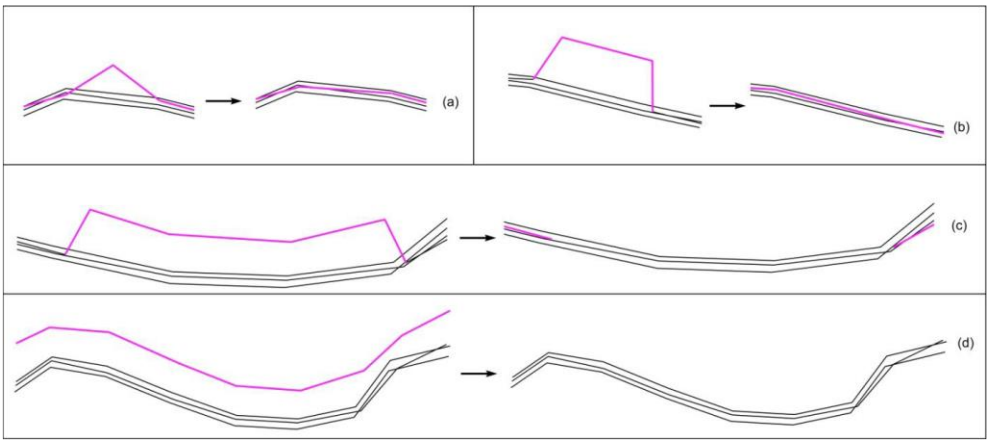


Figure 14 : Reconstruction of the traces after filtering (Gil de la Vega, 2015)

Using 3D axes of trace in detection of outliers of trajectory was also proposed by Etienne (2011). However, the components of this 3D axe are not the same as in previous example. Two components are standard X and Y, whereas the third component is time. For each position in the trace, spatiotemporal limits can be combined so that the 3D zone of spatiotemporal normality can be defined. The zone is modeled by means of quadrangles like presented in the Figure 15 (in green).

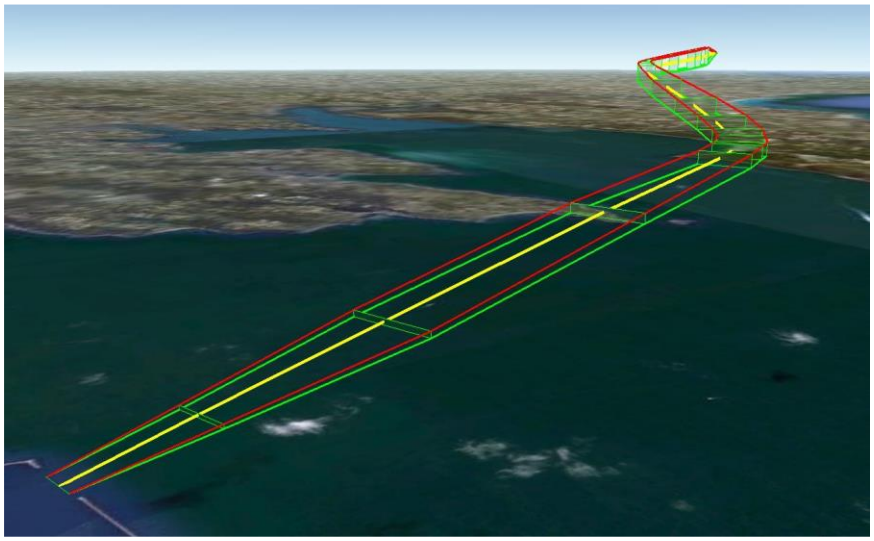


Figure 15 : Illustration of the approach proposed by Etienne (2013)

In Figure 15, the median 3D trace is presented in yellow, interpolated limits of spatiotemporal corridor are presented in red and green. Outside the zone bounded by red boundaries the positions in the traces are late compared to time component, whereas outside the green zone the position of traces is in advance. Outliers are all positions in the traces outside the 3D zone of spatiotemporal normality.

In conclusion, two main approaches for the detection of outliers in GPS data exist, one considering raw GPS measurements for potential outliers, and the other considering outstanding geometries as outlying. The first approach requires complete and reliable raw GPS measurements, whereas the second one requires sufficient number of VGI traces following the same itinerary. However, in our context, none of these conditions are respected. First, VGI GPS data usually does not contain raw GPS measurements and has a lack of metadata, thus the first approach is not applicable in this situation. Second, in some challenging areas, like mountainous areas there is a lack of VGI traces, that is, the number of traces following the same itinerary mostly varies from 1 to 3. Relying only on 2 or 3 traces in application of the second approach is not sufficient for obtaining reliable and good results.

1.3 A practical study of VGI traces completeness

The goal of this section is the evaluation of some existing GPS data quality elements, focusing on attribute completeness and data heterogeneity. Let us notice that, in practice, our input data is coming without metadata like transportation mode, spatiotemporal resolution etc.

1.3.1 Data description

In total, we downloaded 437 traces composed by 292,337 points from French web-sites specialized for sharing traces between hikers and mountain bikers. The web-sites will be fully presented in Appendix 1.

The traces were collected in Vosges Mountains (France) during sport activities such as walking, running, hiking and cycling. However, since the traces are coming without metadata, they are not classified according to transportation mode. That is, only global information about sport activities exists but not specific information for each trace. In addition, spatiotemporal resolution of traces is unknown.

The position of Vosges Mountain is shown in the Figure 16 (bounded by red line).

Figure 16 : Location of our test area

Therefore, in order to examine collected data and assess its quality, the completeness and accuracy of attributes was first examined. Second, the heterogeneity of the dataset was tested statistically.

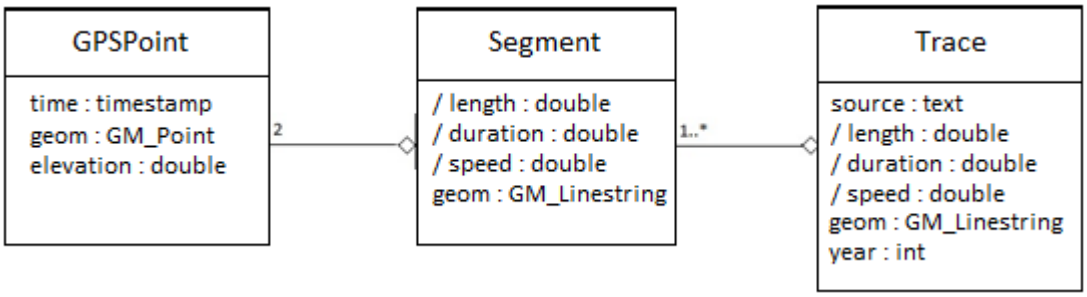


Figure 17 : UML model of VGI traces

In order to evaluate its quality and apply different analyses, our data are modeled as presented in Figure 17 : UML model of VGI traces. The data is represented through three different classes: points come directly from GPX files, whereas segments and traces are subsequently derived from points so that different analyses can be applied. The quality of data can be evaluated from different aspects. For instance, segments are important for calculation of indicators that will be used in detection of outliers and low accuracy points, as well as the traces.

1.3.2 Results of analysis

GPS points are theoretically described by, at minimum: 2D coordinates (WGS-84), timestamp, and elevation.

Practically, we notice that some points have missing attributes (omission), that is, missing elevation or timestamp or even both. We computed six indicators measuring the completeness of the attributes, at the point and trace level (see Table 2). We found that 106,206 points (36.3%) lack timestamps, whereas the situation with elevation is better – only 6,580 points (2.2 %) lack elevation. Regarding the traces, we noticed that 157 (35.9%) have no timestamps at all, whereas 287 (65.6%) have at least one missing timestamps, which is less crucial for further analyses.

Table 2 : Missing attributes in GPS data

	Numbers of points	Number of traces	%
Missing timestamp	10,6206		36.33
Missing elevation	6,580		2.25
Traces with at least one missing timestamp		287	65.67
Traces with at least one missing elevation		22	5.03
Traces without timestamps		157	35.93
Traces without elevation		10	2.29

Timestamps are an important attribute that characterize VGI traces which represents the time where a point is collected. Much analysis can be done based on them such as speed based analyses. Such random presence of missing timestamps and especially a huge percentage of traces without timestamps at all 35.9% are a significant issue in using this data and assessing its quality. Those issues will be more discussed in following sections.

We already mentioned that heterogeneity of VGI GPS data is an important issue for using those data as a data source. To evaluate that, we identified segment speed and length as most representing aspects for measuring VGI traces heterogeneity. The heterogeneity was examined by using normality tests.

First testing on normality was done by means of Shapiro-Wilk test (Shapiro and Wilk, 1965). Having an ordered random sample, $y_1 < y_2 < \dots < y_n$, the original Shapiro-Wilk test is defined as:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

Where x_i is the i^{th} order statistics,

$\bar{x}$ is the sample mean,

$a_i = (a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$

and $m = (m_1, \dots, m_n)^T$ are expected values of the order statistics of independent and identically distributed random variables sampled from standard normal distribution and V is a covariance matrix of those order statistics.

The test was originally designed for small sample size data, less than 50. In addition, Royston (1992) was proved that original Shapiro-Wilk test was not efficient for samples bigger than 50. Therefore, in our testing we applied a modified version of the test (AS R94 algorithm) proposed by Royston (1995).

For assessing the heterogeneity of speed, 163 traces not affected by missing timestamps were tested for the statistical significance p=0.05. For only 7 (4.3%) of them Null hypothesis was not rejected, in other words, they were found normally distributed by the test. Figure 18 illustrates a trace with normally distributed speed

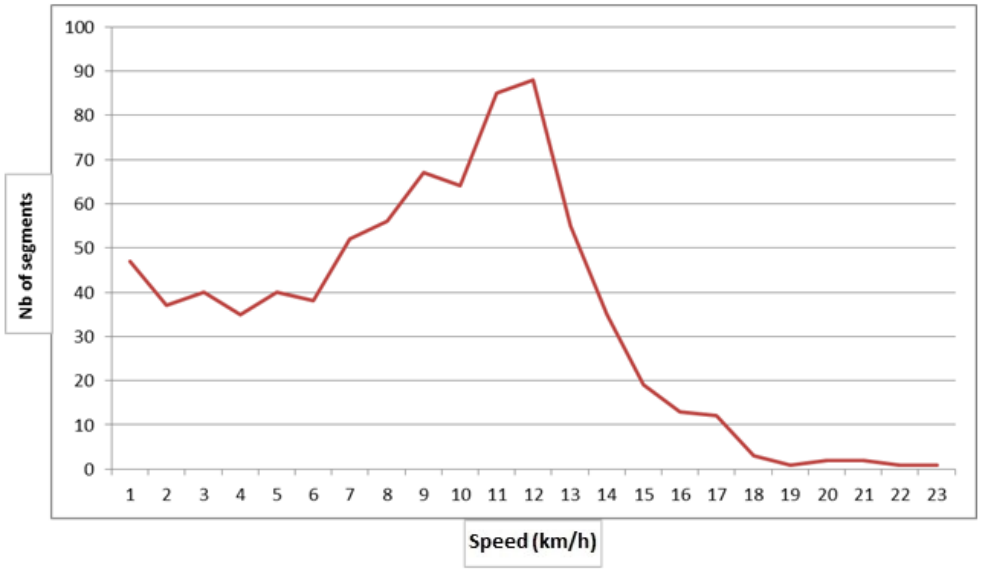


Figure 18 : A VGI trace with normally distributed speed according to Shapiro test

In terms of segment length, Shapiro test found no traces with normally distributed segment lengths for the same statistical significance: p=0.05. This finding proves spatial resolution of the traces not uniform.

In order to confirm obtained results, the distributions were also tested on normality by measuring skewness and kurtosis. Skewness represents a measure of asymmetry of the probability distribution of a random variable around its mean. A unimodal distribution can be symmetric normal, positively or negatively skewed depending on relations between mean, median and mode, as illustrated in the Figure 19:

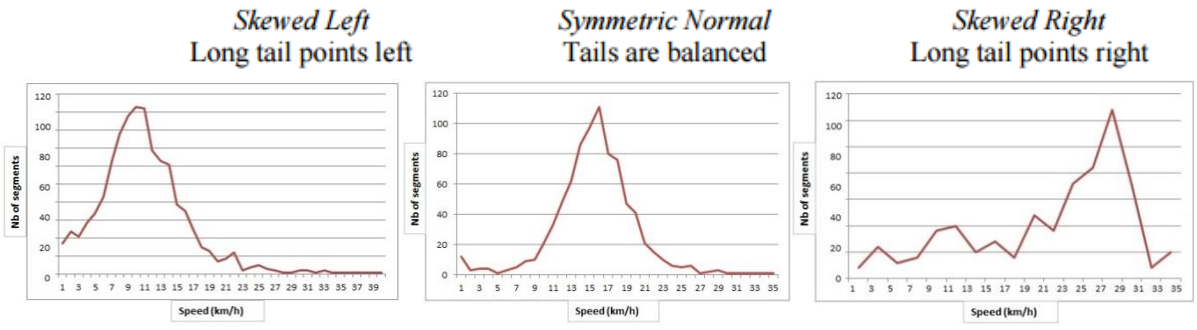


Figure 19 : A unimodal distribution according to skewness (VGI traces speed distribution)

Kurtosis describes the shape of the peak in a distribution of random variable, i.e. how sharp it is.

Data having both skewness and kurtosis between -2 and 2 can be considered normally distributed (George and Mallery, 2010; Trochim and Donnelly, 2006; Field, 2000 and 2009; Gravetter and Wallnau, 2014).

Figure 20 illustrates differences between distributions with lower and higher kurtosis.

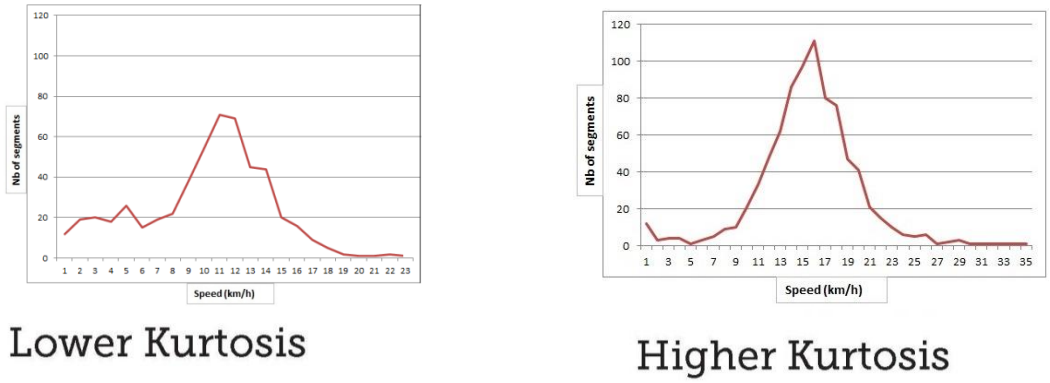


Figure 20 : Lower and higher kurtosis distributions (VGI traces speed distribution)

Regarding speed distributions, skewness varies from -0.39 to 3.17, whereas kurtosis ranges between 1.68 and 12.02. After testing trace by trace, we discovered that, in total, only 16 of 163 traces (10%) tested, fulfill that condition, which confirm the results from the Shapiro test.

Similarly, only 2 traces from the entire dataset were found normally distributed regarding segment length distribution.

Not normal speed distributions can be explained by frequent changes of the speed and even transportation mode during an itinerary required by a very rough relief in mountainous area such our test zone is.

It is important to assess its heterogeneity especially before proposing a method and criteria for detection of outliers or detection of low precision points. The criteria are very sensitive on the data heterogeneity, especially how the thresholds are set, since in case of high heterogeneity significant over and under estimations can be caused.

1.4 Proposal of an approach for assessing spatial quality of VGI traces

Using VGI traces in update of high quality referential road network put a significant importance on their quality issues. Thus, the quality of traces has to be satisfying and also well known.

The focus of this chapter is to describe the approach we have proposed to assess the quality of VGI traces. Most of the approaches for assessment of GPS data quality have been adapted to regular GPS data. Thus, very frequently they are not applicable in case of VGI traces due to the lack of protocol and metadata required for such type of assessment.

In addition, use of referential data like proposed in extrinsic approaches seems to be a bit problematic and inconsistent in this specific case – when VGI traces are supposed to be used for updating referential road network. A fact that a new road appeared means there has not been a corresponding referential

road the new road could be compared to. Indeed, some general prediction of quality distribution of VGI traces in different areas can be done by assessing extrinsically those traces having corresponding referential roads to be compared to. However, this can also be considered an issue, since using referential data for evaluation of data used in updates of that referential data seems to be an inconsistent and biased approach.

Therefore, the approach we have proposed tends to be intrinsic as much as possible. First, to avoid using referential road network due to the above elaborated reasons. Second, the approach less depended on referential data is more interoperable, thus can be applied on various VGI traces dataset.

In general, the approach includes filtering and evaluation of each point of each trace, point by point, trace by trace, only relying on its geometry and metadata available, as well as very few pieces of external data describing the environment. Those found influencing in the state of the art of GPS traces quality.

1.4.1 Approach

We propose an approach composed by five main steps presented in the Figure 21.

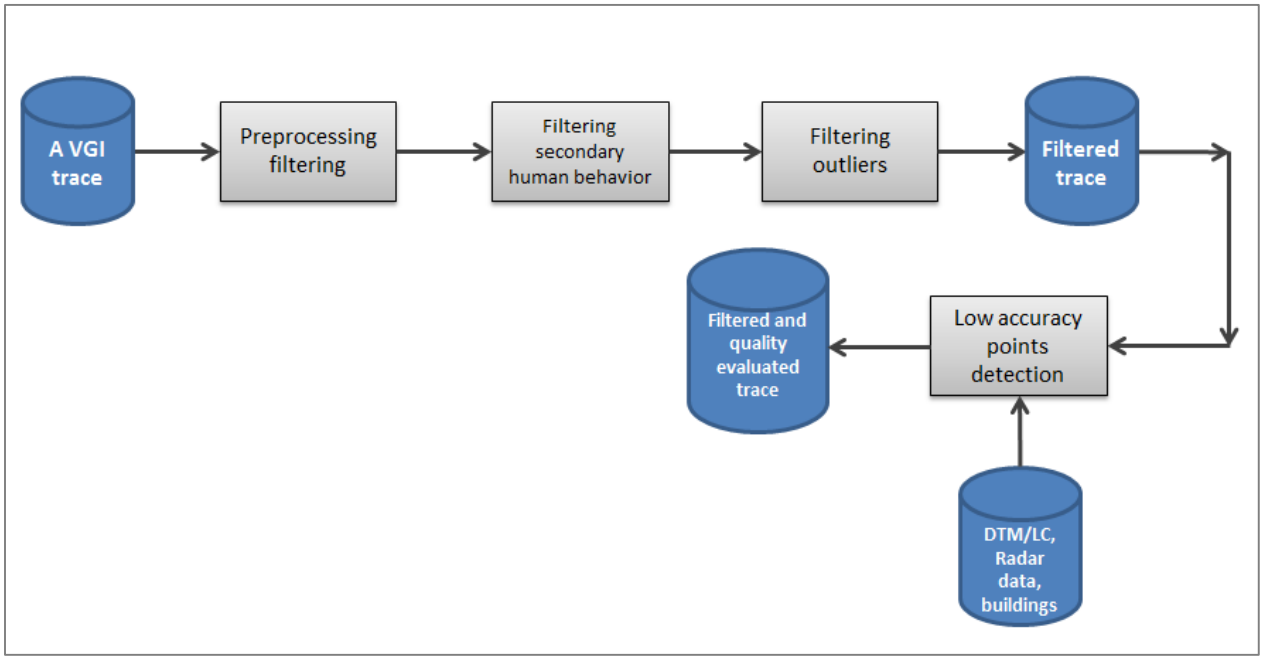


Figure 21 : VGI trace's quality evaluation

First of all, each VGI trace is filtered in terms of aberrant speed, secondary human behavior and outliers. The result of three filtering steps is a trace of better quality more suitable for being used in data matching with referential data. Next, the remained points of filtered trace undergo evaluation of their accuracy by using additional referential data like land cover maps or Digital Terrain Model (DTM). As a result, all points are classified according to their accuracy.

1.4.1.1 Preprocessing filtering

Each VGI trace needs to undergo a prelaminar filtering step that consists of redundant points and negative speed values filtering, as well as filtering of irregular values of some attributes. Redundant

points are overlapping points created due to the stops in movement or not proper work of GPS receiver. We detected them by means of distance values very close to zero. Negative speed values are very strange phenomenon, however they exist in VGI traces due to the errors of GPS clock or when GPX file are created, when a preceding timestamp has a value like it was recorded after a succeeding timestamp.

When it comes to attribute values filtering, significant discrepancies are spotted in values of elevation as well as timestamps. Apart from regular and NULL values of elevation, two other values were found: zero value and the same value for all points in a trace. Besides regular and NULL values for timestamps, there exist traces with the same value for all points in a trace. Since those values of the attributes elevation and timestamps are not correct and possible in the reality, they were all replaced with NULL values. As a result, errors in analyses based on point attributes were avoided.

Final results of preprocessing filtering will be presented in Section 1.4.2.1.

1.4.1.2 Filtering secondary human behavior (SHB)

Traces are coming from human sport activities. In majority of cases mentioned activities are conducted along existing roads and according to planned itineraries. However, taking off from the main itinerary or even existing road (e.g. off road movements) happens sometimes due to different reasons. It may be a break in the nearby meadow or drinking water from a spring next to the main road, visiting an interesting point (e.g. a monument) etc. We will not go more deeply into the reasons, since this research is not devoted to analyses of human mobility and behavior, but we are interested in effects of this behavior on the quality of VGI traces for our purpose. Results of such behaviors are thought of as local geometric anomalies of a trace. Sometimes they are complex like illustrated in Figure 22

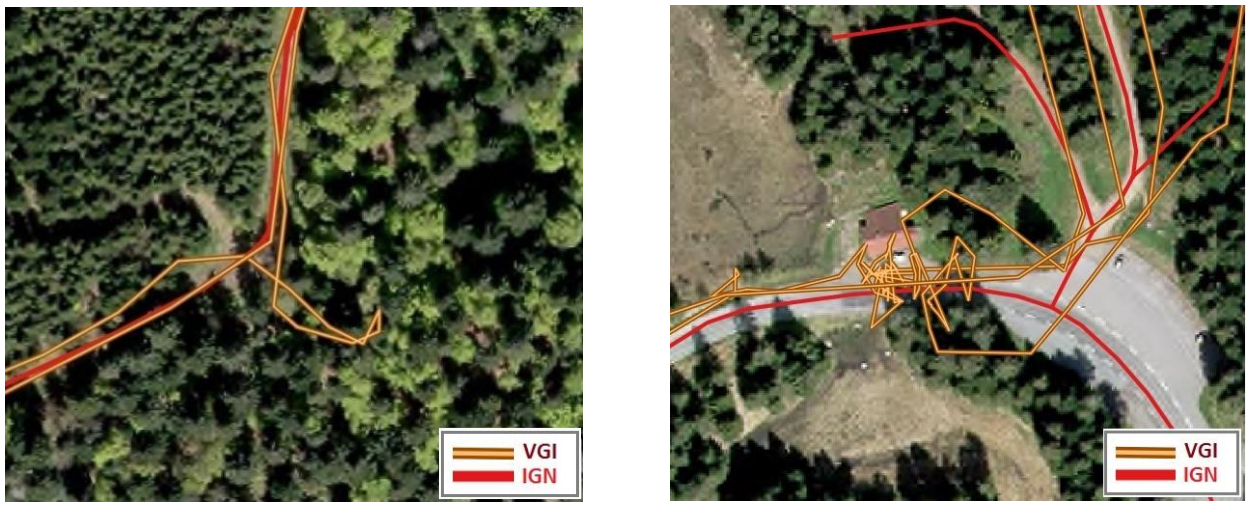


Figure 22 : A simple secondary human behavior (on the left) and a complex secondary behavior (in the right)

Apart from negative effects on geometry of a VGI trace and on the rest of the process for evaluating the quality, secondary negative effects could be caused when matching traces with referential road network. Firstly, a lot of segments of a VGI trace will not be matched with referential roads because of inconsistent shapes. Secondly, a risk that parts of irregular geometries will be mismatched with some referential roads is high, especially in areas with denser road network. For that reasons and because more generally geometries produced as a result of secondary human behavior (in the rest of the text SHB) contain irrelevant information for our work, we propose to filter them.

This phenomenon is mainly due to stops in human mobility. Detection of stops is very well known process in human mobility study (Thierry et al., 2013). Most of the studies to detect stops are relying on spatial information (Buard, 2013), spatiotemporal information (Alvares et al., 2007; Zimmermann et al., 2009; Rocha et al., 2010; Olteanu-Raimond et al., 2012) and temporal information (Palma et al., 2008; Zhixian et al., 2010; Buard, 2011). However, a significant number of traces affected by missing timestamps in our test data show that this is a limitation for the applicability of the approach.

Moreover, the significant spatial resolution heterogeneity presented in Section 1.3 does not allow the application of techniques based on local density of points like DBSCAN (density-based spatial clustering of applications with noise)(Ester et al., 1996) since that would cause a lot of overestimation in stop detection and subsequently leads to significant data loses. Figure 23 illustrates an example of significant spatial resolution heterogeneity on the level of trace number 7. Mean resolution is about 20m. Most of the points are placed 20m from each other, plus-minus few meters, whereas there are some parts of the trace with a group of points that are significantly closer to each other. Those points are labeled in yellow and have a distance between each other from 3-5 meters. By adjusting both parameters in DBSCAN, distance and a number of points composing a stop, this problem remains permanent.

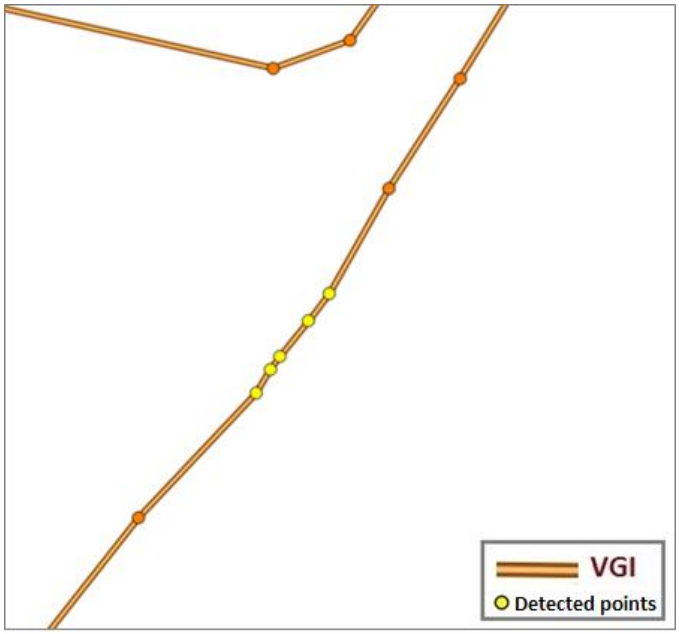


Figure 23 : Limitations of DBSCAN

Therefore, our method is designed to suit the traces randomly affected by missing timestamps, and with high heterogeneity of spatial-temporal resolution like VGI traces are. After an observation of the phenomenon, we noticed that SHB are characterized by two main characteristics: huge direction change between segments and trace's self-intersection. However, since regular geometry of roads in mountainous areas can be very complex and sinuous (huge direction changes), a change of direction only seems not to be a sufficient discriminating criterion for SHB detection (see Figure 24).

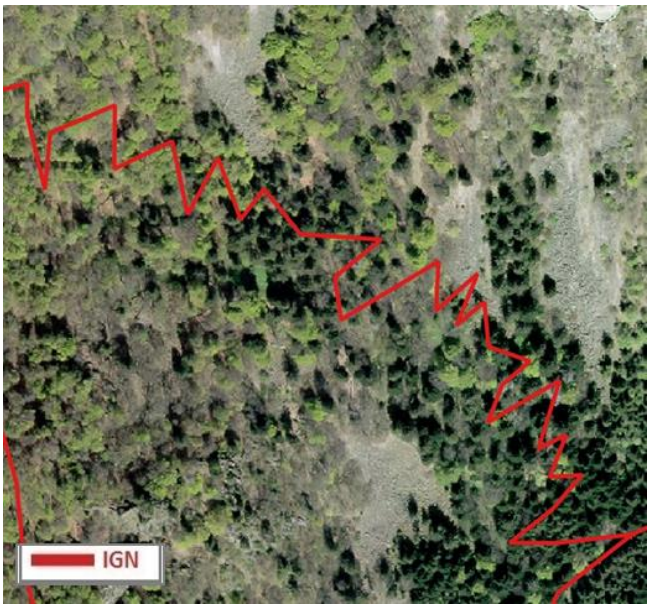


Figure 24 : Complex geometry of referential roads in mountainous area

The ‘no self-intersection’ principle seems to be a more efficient criterion. First, a vast majority of examples we found are characterized by self-intersection. Second, this criterion does not depend on sampling effect, and thus seems adapted for heterogeneous data.

More precisely, we propose to build polygons created from parts of the trace that self-intersect. All segments composing created polygon will be considered as parts of SHB and eliminated.

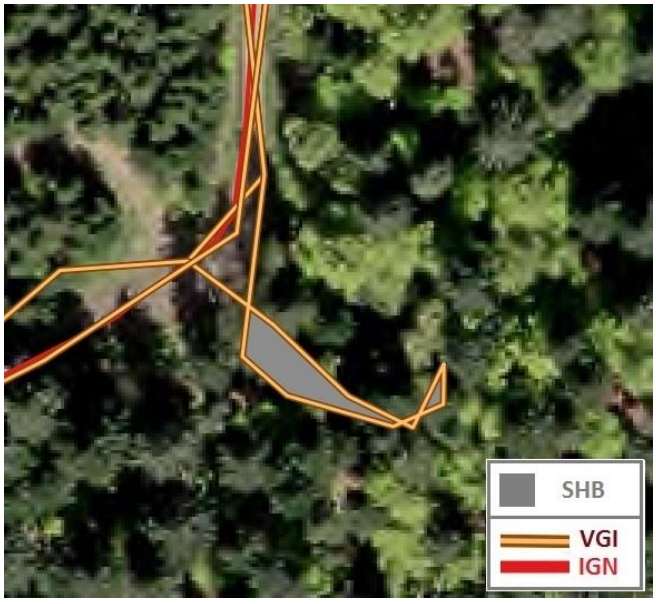


Figure 25 : Trace with a loop

Nevertheless, application of this criterion alone will not bring expected results. First, closed path traces (starting and finishing at the same point) could be falsely recognized as SHB as well as some big parts of itinerary that self-intersect. Thus, to refine that, we only consider polygons with a limited surface (less than 200 square meters). Second, a need for additional conditions is also very visible in case with round

trip traces. A round trip trace could have certain self-intersections since some parts of itinerary are the same. If it was threated only by proposed criteria, a lot of parts the trace would be lost.

Figure 26 on the left shows an example of a round trip trace. This example is drastic since it represents a part of itinerary that was passed 3 times (all tree lines are part of the same trace). As it can be observed, there is no SHB in this part of the trace, even if there are plenty of self-intersections, which would result in erroneous trace after filtering (see figure on the right).

Figure 26 : A Round trip trace with intersection – before and after wrong filtering

In order to avoid such a problem, criteria have to be defined to identify round trip traces. As we do not want to use unreliable timestamps to do that, we propose to look at the elongation of polygons created by self-intersected lines. More precisely, we compare the ratio of their areas and their corresponding minimum bounding circles - MBC (see Figure 27). This ratio is called elongation of SHB polygon.

Figure 27 : Secondary human behavior polygons and their corresponding minimum bounding circles

Figure 28 illustrates the difference of elongation in case of round trip traces and SHB. Here we can see that the elongations of real SHBs (gray polygons on the left) are significantly greater than those of round trip traces (gray polygons on the right).

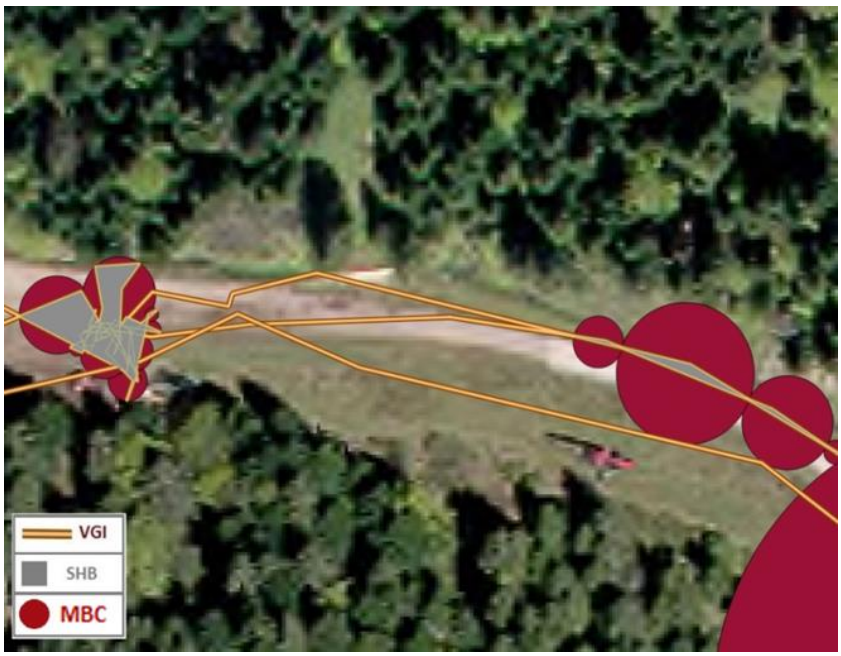


Figure 28 : Round trip traces polygons (on the right) and secondary human behavior polygons (on the left)

After filtering points recorded during secondary behavior, in some very rare cases, the results of reconstruction of a filtered trace can be new self-interactions which are in turn problematic. Thus, the process is repeated iteratively until there are no more self-intersections. In that way, side effects are minimized, and a global successfulness of the approach is increased.

It should be stressed that traces containing effects of SHB, if converted directly to polygons cause irregular geometry of polygons. Reasons for that are self-intersection of traces and presence of dangling ends. Polygons obtained in that way are not suitable for our purpose. Figure 29 presents result of converting such traces directly to polygons (on the right). As it can be noticed the result of lines to polygons operation is not a polygon related to the SHB that we need (presented on the left), but is a polygon created based on the rest of the trace – parts of the traces not involved in self-intersection.

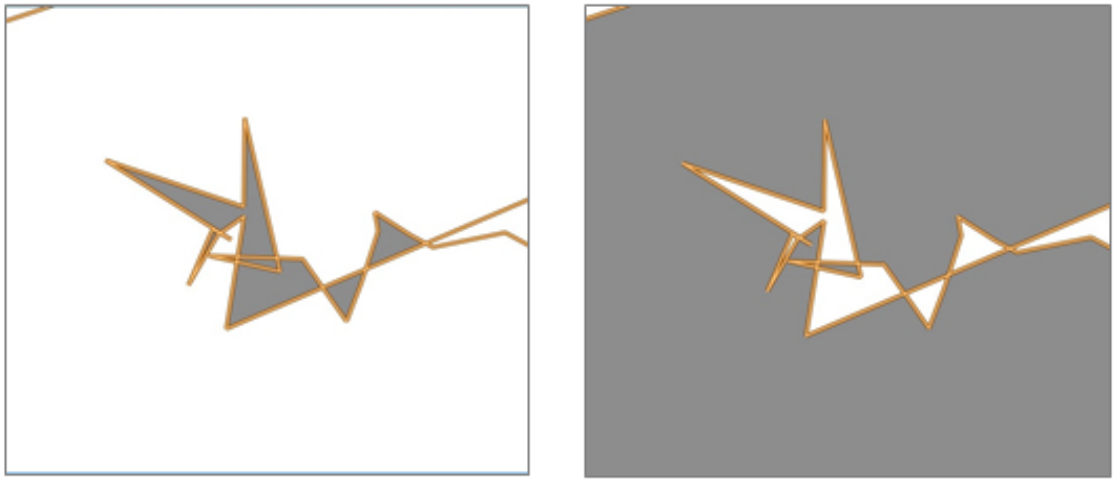


Figure 29 : Due to irregular geometry of a trace, polygons are not created properly

Therefore, the process of filtering requires a pre-processing part. Detailed description of steps conducted is presented below. First four steps belong to pre-processing part, whereas the rest are processing steps.

Step 1: Buffer of traces is created
Step 2: Buffer to lines
Step 3: Lines to polygons
Step 3a: Polygons with a surface > N square meters are eliminated
Step 4: Polygons are enlarged for a buffer zone to ensure that all points needed are included
Step 4a: All buffer surfaces bigger than N square meters are deleted – to eliminate secondary effects of previous operation

Step 5: Minimum bounding circles are created around remained polygons
Step 6: Join of minimum bounding circles with corresponding stops' polygons and calculation of elongation:

$$elongation = \frac{polygon\ area}{MBC\ area} \quad (2)$$

Step 7: Delete all stops' polygons having elongation < S, where S is a threshold
Step 8: Delete all points composing remained stops' polygons

Final results of detection of the method will be presented in 1.4.2.2.

1.4.1.3 Filtering outliers

As explained before, existing approaches for identifying outliers based on the analysis of GPS measurement errors, comparison between traces following the same path, or comparison to referential data are not adapted to our context.

Therefore, we are proposing to define an outlier as a GPS point whose metrics and geometrics characteristics differ significantly from the characteristics of other points composing a trace. Two examples of such points are presented in Figure 30. Points in yellow are examples of outliers whose positions cause values of speed, distance and angle between them and their consecutive points significantly different compared to other points of the same trace. For example, the outlying point on the

right has corresponding speed, distance and direction change values 31.4 m/s, 98.7m, 135° respectively, whereas trace's average values of those indicators are 2.4 m/s, 23.1m and 19° respectively.

Figure 30 : Examples of outlier points

The approach we propose is very important for the fulfilment of our final goal – detection of updates in authoritative road network. An update case "creation of a new road" by default means there has not been a corresponding referential road so that geometry of a new road (VGI trace) could be examined on presence of outliers. Figure 31 illustrates one of such situations:

Figure 31 : Outliers in new-created road

The trace in orange, shown in Figure 32, represents a new created road unfortunately endured by a significant outlier points. Like this, it cannot be proposed as geometry of a road to be updated in referential road network. After application of the approach, the points are detected and removed, and

geometry of a trace is significantly improved. More results are presented in the Section 1.4.2 devoted to the results presentation.

Figure 32 : New-created road without outliers

Criteria for detection of outliers

Due to the heterogeneous nature of VGI GPS data confirmed in section 1.3.2 and more random than systematic influences of environment like multipath, the detection of outliers is a complex task. Many intrinsic and extrinsic indicators may be computed to detect outliers. However, determining relevant criteria and thresholds as well as how to combine them is a challenging task. This can be observed while using some highly relevant trace indicators in modeling and detection of outliers that have been widely used for that purpose, but in regular VGI traces, such as speed and distance.

Since our traces are not classified according to the transportation mode and do not have a uniform spatiotemporal resolution, an application of fix thresholds for speed and distance criteria is inefficient. Thus only something that we call “adaptive threshold” can be tested such as Box plot (Tukey et al., 1977). Theoretically speaking, criteria relying on adaptive thresholds should fit each individual trace better than criteria relying on fixed threshold. In addition, Box plot is a convenient solution, since it is non parametric. It displays variation in a sample without creating assumptions about statistical distribution of the sample. Box plot is usually used with whisker so that the boundaries of the box are first and third quartile, bottom and top respectively. Boundaries of whiskers can be defined in several different ways. However, we decided to use interquartile range (IQR) for that. More specifically, we define outliers as points out of the range defined by median +/-1.5 IQR, the solution known as Tukey boxplot (Frigge et al., 1989). The boundaries are calculated in as defined by equitation 3:

$$Bottom = Q_1 - 1.5(IQR)$$
$$Top = Q_3 + 1.5(IQR) \qquad (3)$$

Where Q_1 is a first quartile and Q_3 is third one

Figure 33 shows application of this method on a randomly chosen trace speed profile:

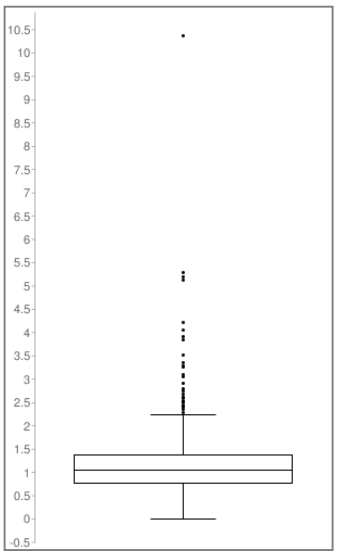


Figure 33 : Box plot of segments speed

In the Figure 33 outlying values of speed presented as dots are plotted. They are very concentrated close to the threshold, whereas the concentration decreases almost gradually while moving away from the threshold. In presented case median value of speed is 1.05 m/s and IQR is 0.6. First quartile is 0.77 m/s whereas the third is 1.37 m/s. Obtained threshold (top boundary) is 2.27 m/s.

As a result, 32 outliers were detected (4%). However, after visual checking, most of the points were not found as outliers despite having outlying speed values according to the presented method. Only five points having a speed value bigger than 5 m/s were confirmed outliers. At the same time, four points having regular speed values according to Box plot are outliers. Thus, many under and over estimations are produced by this method applied on speed as input parameter.

It is obvious that due to the non-uniform distribution of speed, the thresholds set in this way are not very efficient since they produce a lot of overestimations, even in case where dispersion is relatively small like a dispersion of speed in presented case (0.66 m/s). It could be explained by very rough terrain and changes of sport activity (e.g. walking to running) or even transportation mode (walking to cycling) during one single itinerary. This can be observed by looking at speed profile of the trace in question illustrated in Figure 34 that presents the speed profile of the trace. Red line is a value of median speed 1.05 m/s. Frequent and significant changes of speed, especially over median speed can be very easily observed.

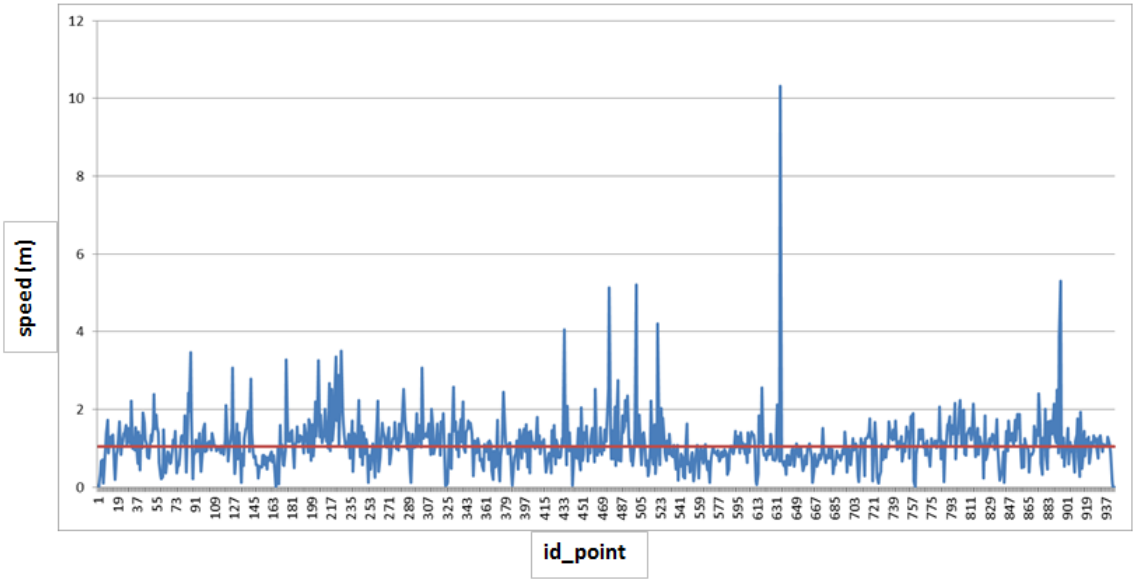


Figure 34 : Randomly chosen trace speed profile

When applied on distance, the results are very similar. Box plot of segments' distances values are presented in Figure 35 where we can observe outlying values of segments' distance presented as dots. They are also very concentrated close to the threshold, and gradually less concentrate while moving away from threshold. In presented case median value of distance is 17.9 m. First quartile is 12 m whereas the third is 26.2 m. Obtained threshold (top boundary) is 47.4 m. As a result, 37 outliers were detected (8%). After visual checking, all points detected as outliers by this method were not confirmed as outlying, even 8 points with segments' distance bigger than 70m. They are regular points with a long distance due to the signal lost. More importantly, a point that is a real outlier was not detected as outlying by Box plot. Like in the speed case, profile of distance presented below show how it is difficult to apply this method on very non-uniform spatial resolution traces.

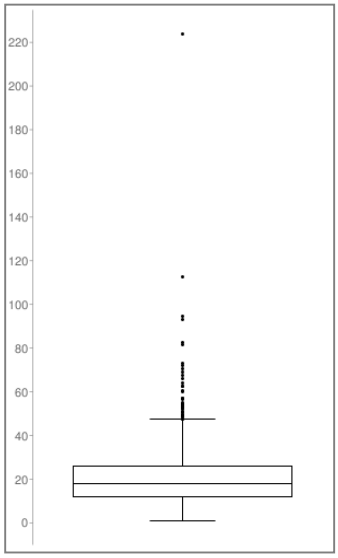


Figure 35 : Box plot of segments distance

Figure 36 illustrates the distance profile of the trace in question. Red line is a value of median segment distance 17.9 m. Frequent and significant changes of distance, especially over median distance are evident. The outlier point has id 456 and is below the median value.

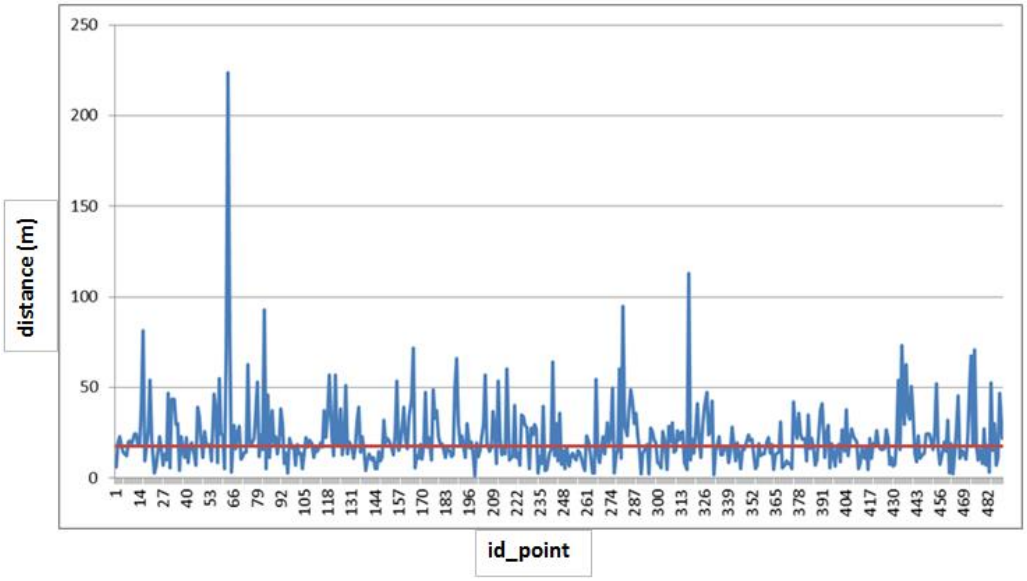


Figure 36 : Randomly chosen trace distance profile

Furthermore, it has to be taken into account that almost 2/3 of the traces are affected by missing timestamps, which considerably limits usability of speed as a criterion, since it cannot be applied on all traces.

Having considered above mentioned facts, four main questions arise:

- How to deal with incompleteness?
- How to determine relevant criteria?
- How to set thresholds for each criterion?
- How to make final decision?

To address listed issues, we propose an approach based on supervised machine learning techniques. Furthermore, since points considered as outliers according to our definition cause a wide range of geometric anomalies of a trace, and since the reasons for them are very heterogeneous (environmental, technical, behavior, etc.), the main goal of the approach is to model them by learning from the examples. This will allow us to extract most dominant features that sampled outlier points have in common, as well as links between outliers and most dominant indicators.

Rule based algorithm such as Repeated Incremental Pruning to Produce Error Reduction (RIPPER) are very convenient not only for characterization of outliers but also for prospective modifications or improvements of generated rules, which is not possible in other classification algorithms (Cohen 1995). The RIPPER algorithm was successfully applied in many studies e.g. (Mustiere, 2001). Let us notice that the choice of the best learning algorithm for our task is out of the scope of our work: this is a secondary point with regard to the good definition of the learning task and the description of examples. As RIPPER is known to be efficient and is able to deal with our examples (described by numeric, qualitative, and

possibly missing attributes), we chose it without deeper study. Quick experiments with other algorithms gave similar results.

Indicators

To formalize outliers' detection, we propose both intrinsic metrics and extrinsic indicators. The former are calculated from VGI traces only, such as distance, speed, direction and elevation between consecutive points. The later are calculated based on the analysis of the spatial context in which GPS points are recorded such as the type and density of forest, the slope and its curvature, the proximity of obstacles (e.g. cliffs, buildings, forest) and other features (e.g. river, lake, building). The choice of extrinsic indicators has been guided by the causes of GPS errors, as identified in the state of the art.

We propose proposed fifteen different indicators. First eight are intrinsic presented in Table 3.

Table 3 : Intrinsic indicators

Indicators	Description	Formula
AngleMean	Average value of 3 direction change	$(\alpha_{i-1} + \alpha_i + \alpha_{i+1})/3$
DistDiffN	Normalized distance	$(D_{i-1} - D_i) / (D_{i-1} + D_i)$
DistDiffMed	Relation between distance and median distance of a trace	$\frac{D_{i-1} + D_i}{2\text{Median(Trace)}}$
DistMean	Mean distance	$(D_{i-1} + D_i)/2$
SpeedDiffN	Normalized speed	$(V_{i-1} - V_i) / (V_{i-1} + V_i)$
SpeedMean	Mean speed	$(V_{i-1} + V_i)/2$
SpeedRate	Speed rate	$(V_{i-1} + V_i)/V_i$
DiffElev	Maximal height difference	$\max \lvert Z_{i+1} - Z_i, Z_i - Z_{i-1} \rvert$

AngleMean is evaluating a change of direction of a trace between consecutive segments. It is an important characteristic of outlaying points, since most of them make a significant direction change in a trace. Since a regular geometry of roads can have significant direction changes e.g. turnings, our indicator is designed to take into account not only a direction change of the point in focus but also changes of direction for the preceding and succeeding point in a trace. In this way the indicator is less sensitive on regular significant direction changes such as very strong road curves. Figure 37 shows how AngleMean is defined and computed. For the standing point i, indicator AngleMean is the mean angle of the three direction changes presented in red, i.e. on point i and on succeeding and preceding points.

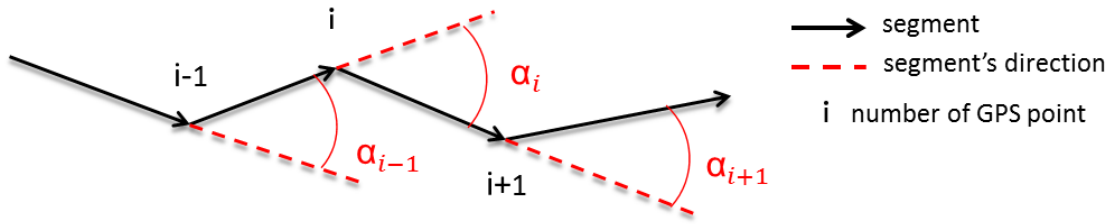


Figure 37 : AngleMean calculation

DistDiffN and SpeedDiffN are representing normalized values of distance and speed respectively based on preceding and succeeding segment. They can register sudden changes of speed and distance between consecutive points.

Similarly, SpeedMean and DistMean represent average values of speed and distance of preceding and succeeding segment. It allows detection of outlying speed and distance values.

Elements by means of which previous four indicators are calculated are illustrated in the figure below:

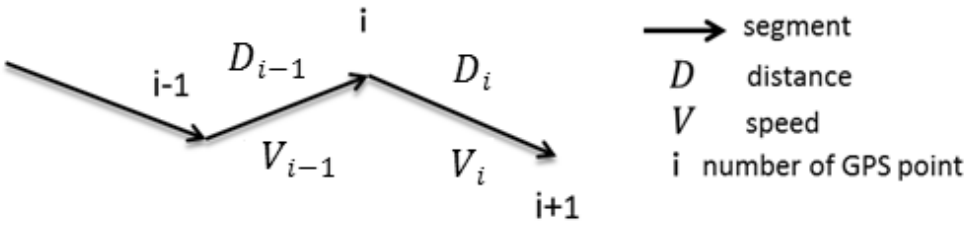


Figure 38 : Elements for calcuation of DistDiffN, SpeedDiffN, SpeedMean and DistMean

DistDiffMed is the ratio between the local segment lengths (average of lengths of preceding and succeeding segments) and the median length of segments of the trace. In this way spatial resolution of a given trace is taken into account and discrepancies from it are modelled.

SpeedRate represents the velocity change rate being proposed by Van Winden et al (2016). It is useful in discovering sudden changes of speed.

DiffEle takes maximal change of elevation based on preceding and succeeding point, as determined by Z values of GPS data.

Other seven extrinsic indicators are presented in Table 4.

Table 4 : Extrinsic indicators

DiffElevDTM	Correlation between elevation (GPS and DTM)	$\|Z_{DTM} - Zgps\|$
Slope	Gradient of line	$\tan(\Theta)$, $-90° < \Theta < 90°$
Obstacles	Proximity of obstacles	true if close to obstacles, false otherwise
Curvature	Convexity of slope	1/R
Vegetation	Type of forest	f (BDTopo® Landcover)
CanopyCover	Point in the forest or not	Boolean (true/false)
InBuildingWater	If the point is in building or water	Boolean (true/false)

DiffEleDtm is the first extrinsic indicator and evaluates the accuracy of GPS elevation on the point in focus. This indicator is based on the idea that significant errors in elevation are usually linked with significant errors in 2D position (Heselton, 1998). It requires using a precise DTM of the area to evaluate it: altitude of GPS data is compared to the altitude at the same (X, Y) position on the DTM.

Slope is a gradient of line and is calculated based on a referential DTM.

The 'obstacles' indicator estimates multi-path effect, one of the most influencing factors in GPS accuracy. It takes into account proximity of obstacles that can be most influencing in our test zone environment type, that is: building and forest. For that purpose, layers Buildings and Land cover from BDTopo® were used. The threshold for defining close features is determined by using a propagation of errors formula taking into account the average reported accuracy of smartphone GPS and the accuracy of BDTopo®, like presented in equitation 4.

$$\partial^2 = \partial_{AGPS}^2 + \partial_{BDTopo}^2 \qquad (4)$$

Where ∂ is a total imprecision of calculation based on smartphone GPS and BDTopo® data that is a size of a buffer that is created around buildings and forest. The calculated size is 11m. The indicator is Boolean, which means that GPS points close to obstacles have a value 'true' whereas GPS points outside buffer zone have values 'false'. This could be refined with distance thresholds related to the height of obstacles, if the information is available.

CanopyCover reports if the GPS point is in the forest or not, according to a land cover map.

Vegetation represents the land cover at the position of the GPS point, as defined by a land cover map.

Owing to imprecision of GPS, some points may be miss-positioned in water or building surfaces according to a reference database, which reflects an inconsistency with a normal hiking/cycling activities, and is then a clue for an outlier. Thus, an indicator carrying this type of information was created and named InBuildingWater. However, presence of bridge can induce a lot of points in the surrounding water surface due to the imprecision of GPS or due to the imprecision of the bridge in BDTopo®. Thus, a tolerance around bridges based on accuracy of BDTopo® should be taken into account and points lying

within this tolerance are not considered as inconsistent. The tolerance in this case is the same as introduced in equitation 4.

Training area for the learning task

In order to apply the supervised machine learning techniques, training area has to be defined. We chose a sampling zone that represents as much as possible the state and heterogeneities of the entire test area and the entire pattern of GPS points. The representatively of the test area has been evaluated by comparing the percentage of missing attributes (timestamps and elevation) of sampled points compared to the total percentage of missing attributes. Those percentages differed for only 2%. In total 2342 points were sampled and manually classified as outliers (3%) or regular points (97%). Sampling zone as well as sampled points are presented in Figure 39.

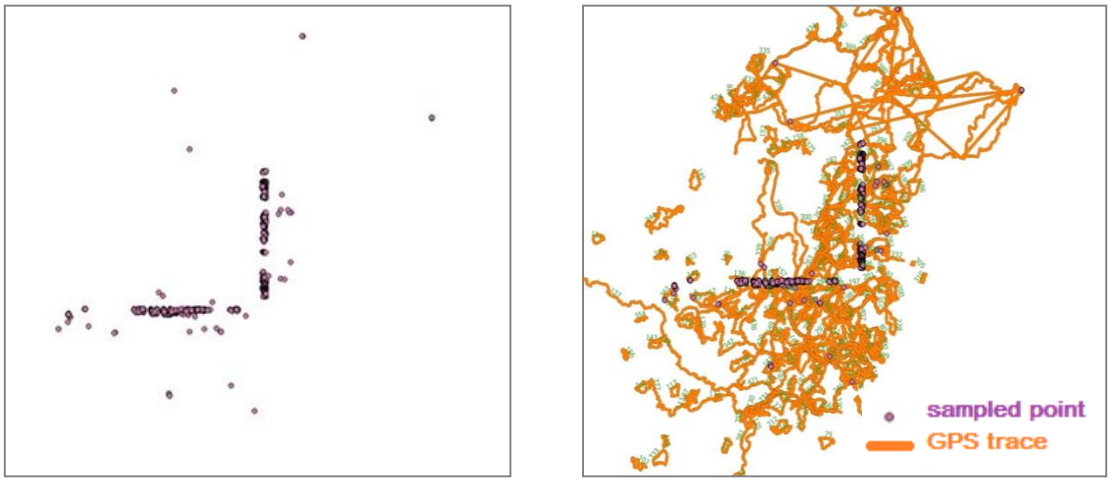


Figure 39 : Sampling zone and sampled points for the supervised machine learning

In the sampling zone, 35 points were interactively classified as outliers. Compared to 2,265 of points classified as regular points, it is obvious the learning dataset is imbalanced (1 outlier for 65 regular points). Applying machine learning techniques on imbalanced dataset was reported as an issue many times (Fawcett and Provost, 1996; Lewis and Catlett, 1994; Kubat et al., 1998). Imbalanced datasets are encountered when there is a predominant class in a phenomenon, e.g. healthy people vs. rare diseases detection. There are two most usual ways to address this issue. The first one is assigning distinct costs to training examples (Pazzani et al., 1994; Domingos, 1999). The second one is re-sampling the original dataset (Japkowicz, 2000; Ling & Li, 1998, Chawla et al., 2000): rare cases may be over-sampled, or majority cases may be under-sampled. The two approaches may be combined. Other authors reported that learning based on imbalanced data not a significant issue when the classes are diametrically opposed according to their characteristics, with no class overlapping (Batista et al., 2004), which is the case in our task.

In our case, we decided adding examples of outliers beyond sampling zone. By this way, we lose the statistical representatively of the examples, but we gain more examples of outliers with a reasonable interactive effort (we only interactively add examples of outliers and not examples of regular points). The number of outliers in the training dataset was thus more than doubled to 77 examples.

Results of the approach will be presented in 1.4.1.3.

1.4.1.4 Low accuracy points detection

As it has been already presented, VGI traces usually do not come with metadata about its measurement process and accuracy. More specifically, very important information associated to the precision such as PDOP or number of visible satellites is missing. The goal of this subsection is to present an approach for assessing VGI traces accuracy (low accuracy vs good accuracy points), based on supervised machine learning using relevant intrinsic and extrinsic indicators, similarly to the approach presented for outlier's detection.

Description of method

At the hearth of the approach is supervised rule based machine learning classification, the same applied in detection of outliers, and the same learning algorithm RIPPER is used. Points were sampled and classified according to their spatial accuracy which was calculated based on referential roads and the accuracy of smartphone GPS according to the equation (4). Final value obtained was 11 meters. Therefore, all points with distance to corresponding referential roads bigger than 11 meters were classified as low accuracy points, whereas others as good accuracy points. However, this time three strategies were applied: manually generated sampling points, automatically generated sampling points and semi-automatically generated sampling points.

In manually generated sampling points 2,008 points were manually taken as examples and also manually classified as good (76%) and low accuracy (24%) points.

Matching and measurements were done manually in order to minimize negative effects of automatic mismatching. An example of low accuracy points according to this criterion is illustrated in Figure 40.

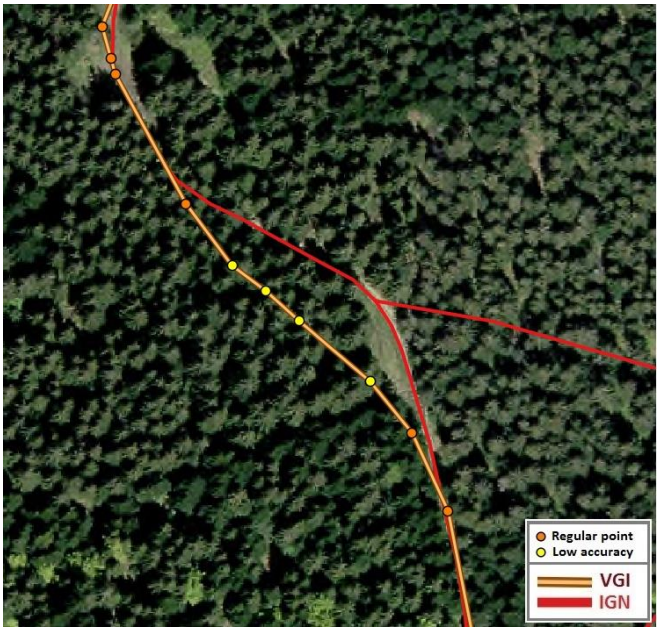


Figure 40 : An example of low accuracy points

In automatically generated sampling points, 14,891 were taken automatically, from which 10,787 were classified as good accuracy points (73%) and 4,034 as low accuracy points (27%) automatically as well.

Semi-automatic sampling was done on 7,745 points. The points were selected automatically, but then filtered depending on the fact if they belong to new or existing roads. This is important, since in pure automatic generated sampling all points belonging to missing roads are categorized as low accuracy since the distance to closest referential road is certainly bigger than 11m. From 7,745 points, 5,358 points were classified as good accuracy points (70%), whereas the rest were classified as low accuracy points (30%).

In both automatic and semi-automatic generated sampling, points were selected by buffers using the approach of Goodchild and Hunter (1997).

Indicators

The indicators proposed for the detection of outliers in the section 2 are also used in this classification.

Results of the approach are shown in subsection 1.4.2.4.

1.4.2 Results of the data quality assessment

This subsection is devoted to presentation of result of the approach on test data described in 1.4.1, step by step.

1.4.2.1 Pre-processing filtering

As a result of preprocessing filtering, 4,587 redundant points were eliminated (2%). Seven points with "negative speed" values were also eliminated.

Zero values of elevation were found 1,135 times, whereas 2 traces have the same values of elevation for each point. Elevation of all those point was set to NULL. Similarly, two traces had a constant timestamp for all points, and this was set to NULL.

1.4.2.2 Detection of secondary human behaviour

Parameters that we used in applying the method for SHB detection are:

Step 1: buffer size = 0.0001m
Step 3a and 4a: N=200 square meters
Step 4: buffer size 0.1m
Step 7: S = 0.13

After applying the method, in total 11746 (4%) points were detected and subsequently eliminated as results of SHB. The algorithm was run until no results of SHB remained. Three iterations have been necessary to do rich the convergence: 9,006; 2,740 and 475 points were detected within first, second and third iteration respectively. This confirms the need for keeping the process iterative. Four percent of points (among all points) detected seems globally reasonable, if one considers it as a very rough indicator of percentage of time devoted to stops during activities. In his PhD research Colin Kerouanton (on-going thesis) classified stops according to their duration concluded that 75% of stops are up to 15 minutes, whereas the rest are longer.

The method is successfully applied on various types of SHB, from simple to very complex. Validation of the results was done on randomly selected 14 traces (265km). In total, 204 SHB were found by the method, 3 falsely, whereas 16 SHB were not detected. Thus, both precision and recall were found high, 98% and 93% respectively. Practical results will be presented in the following figures.

Figure 41 shows application of the method on simple SHB.

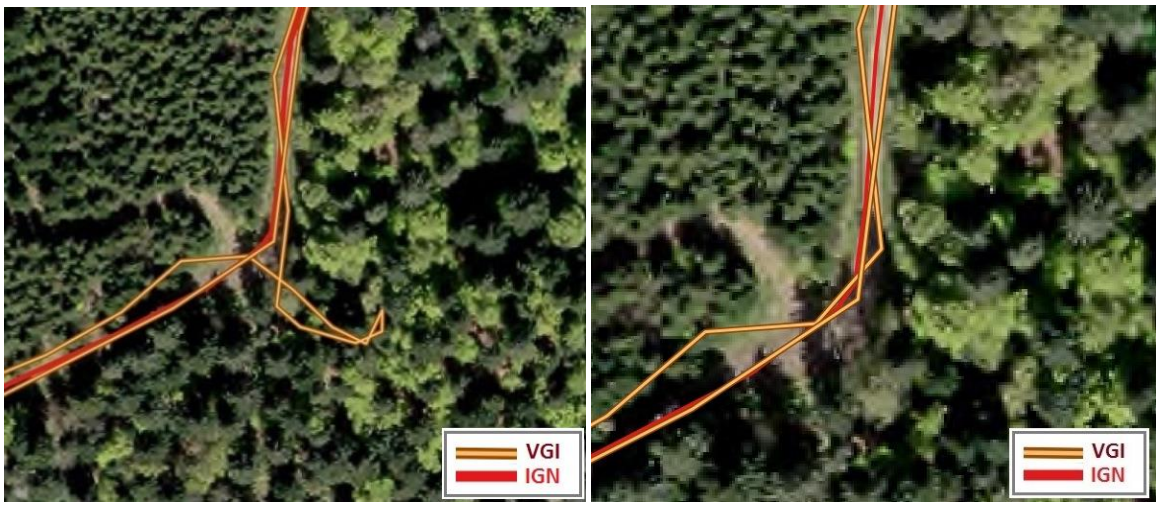


Figure 41 : A simple geometry VGI trace before (on the left) and after filtering of secondary human behavior (on the right)

Figure 42 illustrates an example of application of method on a complex secondary human behavior that resulted in a huge loop of segments.

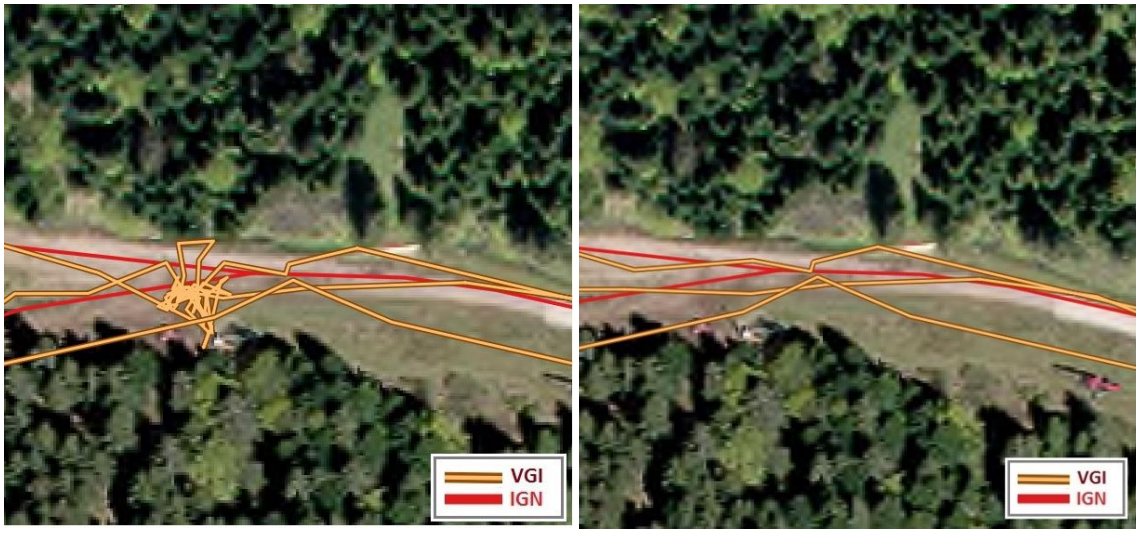


Figure 42 : A complex geometry VGI trace before and after filtering of secondary human behavior

Effects of secondary human behavior are especially frequent in urban areas due to the various activities conducted there. Start of an itinerary is very often from an indoor place or in front of a building which all makes geometric loops in the traces. Figure 43 illustrates the performance of our approach in one of such examples:

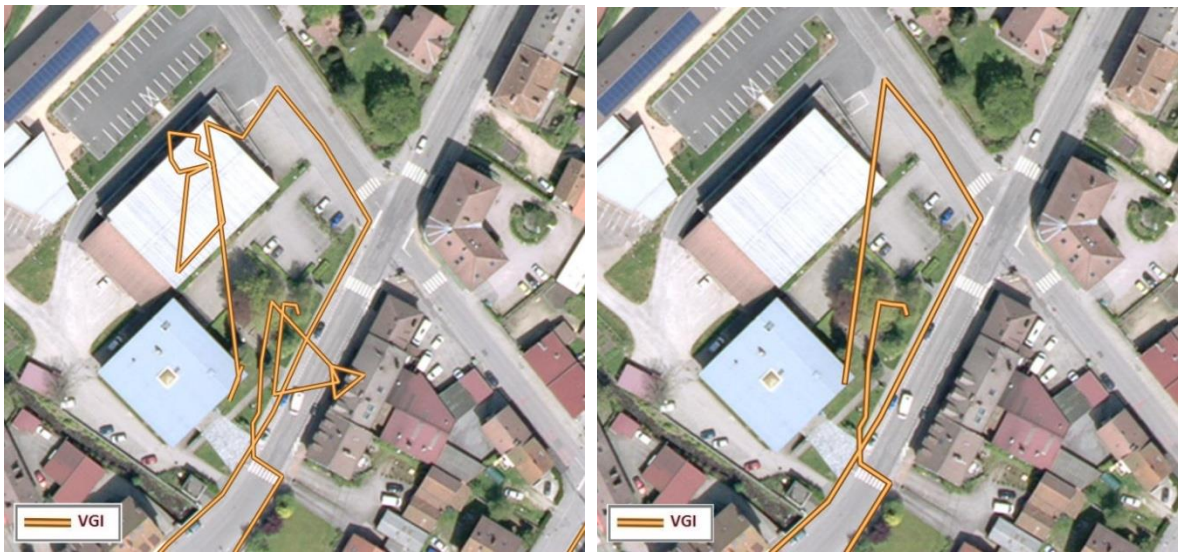


Figure 43 : Secondary human behavior in urban areas before (on the left) and after filtering (on the right)

In the Figure 44 we can see the situation where an interesting point pointed by blue arrow (seems to be an informal parking spot) is a reason for leaving the main itinerary.

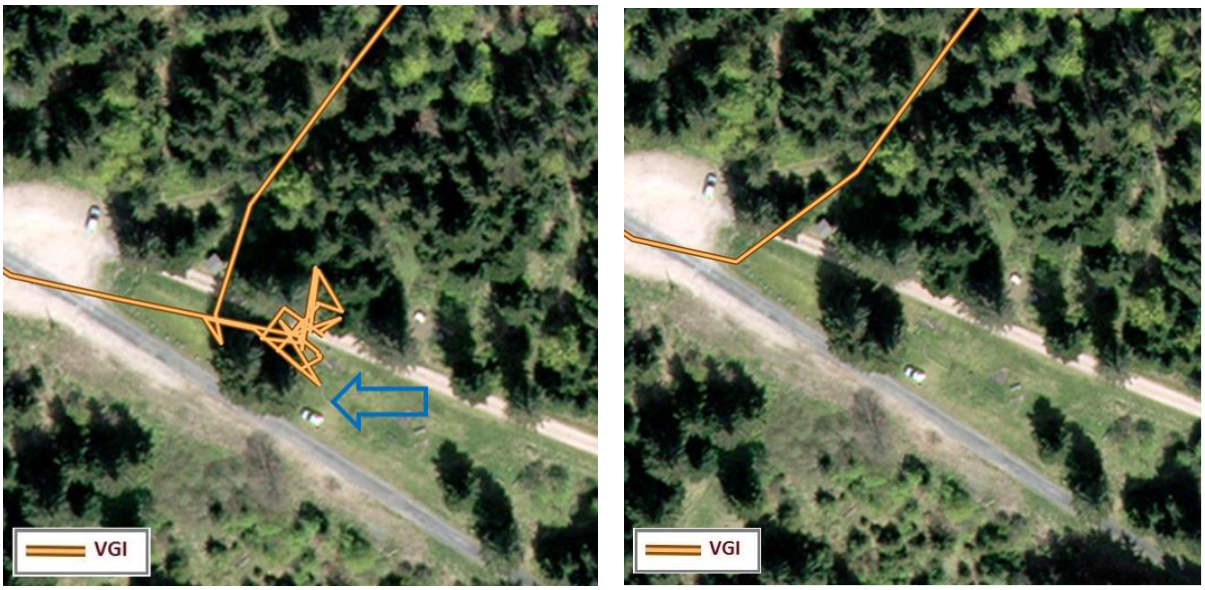


Figure 44 : Secondary human behavior as a result of visiting an interesting point

However, in some cases it is not possible to fully eliminate effects of secondary human behavior. In those cases, the polygons created based on secondary human behavior points are very narrow. Due to that, the presented ratio between the surface and its corresponding MBC is very small, under the threshold N for detection of round-trip traces set on 0.13. Thus, that part of the loop was not recognized as a result of secondary human behavior. Some cases are presented in the Figure 45.

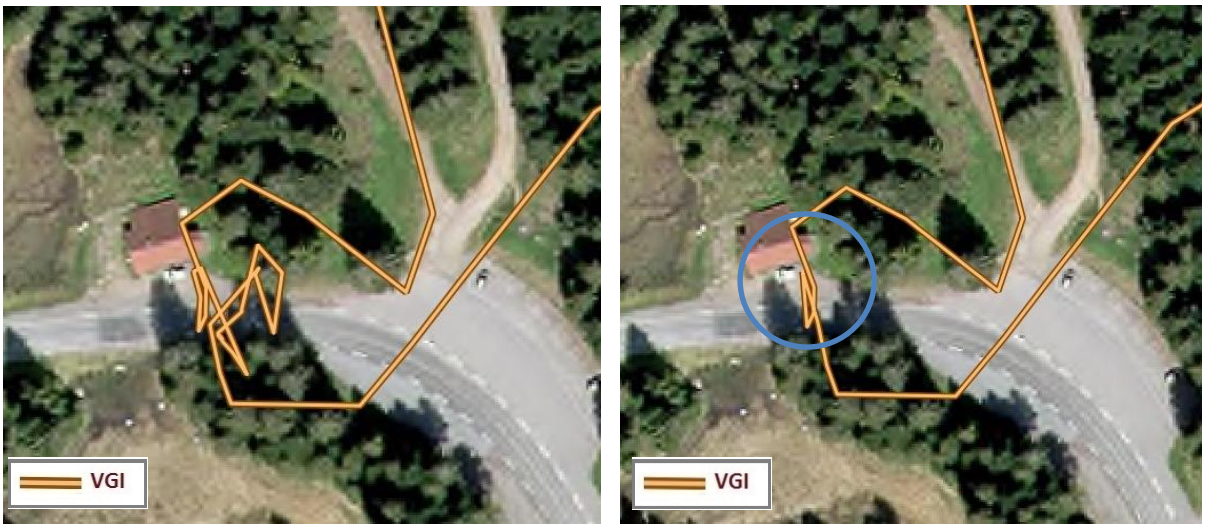


Figure 45 : Limitations of the method

Figure 45 shows the case where the part of a loop labeled by blue cycle remained in trace geometry even if it is a result of secondary human behavior. Due to this limitation, what we can see in the presented figure is that in some very complex cases, even after elimination of SHB the geometry of the trace is still dissatisfying.

However, not always this case is an issue in application of the method. If the overlapping of the polygons created based on SHB is big, even the points composing polygons with N<0.13 will be deleted since they are already within bigger polygon that is detected as SHB. It could be also considered as a conflict situation, since according to main condition S<0.13 they are not recognized as SHB. In our opinion the fact that polygons are overlapping is actually one reason more to consider it as SHB. That is why in such situation, we are giving a priority to greater polygons. This situation is shown in the Figure 46.

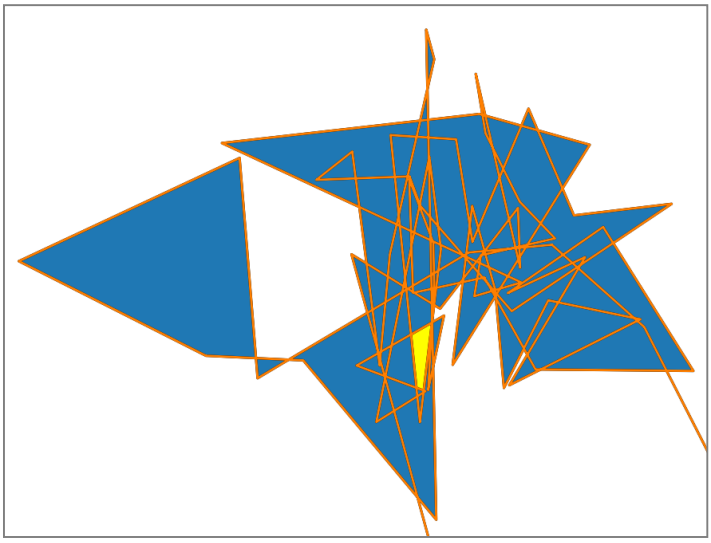


Figure 46 : Conflict situation between secondary human behavior polygons

Figure 46 shows polygons (in blue) created between segments that are results of SHB. Polygon in yellow does not fulfil the above mentioned condition to be eliminated, however it is completely within the

bigger polygon that fulfills the condition. Thus, according to our strategy points composing the yellow polygon will be also eliminated since they are already contained by the polygon that fulfils the condition.

Overall, the results are satisfying since the significant majority of secondary human behaviors were eliminated. The geometry of traces was improved, in most of the cases very much, while in some specific cases less due to the presence of very narrow polygons. Lack of timestamps and a need to detect round-trip traces only by geometric approach caused underestimations, like one presented in Figure 45.

1.4.2.3 Detection of outliers

The learning process was conducted in WEKA software package using RIPPER algorithm proposed by Cohen (1995), through its java implementation is JRIP. Evaluation of results by means of 10 fold cross validation is presented in Table 5.

Table 5 : Confusion matrix

	Predicted		
Actual class		Outlier	Not an outlier
	Outlier	61	16
	Not an outlier	16	2248

The number of correctly classified outliers is almost four times bigger than the number of misclassified, while overestimation and underestimation are balanced. Precision and recall are the same, 79%, while F_1 measure is 0.79. In addition, a global precision of the approach calculated from the matrix is 98%. The results confirm the performance of the approach and generated rules, despite the presence and random distribution of missing attributes.

In total, five rules and thresholds were generated and presented below:

DistDiffMed >= 1.05 and AngleMean >= 87.54 => outlier or

AngleMean >= 71.25 and SpeedRate >= 1.50 => outlier or

AngleMean >= 74.80 and DistDiffN <= 0.21 => outlier or

AngleMean >= 83.15 and SpeedRate <= 0.85 => outlier or

AngleMean >= 56.43 and DistMean >= 8847.31 => outlier

We can observe five different rules combining five different indicators which confirm that detection of outliers in VGI traces is a complex task that cannot be solved through a single threshold or only one indicator.

One important finding from these results is that only intrinsic indicators were recognized as efficient in outliers' detection. In other words, they were found most characterizing for differentiating between outliers and regular points. This may be counterintuitive regarding the state of the art on the effect of external condition and particularly land cover. However, this may be explained by the lack of accuracy and not sufficient resolution of referential data used (DTM and Land Cover map).

Concerning the influence of external (environmental) factors, it is obvious that based on the generated rules, the impact and importance of the external factors on detection of outliers were not discovered. It is likely that a link between them exists, however it is difficult to be detected and modelled in VGI traces due to their heterogeneity, and a lack of two very important information such as accuracy of GPS device and the quality of GPS signal.

For example, the same obstacle produces a multipath and subsequently outlier while collecting data by low precision GPS, whereas it is not a case for high precision GPS under the same conditions. The same situation is with canopy cover. More precise device using better quality signal can produce a trace without outliers even in closed coniferous forest, while lower precision device supplied with low quality signal would cause outliers even in an open area.

Among internal criteria used, AngleMean proves to be the most important indicator, which is quite reasonable since geometric anomalies of a trace have significant direction changes.

The generated rules are applied one by one on unclassified points from test area. As a result, 9,303 points (3%) were detected as outliers.

As we observed in the Table 5, 16 GPS points were classified as outliers whereas they are regular points. The same number of outliers was misclassified as regular points. The first are known as false negative results whereas the latter as false positive. By analysing false positive and false-negative results with sampling point's dataset we discovered that all false positive outliers are result of side-effect cause by AngleMean indicator values. It was expected since outliers affect significantly intrinsic indicators of neighbouring points in a trace, especially direction changes represented by AngleMean indicator. A typical example of false positive outlier is pointed by blue arrow in Figure 47.

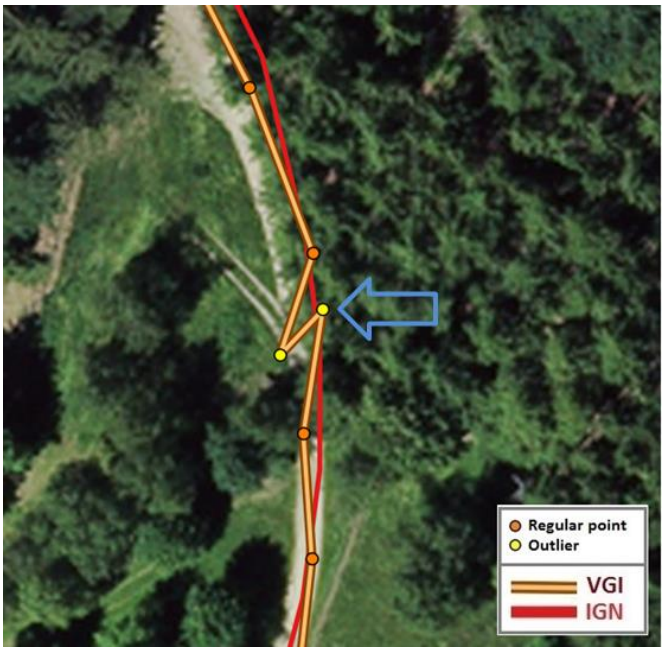


Figure 47 : Side effect – false positive result

False negative results are mainly caused by sensibility of thresholds. The analyses of them will be presented in the next subsection: Sensibility of thresholds.

By analyzing generated rules, one by one, the conclusions are as follows:

Rule 1 deals with outliers having very sharp direction change, close to 90° and corresponding distances slightly bigger than mean value of spatial resolution of the full trace. Figure 48 illustrates an outlier detected by rule 1. Let us notice that this rule is applicable on traces having timestamps as well as traces having missing timestamps. In total 1461 points were detected by rule 1.

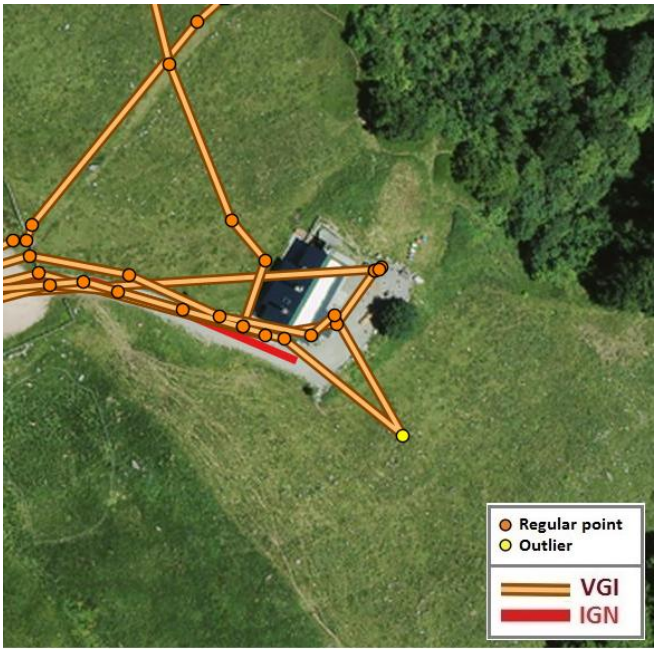


Figure 48 : An outlier detected by the rule 1

Rule 2 recognizes outlying points having significant direction as well as speed changes, as illustrated in the figure below. As a result of application of rule 2, 2,076 points were detected as outliers. Figure 49 illustrates an outlier detected by rule 2.

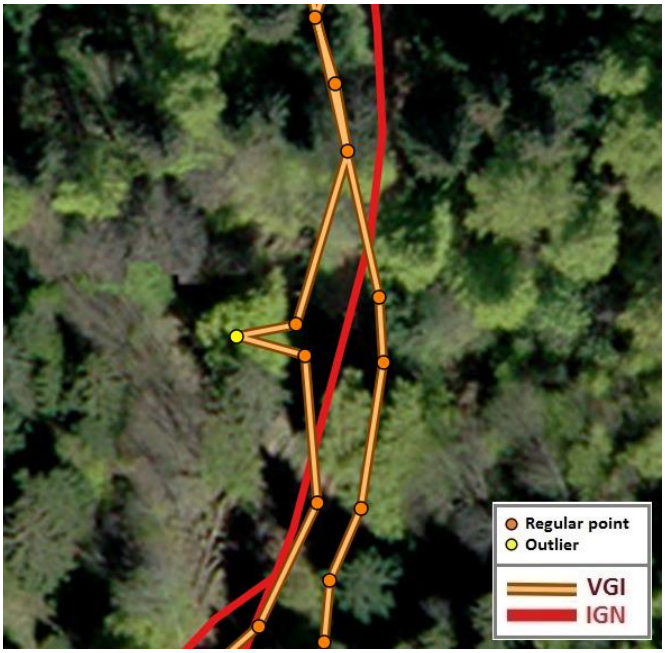


Figure 49 : An outlier detected by the rule 2

Rule 3 detects sudden deformations of a trace's geometry, sharp in terms of direction but not far away from main axis of the trace. The rule is applicable on traces having timestamps as well as traces having missing timestamps. In total, 3,622 points were detected by means of this rule. One example is presented in the Figure 50.

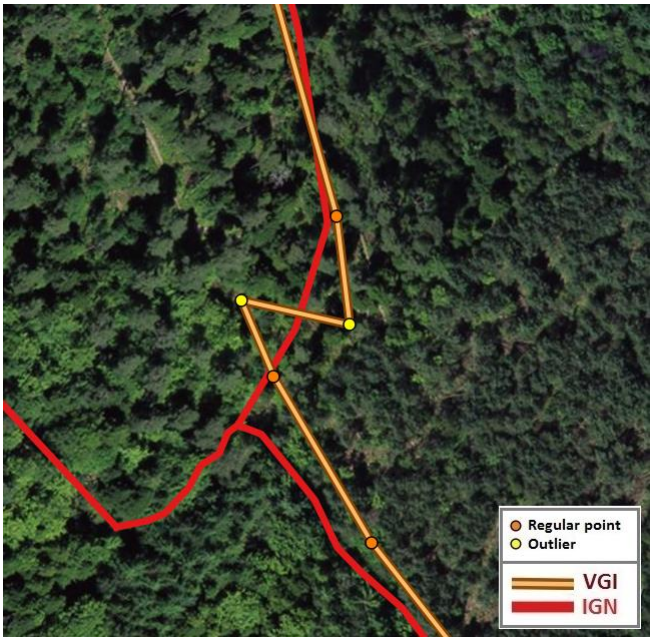


Figure 50 : An outlier detected by the rule 3

Rule 4 deals with phenomenon similar to the one detected by rule 3, small but sudden and sharp geometry changes, but discovers them more successfully in the traces without missing timestamps as illustrated in Figure 51. Finally, 2,135 points were detected as outliers by rule 4.

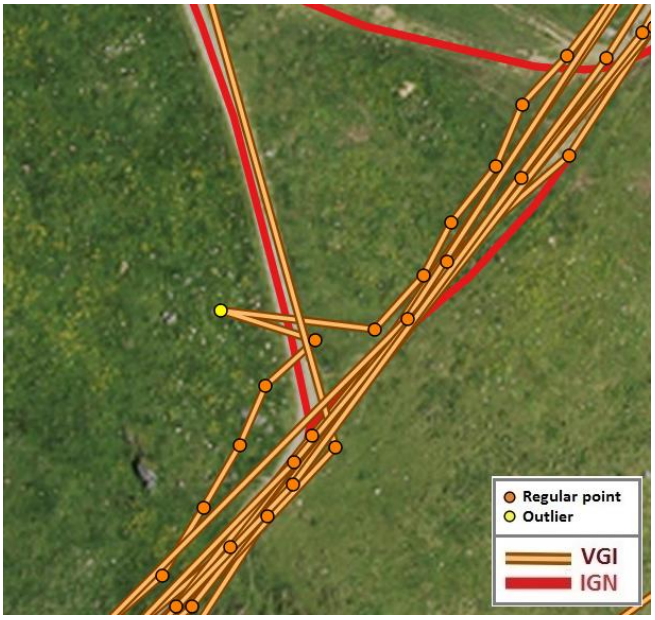


Figure 51 : An outlier detected by the rule 4

Rule 5 is designated to, outlying points very far away from main axes of the trace. They are not frequent but very harmful for geometry of a trace like illustrated inFigure 52. As a result, 9 very long distance points were discovered.

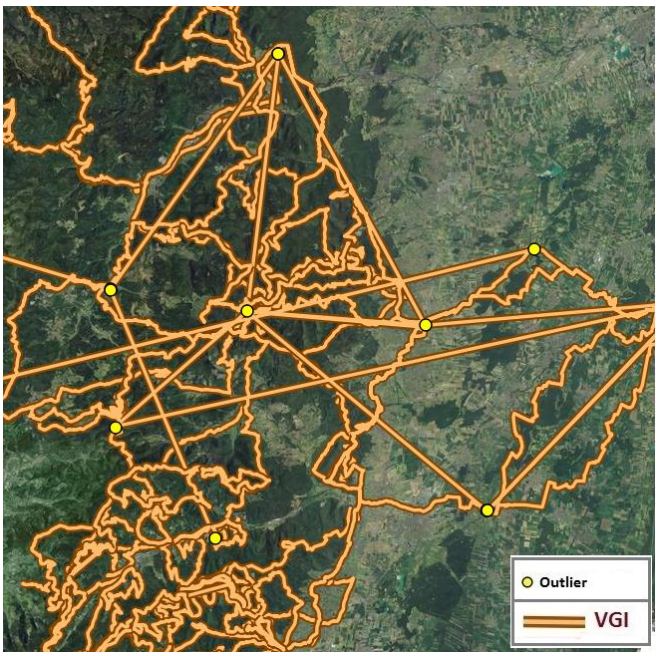


Figure 52 : Outliers detected by the rule 5

Figure 52 outlying points are located at a distance more than 8km from preceding points, which results in very significant geometry deviations as it can be observed.

There is a specific situation where the rules do not perform properly, causing overestimations: this concerns traces with very high sinuosity and low spatial resolution such as illustrated in Figure 53.

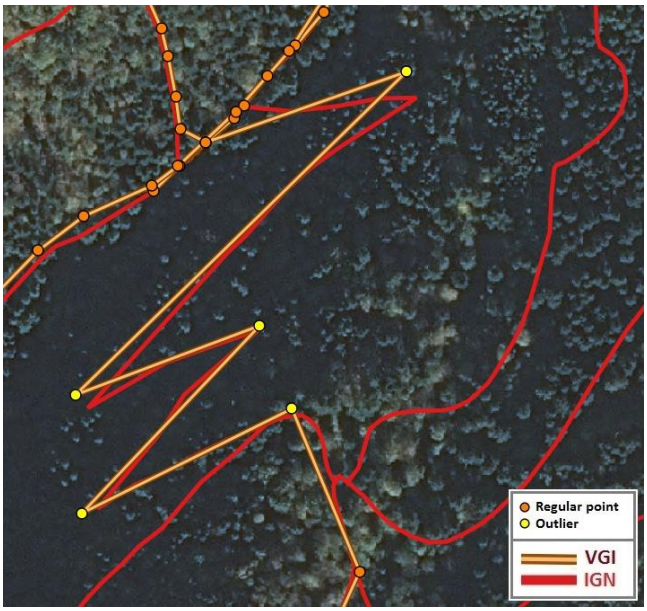


Figure 53 : Misclassified outliers

All five GPS points in yellow of the trace, that actually fits authoritative data, are identified as outliers. Such geometries of roads are not frequent (11 cases in our test area).

A case with high sinuosity of referential road but good spatial resolution of VGI traces is threated significantly better like in the example below, were only one point were misclassified as outlier.

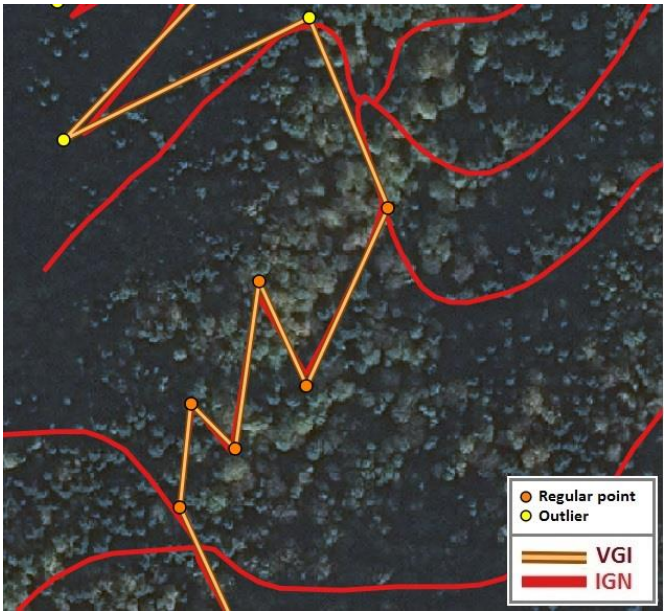


Figure 54 : Misclassified outliers

Sensibility of thresholds

Detailed analysis of false negative results revealed that they mostly appeared due to the sensibility of thresholds. This means that they remained under the radar for sub decimal values of their indicators. For example, a change of threshold for AngleMean in criterion 1 for only 0.1° results with 7 points less detected when the threshold is made higher, whereas 5 more points are detected when the threshold was reduced for the same value.

Figure 55 shows outlying points that were not detected due to slight difference between their AngleMean values and the threshold. All differences are under 1°. As usual, outlying points are presented in yellow, whereas orange points close to them (within blue cycle) are false negative results.

Figure 55 : Side effect – false negative results

Because of the previous considerations, we tested thresholds in rule 1, i.e. for AngleMean and DistDiffMed criteria. The goal is to observe changes in the number of points detected as outliers while thresholds are slightly changed. Slight increase and decrease of thresholds are tested. Criterion AngleMean is set as less strict by decreasing its value, whereas DistDiffMed is tested by increasing its value.

In: Sensibility of threshold diagram Figure 56 we can observe a significant rise of number of points detected as outliers while a threshold of AngleMean is reduced for only 0.5°. As a result of angle reduction of 4°, more than 200 new points were recognized as outliers.

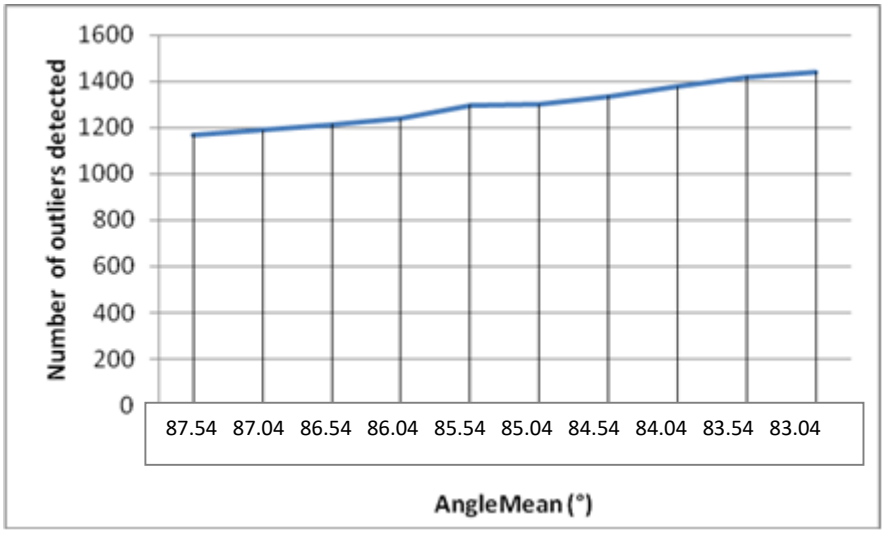


Figure 56 : Sensibility of threshold diagram

Similarly, the examination of other indicator DistDiffMed in rule 1 showed sharp decrease of detected outliers while its vales were reduced for 0.05 as seen in the Figure 57.

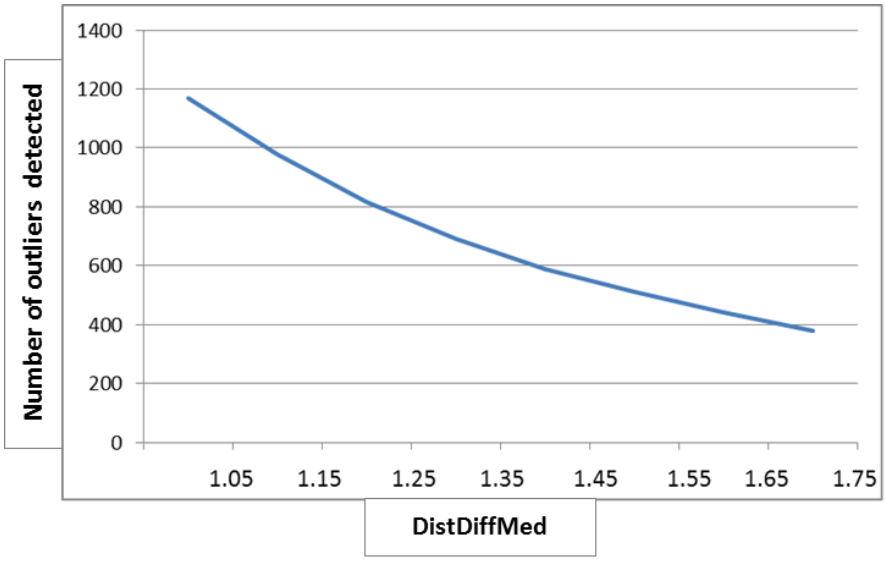


Figure 57 : Sensibility of threshold diagram

Once again it confirmed an issue of setting thresholds very challenging. Thus, the uncertainty of results of the approach is the highest in a tight zone above and under thresholds.

To summarize, in this subsection we presented an approach for formalization and detection of the outliers despite the high heterogeneity of VGI traces and dominantly random influences of external factors on their quality. Even if the approach correctly classifies most outliers, it is sometime mistaken (precision and recall approximately at 80%) and this may be improved by other approaches. However, the approach has the important strength of simplicity, which offers the following advantages:

- Only intrinsic indicators are used: no extra resources such as DTM, Land Cover are required
- It can be successfully applied on non-classified traces according to the transport mode (e.g. walking, cycling)
- It can be applied even if few data exist following the same object in the real world
- It can be applied even on traces with missing attributes

1.4.2.4 Low accuracy points detection

Like for the detection of outliers, the learning process for determining low/good accuracy points was conducted in WEKA software package using RIPPER algorithm proposed by Cohen (1995). Evaluations of results by means of 10-fold cross validation for the three strategies are presented in following confusion matrixes.

First, the results of strategy with manually generated sampling points are presented in Table 6.

Table 6 : Confusion matrix manually generated sampling points

	Predicted		
Actual class		Low accuracy point	Good accuracy point
	Low accuracy point	263	283
	Good accuracy point	158	1304

The global precision of the approach calculated from the matrix is 79%. In comparison, an approach that would classify all points as good accuracy points would have a precision of 75%. This is only a small improvement. The precision of classification of low accuracy points is 48% and recall is 62%. F_1measure is 0.54. Seven rules were generated.

Second, the results of strategy with automatically generated sampling are presented in Table 7.

Table 7 : Confusion matrix automatically generated sampling points

	Predicted		
Actual class		Low accuracy point	Good accuracy point
	Low accuracy point	1,626	2,408
	Good accuracy point	658	10,129

The global precision of the approach calculated from the matrix is 79%. The precision of classification of low accuracy points is 40% and recall is 71%. F_1measure is 0.51. Twenty-three rules were generated.

Third, the results of strategy with semi-automatic generated sampling are presented in Table 8.

Table 8 : Confusion matrix semi- automatic generated sampling points

	Predicted		
Actual class		Low accuracy point	Good accuracy point
	Low accuracy point	1,678	1,009
	Good accuracy point	449	4,609

The global precision of the approach calculated from the matrix is 81%. The precision of classification of low accuracy points is 79% and recall is 62%. F_1measure is 0.69. Twenty-seven rules were generated.

After analyses of the results, rules, and studious visual checking of points detected as low accurate, most important advantages and disadvantages of the strategies have been distinguished.

First, strategy with automatically generated sampling points makes the method fully automatic, provides big pattern and saves time. However, it was trained on the data where all new roads were treated as low accurate points, so the rules generated are likely to falsely detect most of new road as low accurate. And indeed, the precision of this strategy is the lowest.

Second, strategy with semi-automatic generated sampling points designed as a compromise between time saving and precision of the approach has the best scores. Nevertheless, after visual checking it was discovered that the rules for detection of low accuracy points are very depending on the sampling pattern. Probably traces used in sampling were not so heterogeneous (recorded by very similar devices in similar conditions) hence, the rules have very local character and did not perform successfully out of the sampling zone. E.g. one of the most successful rules for detecting low accuracy points is "(Slope >= 14.73826) and (SpeedMean >= 1.57m/s) and (DiffElevDTM <= 30.65m)." Having been tested out of sampling zone the rule completely underperformed giving dozens of misclassified points. Even just by looking at the rule, the values of indicators for low accuracy points seems strange compared to the findings described in Section 1.2.4.1 Sources of errors of GPS traces. E.g. it would be more logical that low accuracy points are those having DiffElevDTM higher than 30.65m than lower.

Finally, strategy with manually generated sampling points kept its performance out of sampling zone. This strategy gave most generic rules, which is most important for such type of the approach. It is most time-consuming, but the errors in sampling are minimized. Even though, the results (precision and recall) are not high, they are satisfying for such type of the approach taking into account the conditions and the data. Therefore, we decided to apply this strategy in our work. As a result, only the rules and performance of this strategy will be presented and discussed.

Generally, the results show only a relative success of the approach, which is less successful that the same approach applied to the detection of outlier. The question arose, why? On explanation of that is certainly that differences between low and good accuracy points are considerably smaller than differences between outliers and regular points. Outliers' characteristics are very remarkable compared to characteristics of regular points, whereas characteristics of accuracy points are not crucially different from characteristics of good accuracy points. This puts the learning algorithm in a challenging situation, particularity in absence of essential quality indicators such as accuracy of GPS device and signal quality. This shows that relying on intrinsic metric and geometric characteristics in this situation cannot provide remarkable results for evaluating a priori the precision of GPS data. Better external data (land cover, DTM) may improve the result, but certainly not significantly. However, the approach detects approximately half of the low accuracy points. Even if not perfect, we believe this may be useful information in our context, if we only consider it carefully as approximate clues about the accuracy.

In total, the following seven rules are learnt:

AngleMean>45.05 and CanopyCover= YES and Slope>19.2 => low accuracy point or

AngleMean >= 23°.45 and SpeedMean <= 1.21m/s and DiffElevDTM >= 14m => low accuracy point or

AngleMean >= 42°.63 and DistDiffMed <= 0.62 => low accuracy point or

DistMean >= 45.37m and DiffElevDTM <= 4.83m and CanopyCover = YES and Slope <= 17.55 and DistDiffMed <= 1.81 => low accuracy point or

AngleMean >= 21.145 and DistMean >= 116.34m => low accuracy point or

Slope >= 25.18 and DistMean <= 11.164266 => low accuracy point or

CanopyCover= YES and Slope>31.2 and AngleMean>21°

Those rules combine seven different indicators. Intrinsic indicators remain dominant (4/7 indicators), like for the outlier's detection. However, three extrinsic indicators were recognized as relevant and efficient in low accuracy point's detection: CanopyCover, Slope and DiffElevDTM. This is especially interesting since those two factors (canopy cover and relief) were reported as most influencing on GPS data quality in presented state of the art. This leads to the conclusion that effects of external factors on VGI traces are really random and mostly unpredictable except two strongest, canopy and relief.

Visual results of detection of low accuracy points are presented in the following figures, rule by rule.

First example in Figure 58 illustrates low accuracy points detected in deep forest having significant slope between each other (bigger than 19.2%) and having considerable direction change (bigger than 44°).

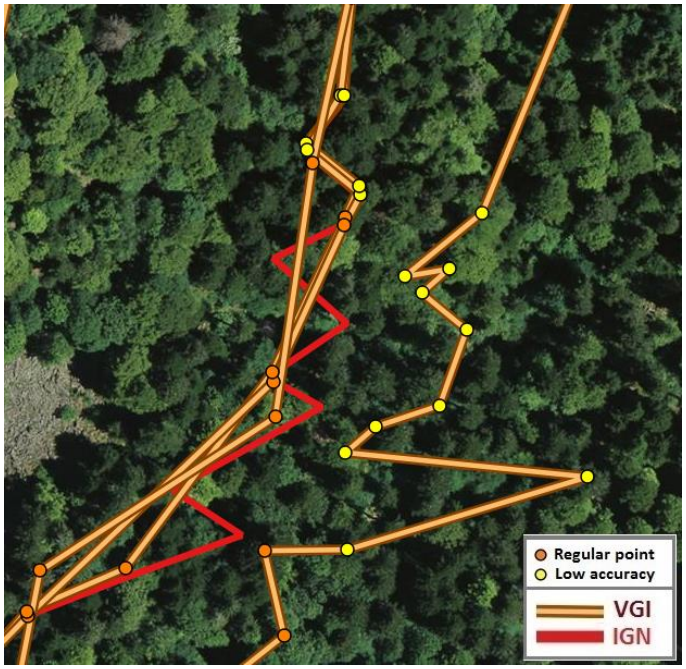


Figure 58 : Low accuracy points detected – rule1

Second example in Figure 59 shows low accuracy points detected by means of direction change, speed and low accuracy of their elevation compared to DTM.

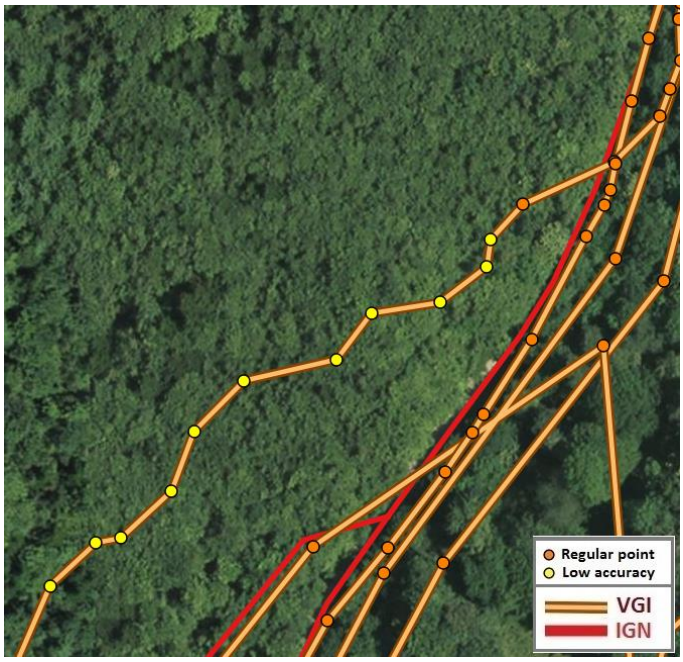


Figure 59 : Low accuracy points detected – rule2

In third example in Figure 60 low accuracy points were detected as a points with significant direction changes (AngleMean> 42°.63) having distance between them much lower than median distance of the trace they belong to (DistDiffMed <0.62).

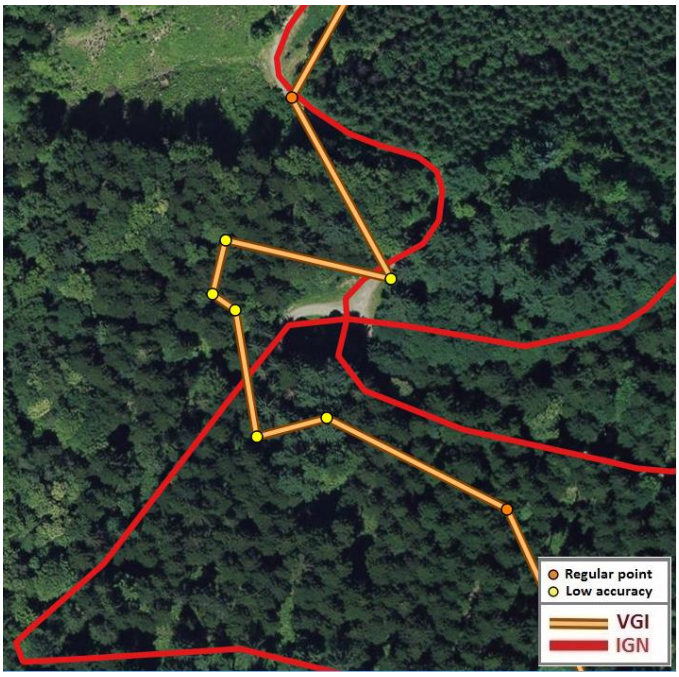


Figure 60 : Low accuracy points detected – rule 3

Forth rule Figure 61 is most complex and formalizes low accuracy points relying on five different indicators, three intrinsic and two extrinsic.

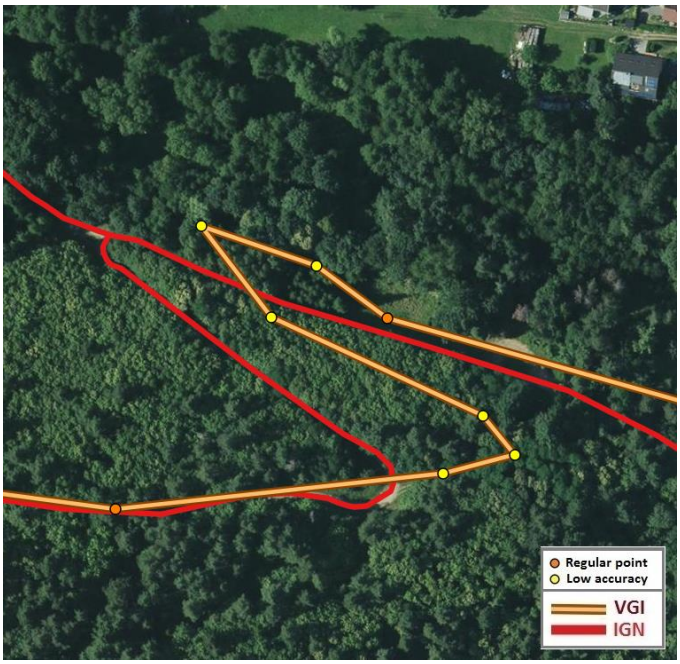


Figure 61 : Low accuracy points detected – rule 4

Rule 5, illustrated in Figure 62, is modelling low accuracy points as points with low local spatial resolution (DistMean >116.34m) and not negligible direction change (AngleMean>21°.14)

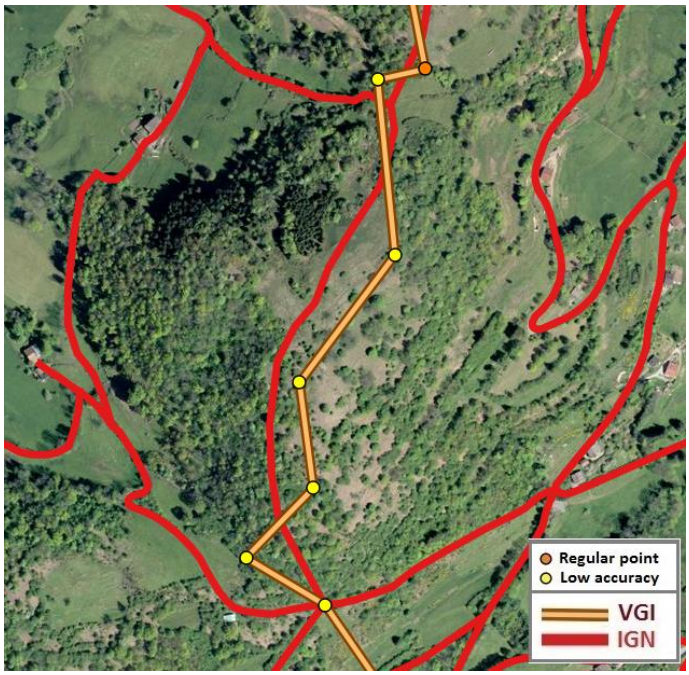


Figure 62 : Low accuracy points detected – rule5

Results of rule 6 are shown in Figure 63. Low accuracy points were found as points with considerable slope between each other (Slope>25.18%) and with relatively small distance, lower than 11.16m.

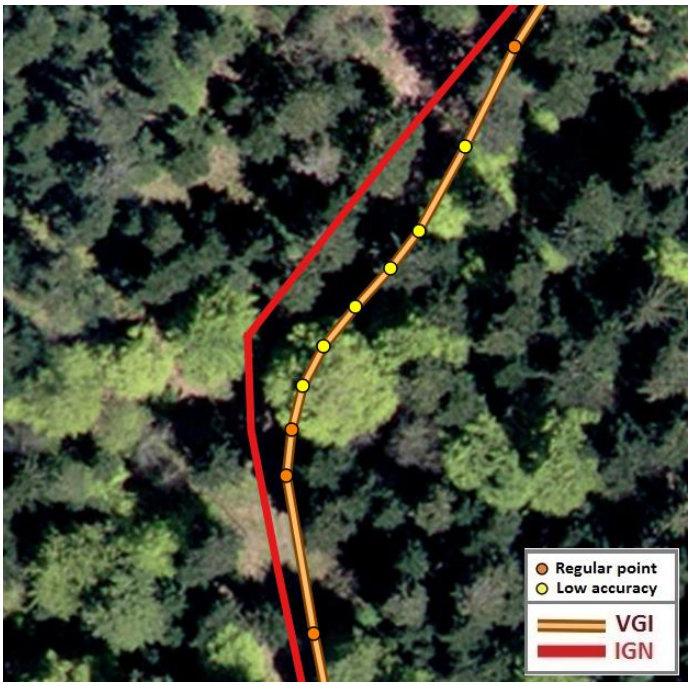


Figure 63 : Low accuracy points detected – rule6

Rule 7 formalizes low accuracy points as points in deep forest with extremely steep slope (Slope>31.2%) and having a certain change of direction, higher than 21° (see Figure 64).

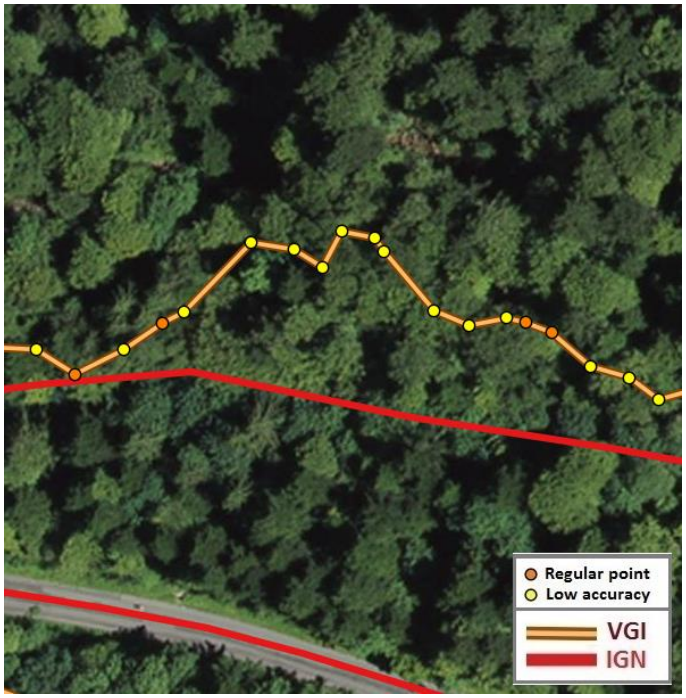


Figure 64 : Low accuracy points detected – rule7

As reported before, the approach has not performed highly due to the discussed significant limitations of VGI GPS data. In next examples, we will illustrate where the approach lacks precision as well as recall.

First example (see Figure 65) presents the lack of precision, a case where four points were detected as low accurate, three of them falsely (1,2,4) and one correctly (3).

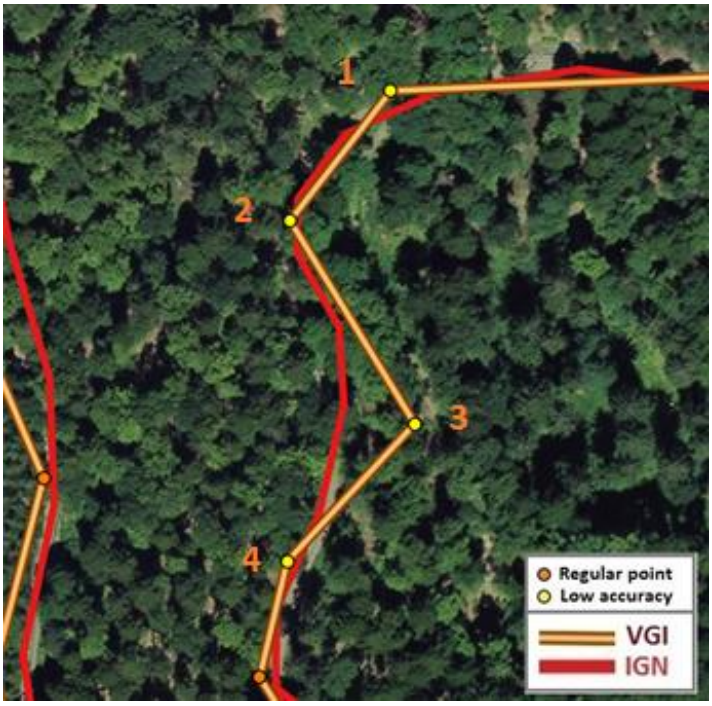


Figure 65 : False positive low accuracy points

The example implements rule 1, which means that points have very sharp direction changes between them, followed by very steep slopes and all of that in deep forest environment. Not only according to our machine learning classification those points are low accuracy, but according to presented state of the art in 1.2.4, they are very likely to be low accuracy, since conditions modeled by rule 1 are found very harmful for GPS device accuracy. Even having very negative conditions for GPS accuracy they are somehow very accurate points. It is difficult to explain, but knowing how the accuracy of VGI GPS data is unpredictable, the situations like this are possible and not very rare. This is a typical example of false positive results for all seven rules.

Second example concerns a lack of recall – low accurate points not detected. A typical case is shown in Figure 66.

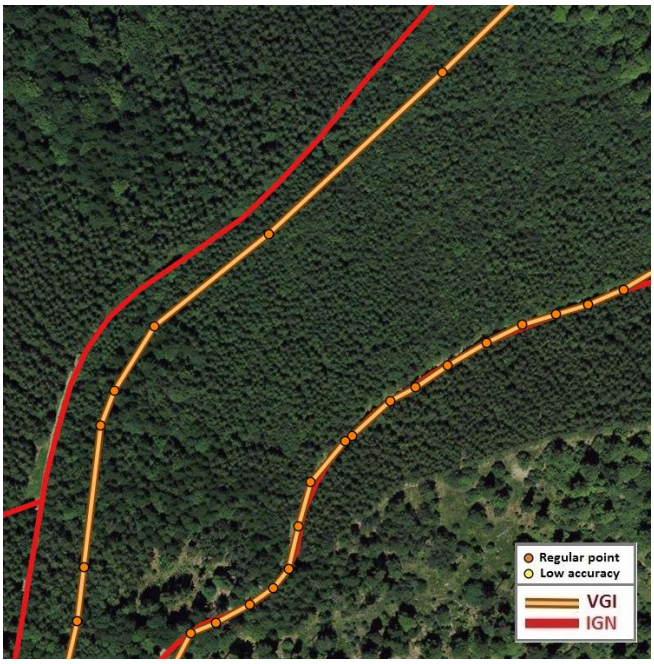


Figure 66 : False negative low accuracy points

Trace on the left is in question. Segments compose the trace on the left deviate 23m in average from IGN road. In this case all indicators (both intrinsic and extrinsic) have regular values common for good accuracy points, except the indicator implying that points are in deep closed coniferous forest. It may be a reason for weak accuracy. However, it is likely that this indicator was not the only reason for weak accuracy, but its interaction with some other very important factors we are not able to consider, like poor geometry of satellites, weak signal, etc. This proves how difficult is to evaluate accuracy of VGI traces without comparing them to referential road network. Systematic influences on GPS accuracy like geometry of satellites are not possible to assess without having corresponding metadata like PDOP. Due to the plenty of systematic factors affecting GPS (presented in Section 1.2.4.1) recall of the approach is generally weak.

In addition, multisource data like our traces that are coming from GPS devices with very various accuracy limit precision and recall of such type of approach. On the one hand, high quality GPS device can record good accuracy points even in very challenging environments, whereas low cost device can produce low accurate points even in very convenient environment. Even in same environment, two different devices can perform very differently like in Figure 66 (low accurate trace on the left close to perfectly accurate trace on the right). Huge variation of accuracy between sources does not allow remarkable results in modelling accuracy of GPS points in presented way. Metadata is of the highest importance for assessing such data accuracy.

We found surprising that the indicator DiffElevDTM was not recognized as an important one in modelling low accuracy points. Heselton (1998) proved a correlation between elevation and 2D position errors, especially in case of huge elevation errors. Thus, we considered it as potentially a very good clue of low accurate 2D position when DiffElevDTM indicator was designed.

Mainly, there are two reasons why it failed to provide clues of 2D position error. First, policy of application providers like Strava GPS consists in replacing elevations measured by GPS device by elevation sampled from external sources like referential DTM. The process is described in the flowing link: https://support.strava.com/hc/en-us/articles/216919447-Elevation-for-Your-Activity. In this case, the elevation error is relatively small even if it is very high in the reality. Second, since there is no protocol for VGI traces collection, sometimes a correction for optometric height is not applied. That results in falsely high elevation errors. Relevant information about elevation corrections (see Figure 67) is not provided due to the absence of metadata, thus this issue was discovered in evaluation of results of our approach.

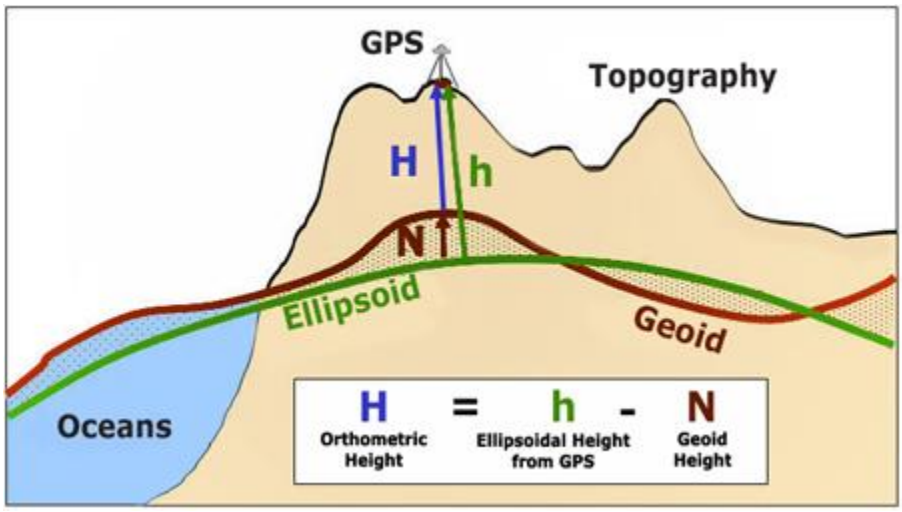


Figure 67 : Relation between ellipsoid (GPS) elevation and topographic (geoid) elevation (by keywordsking.com)

Figure 67 illustrates a relation between two height systems, based on ellipsoid used by GPS and based on geoid used as authoritative data elevation reference system.

In order to confirm our hypothesis regarding missing elevation corrections, two relevant pieces of information were required: average elevation error of smartphone (16m) (Samsung Galaxy S6) and approximate geoid height (N) for our test zone (48-49m) (Service de Géodésie et Nivellement - IGN France, 2010). Smartphone GPS errors were considered in accordance with our hypothesis that most of our data is collected using on smartphones. Therefore, the zone of elevation values without corrections (zone P) taking into account average elevation error was determined as:

Bottom threshold: 48m-16m= 22m

Top threshold: 49m+16m= 75m (5)

Zone P: 22m<Z<75m

To confirm these assumptions, a distribution graph of DiffElevDTM indicator was done. The graph is presented in the Figure 68.

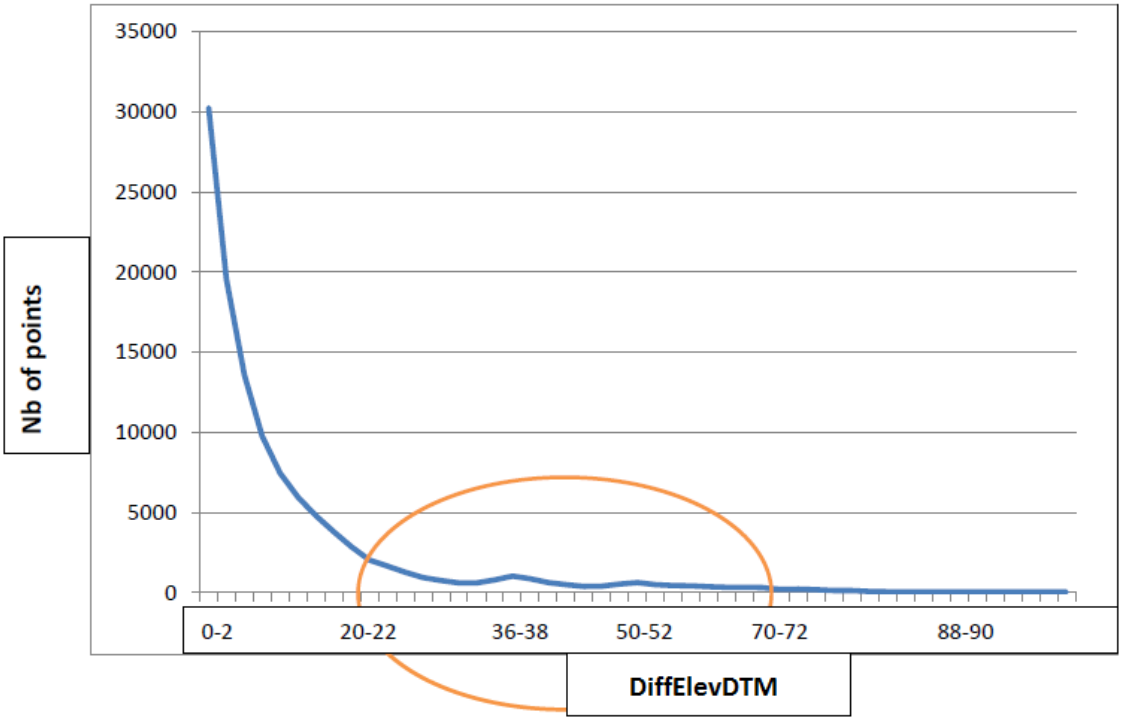


Figure 68 : DiffElevDTM distribution graph

In the graph changes in the zone of our interest (labelled by orange ellipse) can be spotted, but they are not significant compared to all data.

Extremely high number of points with DiffElevDTM <2m – with elevation almost like DTM elevation come mainly from already mentioned policy of application providers. Thus, this very high accuracy of elevation is mainly false. In total 21,812 points have an elevation in zone 22m<Z<75m, so they are potential candidates for the points without elevation correction.

In order to separate points that are without elevation correction from the points that have corrections but have a real elevation error between 22 and 75m, we start from a hypothesis that points without elevation correction come from the same trace or group of traces. Since, when recording and exporting a GPS data by smartphone applications, the correction is applied on all points or not a single point. Thus it is important to find traces with high percentage of points with elevation in the zone P.

We found 11 traces with more than 95% points with elevation in the zone P, and 1 with more than 84% of points in the zone Z. All other traces are far below the 84%. The next one is on 59% and the majority is under 50%. A deviation from 100% is quite normal, since we used average error of elevation of 16m, whereas in some cases it can exceed that level and have an error of 17, 18, 19m and more meters. Most importantly, the majority of points in our interest are in zone P.

Hence, if we consider traces with more than 84% of points with elevation in zone P as traces without elevation corrections we can conclude following things:

- Total number of points in those traces is 11,603, of whom 11342 are points with elevation in zone P (97.7%)
- They are all coming from 3 websites: VTTour, VisuGPX and RandoGPS, 4 from each site.
- No need for repeating detection of outliers and low accuracy points since the presence of points without correction is not significant (11063—4%)
- Potentially we can propose a method for detection of traces without elevation correction and corrections as well.

In order to proceed to the data matching phase, segments are constructed as proposed in UML model in Section 1.3.1. The accuracy of segments is determined based on the accuracy of points composing them. Thus, we have two types of segments according their accuracy: low and good accuracy segments. Good accuracy segments are those composed by two good accuracy points, whereas if one or both points composing a segment are low accurate, the segment is considered low accurate.

1.5 Conclusion

In this chapter we presented current state of the art of VGI data quality assessment, and our approach for evaluation of VGI traces quality. Since VGI data studies are quite recent, the state of the art included the relevant research in a past decade. Special attention was paid special on linear VGI data. In addition, a data source aspect of our VGI data was particular considered, thus an overview of GPS data quality evaluation was presented. Overall, very few research works were dealing with VGI traces in real conditions, with few or no metadata at all, and high heterogeneity. In addition, the performance of GPS devices in different environment was analysed. The findings obtained were after used in the approach for evaluation of VGI traces data quality. The approach proposed consists of three methods, two based on by machine learning techniques: detection of outliers in GPS points and detection of low accuracy GPS points, and a third one based only on geometric operations: detection of secondary human behaviour.

First, the approach, tested on real data in mountain areas, performed efficiently to detect outliers in VGI traces. Simple rules, only based on intrinsic indicators, have been learned. In total about 3% of points are eliminated as outliers. This represents about 80% of actual outliers that are eliminated from raw data. Second, the approach treated trace's geometric anomalies caused by secondary human behaviour with very successfully. Both precision and recall remained high 98% and 93% respectively. Third, a similar learning approach to that used in detection of outliers was applied for identifying the accuracy of GPS points (classified into two raw categories: low and high accuracy). As a result, seven rules were learned involving seven intrinsic but also extrinsic indicators. The performance of the approach remains medium, but shows its potential and may be sufficient for some purposes. The inability of the approach to classify correctly half of points is mainly due to a lack of some of crucial metadata like PDOP, number of visible GPS satellites or knowing the GPS device. Along to lack of the metadata, two other key limitations of the approach proposed are high heterogeneity of VGI traces (different accuracy of GPS devices used, different positions of devices while collecting) and not enough satisfying quality of referential data used in machine learning (DTM resolution, land cover accuracy).

Chapter 2

Data matching

2 Data matching

According to Walter and Fritsch (1999) data matching is a tool that links homologues spatial objects that represent the same reality. In this section, we are first presenting relevant work on spatial data matching paying special attention on works dealing with feature matching since this is in focus of our work. Second, the proposed approach for detection of updates in authoritative spatial databases based on data matching will be described and evaluated by using real authoritative data coming from IGN France. Finally, conclusion and perspectives will be presented.

2.1 Goal of data matching

Data matching is a relevant research field in Geographic Information Sciences with many direct and indirect applications. Among them, we can cite geographic data integration (Devogele et al., 1996; Zielstra and Hochmair 2011; Al-Bakri 2012; Costes 2015), conflation (Smart et al., 2011; Song et al., 2011), quality evaluation (Goodchild and Hunter 1997; Koukoletsos et al., 2012), and detection of updates (Uitermark et al., 1999; Liu et al., 2015; Van Winden et al., 2016).

Geographical data integration implies unifying geographical data from different sources in a unique database in order to ensure unified environment for processing, modelling and visualization (Abdalla 2016). Since data is coming from different sources, usually defined with different specifications and different purposes many issues such as integration of data with different quality (e.g. accuracy, completeness) and level of detail occur. This situation especially affects data matching by causing mismatching results. Additionally, matching results are limited in case when datasets to integrate have different geodetic datum (projection) which is not a rare when integrate multi source data. Effects of projection transformation on matching data from different sources are presented in the work of Siva Kumar (2000).

Above discussed limitations can be especially visible in case when the integration of VGI and authoritative or formal data should be done. Evaluation of integration between VGI data (OSM) and authoritative data (e.g. Ordnance Survey) was done by Al-Bakri (2012). The main conclusion of this evaluation is that there are still significant incompatibilities and divergences in integration of VGI and authoritative datasets, in the first place the accuracy. However, few positive aspects of these incompatibilities are mentioned for example the richness of VGI data. On the other hand, Zielstra and Hochmair (2011) found satisfying the integration of OSM data with formal data coming from TeleAtlas and NAVTEQ. However, this refers to a very small aspect of integration where pedestrian paths of OSM were found useful in enhancing accessibility to bus and metro stations in US and German cities. Thus, generally speaking, the issues are still more numerous than advantages. In his work Al-Bakri (2012) proposed nine recommendations for the improvement of VGI and formal data integration. They are mainly oriented to introducing more VGI sources together (like OSM, Flickr, textual descriptions etc.) into integration and improving the quality of VGI data before its use.

Conflation or map conflation aims to enhance or enrich map content by combining data from different sources. It is a challenging task if datasets have different projections, accuracy levels and resolutions (Chen 2005). Even there are a lot of online maps available, many of them are not geo-referenced which

significantly limits the conflation process. (Chen 2005). With the beginning of the era of digital maps produced from geographical databases (GDB), according to Gillman (1985), and Gabay and Doytsher (1994), map conflation is divided into two phases: the identification of possible correspondences between elements (matching) and the alignment of these matchings. Data matching plays an important role here, since it comes as a first step. The final results of map conflation are thus highly dependent on results of data matching (Walter and Fritsch 1999; Uitermark 2001). In order to overcome that issue, some research works have proposed three phases in conflation (Cobb et al., 1998; Chen et al., 2004). The first phase is feature matching. The second phase is validation which should ensure that there is no inappropriate matching and the differences between matched objects are evident. The last phase consists in correcting spatial data or creating new integrated data so that evident differences are removed. A comprehensive state of the art of digital map conflation can be found in the work of Ruiz et al., (2011).

Every spatial dataset has different quality depending on numerous factors, like accuracy of collection process, granularity, etc. Usually the quality is well-known and already specified as it is the case for authoritative datasets. However, with growing use of the new technologies of Web 2.0, more and more produced spatial datasets have unknown or heterogeneous quality such as VGI. The quality of spatial data is generally assessed by comparing dataset in question to a dataset with known quality that is used as a reference, mostly considered as “ground truth” data. Data matching is used here in identifying which feature from a dataset in question should be compared to which feature in a referential dataset. Many works on this topic has been done to assess the quality of (Goodchild and Hunter, 1997; Girres and Touya, 2010, Al Bakri et al., 2012) as already detailed in the Chapter 1.

Spatial entities are changing continuously in the real world. Their representations are usually stored in GDB also need to follow those changes in order to be up-to-date. This includes modification of objects, adding new objects or deleting existing objects. Data matching is used here as a tool for updating GDBs. The main goal is to follow the changes of objects from the last update by identifying first homologues objects between these two milestones. Once homologues objects are identified, the changes in terms of geometry or attributes can be examined. Also, objects having no their homologues objects are a clue that some spatial entities are created in the real world but not yet represented in a GDB. Cited works in this field dealt with updating of road networks which is also in our focus. Uitermark et al., (1999) proposed a geometric approach based on Constrained Delaunay Triangulation to update Top10vector database, whereas Van Winden et al., (2016) focused on updating Dutch authoritative road attributes (e.g. one-way road or not) by means of VGI traces collected for the research purposes of Delft University of Technology in the Netherlands. Liu et al., (2015) proposed an approach to detect new roads for updating authoritative road network by using OSM data. The approach first applies a data matching that identify new roads. Four scenarios for integration of detected new roads are defined (see Figure 69).

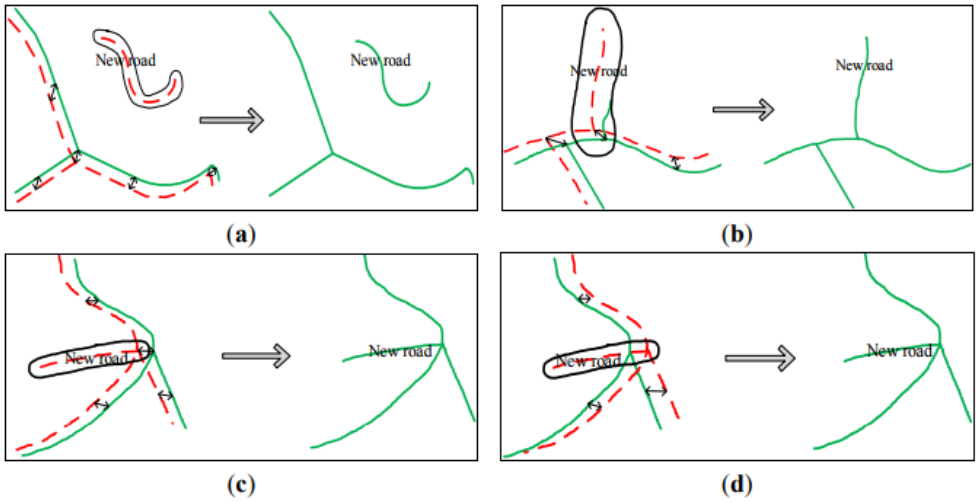


Figure 69: Scenarios for updating new roads into existing road network (Lui et al., 2015)

Scenarios are as follows:

- Isolated new road (see Figure 69 a))
- Removing insignificant reference road (see Figure 69 b))
- New road undershooting (see Figure 69 c))
- New road overshooting (see Figure 69 d))

Whether the goal is data integration, update, conflation or quality assessment, the data matching is used as a tool for defining homologous objects and consider at least two geographical databases.

Very close to data matching, the process of map matching, sometimes wrongly considered as the same process, consists in assigning locations to most probable road within a road network. It is mostly used for navigation purposes based on GPS technologies to keep a GPS user 'always on a road'. According to Lou et al., (2009) there are three types of map matching approaches: local (incremental), global and statistical. The local approach aims to identify local match of geometries between GPS points and road network by taking into account only current position or positions of some neighbouring points. As a result, such strategy is greatly sensitive on GPS measurement errors. Thus, here a fast computation time is counterbalanced by generally low accuracy. Accuracy is especially negatively affected when the approach is applied on GPS data with low sampling frequency (sampling rate >30s) (Lou et al., 2009). The global map matching approaches consists in matching the entire GPS trace with a road network. Thus, they are less sensitive on measurement errors, but still sensitive on low sampling rates. It particularly depends on the speed between points sampling. The greater the speed is, the more harmful low sampling rate is for accuracy in this context. Hence, for example a speed of only 60 km/h with a sampling rate 1 point/minute results in one GPS point every 1 km. Finally, the statistical approach can be used in both previous cases and relies on statistical methods like Hidden Markov Model, Kalman filter etc. Some works propose a solution for overcoming negative effects of GPS measurement errors (Newson and Krumm, 2009) and low sampling rate (Lou et al., 2009) on map matching results, both relying on Hidden

Markov Model (HMM). First work succeeded to deal with measurement errors even in very challenging situations such as to correctly match points placed between slip road and highway. Proposed HMM approach successfully balanced the effects of GPS measurement noise and routing behaviour by calculating optimal path based on measurements and transition probabilities. However, the proposed solution remained unsuccessful in matching low sampling rate points starting from 30 seconds. On the other hand, the approach proposed by (Lou et al., 2009) performed better with low sampling rates points (sampling rate from 30 seconds to 5 minutes) coming from on taxi drivers GPS data in Beijing urban area. Even successful in matching low sampling rate points, the approach is not generic since it was calibrated on very specific taxi vehicles case study in an urban area. Some general hypothesis of the approach such as: "True paths tend to be direct, rather than roundabout" are made based on taxi drivers' behaviour. Favouring direct paths over roundabout has sense in taxi data matching problem in an urban area, but is not applicable in various other real world situations.

2.2 State of the art of data matching

Up to now different data matching approaches and methods have been proposed depending on many aspects, like type of data to match, final goal of matching, type of criteria applied.

Research works on global taxonomy on matching approaches and methods especially regarding vector spatial data are not numerous. One of the most recent and comprehensive is a work of Xavier et al., (2016), where data matching approaches are classified function of two aspects: measures and methods.

According to Xavier et al., (2016) two types of measures are proposed in the literature: single measures (Beeri et al.,2005) and multiple measures which are combined based on different techniques such as normalized score (Pendyala, 2002), weighted combination of measures (Zhang and Meng, 2007), probabilistic theory (Tong et al., 2009), Belief Theory (Olteanu-Raimond et al., 2015). In some other works, the measures are also seen as a basis for defining criteria for data matching (Olteanu, 2008). In this last work four different groups of criteria are proposed: geometrical criteria based on geometric measures (e.g. Euclidian distance, line orientation, etc.), topological and neighbourhood criteria based on topological relationships, semantic criterion based on semantic measures and finally attribute criteria based on string measures.

Concerning the type of methods, classifications of matching methods exist depending on various aspects of matching. One of the first was a classification of Rosen and Saalfeld (1985) that categorised matching methods according to criteria they are using as: discrete or continuous, geometric or topological, local or semi-local, independent or dependent. In their work, Devogele (1996) proposed a more general classification into semantic, topological and geometrical matching. Yuan and Tao (1999) proposed basically the same classification, except the name of the second type of matching, ('attribute' instead of 'semantic'). On the other hand, Chehreghan and Abbaspour (2017) classify methods for linear object matching into two categories: criteria-based and optimization-based approaches. On the opposite, Volz (2006) classify the matching methods with respect with the type of the object to match into four categories: point-based, line-based, area-based, and mixed approaches.

In their review paper, Xavier et al (2016) argued that proposed taxonomies and classifications of matching methods are not adequate. First, because the methods are classified according to a single perspective (e.g. type of objects to match), however usually more perspectives are combined during the matching process (e.g. matching various types of objects within one approach).

Following the taxonomy of Xavier et al. (2016), relevant research works for vector data matching is presented with respect to methods and similarity measures.

2.2.1 Data matching methods

In this section relevant data matching methods are described following the classification of Xavier et al (2016). They proposed a general classification of matching methods according to two points of view: level of actuation and case of correspondence.

2.2.1.1 Level of actuation

Regarding the level of actuation, Xavier et al (2016) classified matching methods as methods for schema matching, feature matching and internal matching as illustrated in Figure 70.

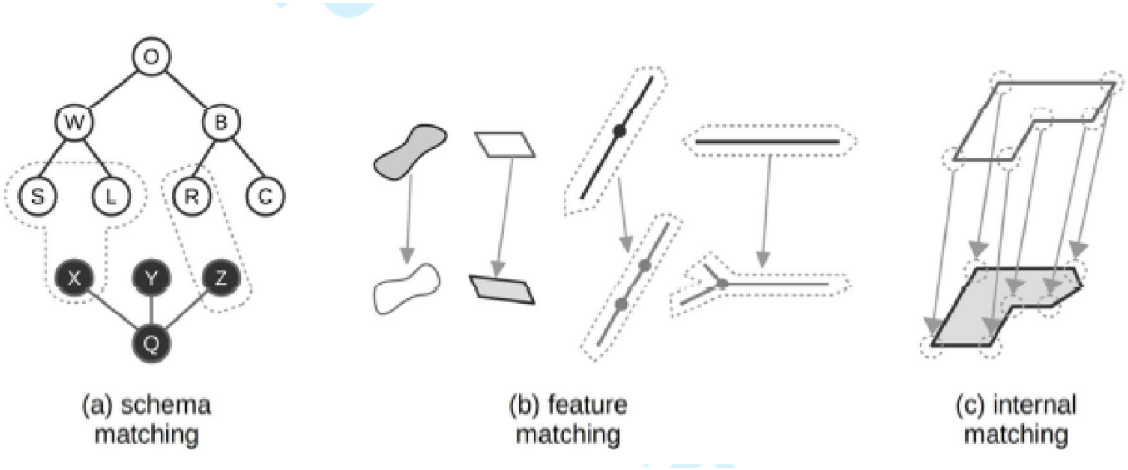


Figure 70: Classification according to level of actuation (Xavier et al., 2016)

Schema matching refers to matching of two datasets according to their data models. In other words, matching is done between classes. The goal is to find corresponding classes so that matching on the lower levels of hierarchy (e.g. feature matching) can be carried out. Matching classes is based on finding semantic similarity between classes that represent the same real world phenomenon or finding semantic correspondences between elements of two schemas as defined by Milo and Zohar (1998). In terms of integration of VGI and authoritative data, a data matching on schema level was done by Al-Bakri and Fairbairn (2012). In this work, matching schemas were conducted by means of three different measures such as name similarity (Lin 1998), data type similarity (Hong-Minh and Smith 2007) and structural similarity (Amarintrarak et al., 2009).

When homologous classes are found, the process of matching their instances can be done, process named feature matching or more currently data matching, as defined in the beginning of this chapter. There are various methods on this type of matching, for instance buffer growing matching. One of the first works using this method is the work of Walter (1997) that proposed a buffer growing matching for network matching.

Buffer growing principle matches objects from a dataset A to objects in a dataset B if objects from the dataset B are completely within buffers constructed around objects in the dataset A. The process works as represented in Figure 71. First a buffer is created around an object a2 from a dataset A. If there is no object from dataset B within the created buffer, the buffer is extended to the next object. Thus, elements a2 and a3 are combined into a new logical object a2a3 and can be matched to the element b1 since it is within buffer of a2a3. Similarly, objects b1 and b2 can be combined in the object b1b2 and matched to the logical object a2a3. In order to avoid redundancy, it is not necessary to consider the matching a2a3a4 to b1b2b3 because it was already done by combination of matching a2a3 to b1b2 and a4 to b3 (Walter and Fritch 1999).

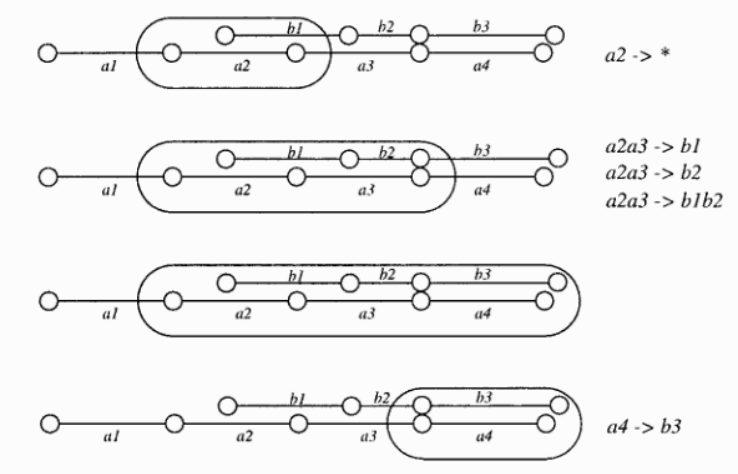


Figure 71 : Buffer growing method (Walter and Fritch 1999)

Buffer growing is not unconditional. It has a condition in terms of the number of edges which are connected to the node. Thus, on the one hand if number of edges is two, buffer can grow to the next element. On the other hand, if the number of edges is higher, a buffer can grow to the next object only if an angle between two objects is less than 45 degrees as illustrated in Figure 72.

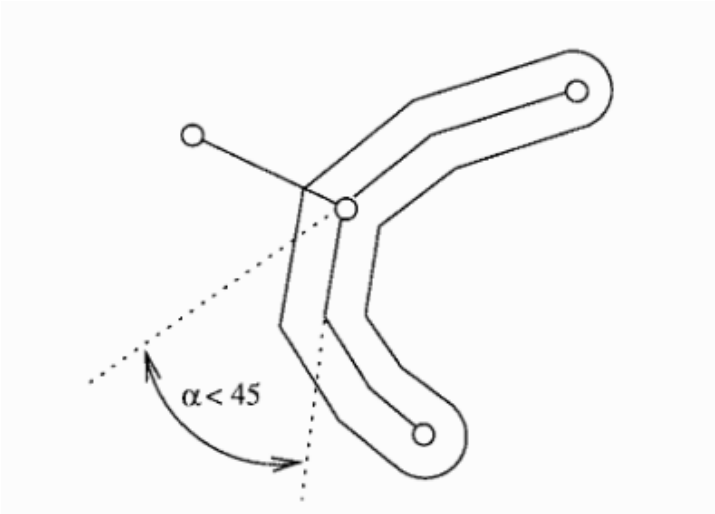


Figure 72 : Angle criterion in buffer growing method

Later studies were mainly based on this principle (Volz 2006, Zhang and Meng 2007, Chen and Walter 2009, Ying et al., 2011). Volz (2006) proposed an iterative node and edge matching method relying on buffer growing to handle one-to-many cases. Zhang and Meng (2007) adapted buffer growing principle by extending buffer growing in unsymmetrical way. Buffer growing was also extended by Chen and Walter (2009) so that it can deal with matching lines and points.

In terms of VGI data, buffer growing principle was applied in the study of Liu et al., (2015) where the goal is to update road map by using OSM data. The study particularly pointed out a dependence of detection of new roads on the buffer size used (radius). Thus, they proposed a progressive two steps process for determining optimal buffer size. First step refers to determining the range of the buffer radius based on the accuracy of a road network and on specificities of road network in the test area (e.g. width of traffic lines according to the type of road). Second step is determining optimal buffer size that must be within the calculated range. Each buffer size was tested and evaluated based on visual checking of the new roads detected. Finally, the optimal buffer size is found by dividing buffer sizes in overlapping groups and applying min-max strategy. The group with most stable (minimal) changes of results was chosen, and the optimal size is calculated by averaging sizes from the group. Besides buffer growing principle used in a selection phase, topological and attributes similarity measures were also used in identifying best matching candidates.

Various other methods were used in feature matching, like mutually-nearest, probabilistic and normalized-weights (Beeri et al., 2004), Belief Theory (Olteanu-Raimond et al., 2015).

Internal matching is on the bottom of hierarchy proposed by Xavier et al., (2016) and comes usually feature matching. It considers matching interior components of features, like matching vertices of polygons or lines. However, particular methods are usually composed by elements related to feature matching and elements related to internal matching together (Ruiz-Lendínez 2013, Fan et al., 2016).

2.2.1.2 Case of correspondence

Xavier et al. (2016) categorized matching methods according to the case of correspondence known in the literature as matching links cardinality. They proposed only three cases: one to one (1:1), one-to-many (1:n) or (n:1) and many to many (m:n) (see Figure 73).

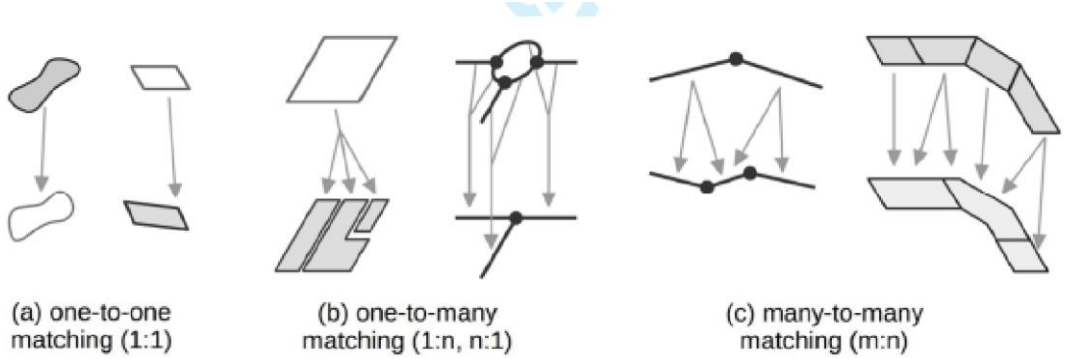


Figure 73 : Matching according to case of correspondence (Xavier et al., 2016)

Methods that recognise only cardinality one to one are (Saalfeld 1988; Beeri et al., 2004, Safra et al., 2010, Song et al., 2011). There is an extensive discussion if such methods can be sufficient in all real word situations. For example, Li and Goodchild (2010) argued in their work that such cardinality model would failed if an object from one dataset is represented with few parts in the other dataset i.e. should have few homologues object, which is not rare case especially in matching datasets with different granularity levels.

Unlikely one-to-one methods, one-to-many methods are able to successfully match datasets with different level of granularity. In terms of linear features matching, we pay attention on three works Olteanu-Raimond and Mustiere (2008), Kieler et al., (2009) and Li and Goodchild (2010). One-to-many cases were successfully treated in network matching in the study of Olteanu-Raimond and Mustiere (2008), who proposed a method based on Belief Theory (Dempster 1967, Shafer 1976). On the other hand, even if rivers in a test dataset of Kieler et al., (2009) were represented as lines and polygons (rivers having width bigger than certain threshold of representation), matching is done only between line features. Polygons are decomposed to center-lines using a skeletonization algorithm which preserves topology. Subsequently, one-to-many strategy was required. Li and Goodchild (2011) measured similarity between line features having one-to-many, and many-to-one relation.

Many-to-many relationship is adapted for managing most complex matching tasks in terms of results cardinality. This is mostly a case when matching two datasets having different data capture specifications, i.e. when same real world feature was captured following different strategies. For instance, a same block of buildings is represented with three objects in one dataset, whereas in the dataset five objects were used in its representation, like shown in the Figure 74:

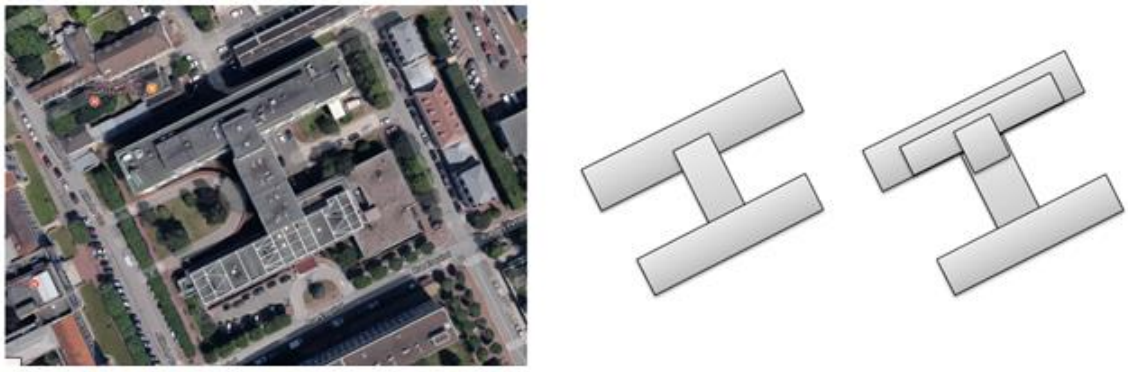


Figure 74 : a) Real world situation (Google Maps); b) Modelling same block of building according to two different data capture specifications

In the Figure 74, we can observe different representation of a block of buildings shown on the left, according to two data capture specifications: first (Figure 74 b) on the left) represents capture for maps with a larger scale and the second (Figure 74 b) on the right) for the maps with a smaller scale. Most methods designed to identify many-to-many relationships were developed to address polygons matching, from van Wijngaarden et al., (1997) to Zhang et al., (2014). In terms of linear data matching, many-to-many cardinality was allowed in works of Walter and Fritch (1999), Yang et al., (2013), Tong et al., (2014) etc. This cardinality was also applied in matching nodes (Mustière and Devogele, 2008).

The classification proposed by Xavier et al., 2016 does not include 1:0 and 0:1 cardinality. However, they are very important in case of detection of updates. Detection of updates relies on finding differences between two datasets, which would not be possible if cardinalities 1:0 and 0:1 are now allowed. Therefore, it is important to stress here that detection of updates requires all six possible cardinality cases in matching. Let us define database A and its different states, database A1 in the moment T1 and database A2 in the moment T2. Supposing that T2 is later in time then T1, updating database A1 by means of database A2 would have following cardinalities of matching between A1 and A2:

- 1:1 - the object exists in both databases
- 1:0 - the object is deleted from database A1
- 1:n - the object in database A1 is modified being split into n objects
- n:m - modification in both databases
- n:1 - modification in database A1, a fusion of objects
- 0:1 - a new object is added in A2

2.2.2 Similarity measures used in data matching

Similarity measures are an important element of matching in deciding to which extend two features are homologous (i.e. represent the same entity in the real world). Weather two objects are homologous or not depend on similarity measures used. There is no official classification of similarity measures, however, some are very cited like geometric, attribute and spatial relationship measures proposed by

Tong et al., (2009) or geometric, semantic and contextual proposed by Zhang et al., (2012). Figure 75 illustrates the taxonomy for similarity measures proposed by Xavier et al., (2016).

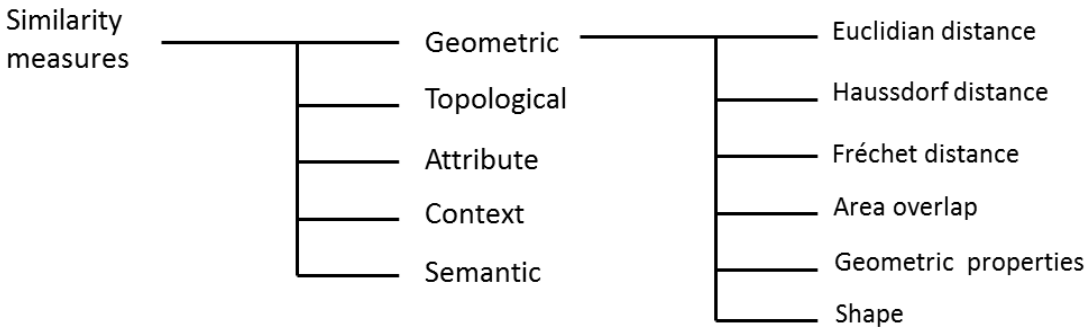


Figure 75 : Similarity measures taxonomy (Xavier et al., 2016)

As seen in Figure 75, the most dominant similarity measures are geometric measures, which is logical since vector data is in question. In addition, vector spatial features have most of the time no unique and shared identifiers, and that spatial characteristics are thus essential for identifying homologous object (homologous is first thought of as "same place" and "same shape" and "same spatial configuration" even if other characteristics may also be taken into account. In the following subsections listed geometric measures will be presented more in detail since they are widely used in spatial data matching, especially distance measures.

2.2.2.1 Geometric similarity measures

Mostly used geometric measures for matching linear features will be presented in detail such as:

- Euclidian distance
- Hausdorff distance
- Fréchet distance
- Mean distance
- Orientation
- Epsilon band

Euclidean distance

Euclidian distance is very well-known term in geometry that represents straight-line distance between two points in Euclidian space. Thus, it has been very frequently used in data matching, especially in matching points (Beeri et al., 2004, McKenzie et al., 2014). Even if it is designed for point distance calculation, it could be used for calculation of distance between linear features e.g. Euclidian distance between midpoints of two lines.

Hausdorff distance

Hausdorff distance between two lines is defined as a measure of the maximum of the minimum distances between two sets of linear features (Nutanong et al., 2011). Taking into account two lines L_1

and L_2 as illustrated in the Figure76, the Hausdorff distance (d_H) equals to the maximal value of distances d_1 and d_2 as shown in Equation 1.

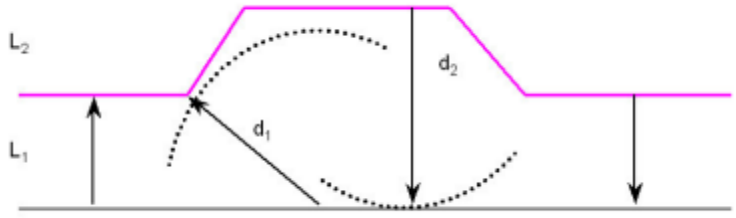


Figure76 : Hausdorff distance (Olteanu, 2008)

$$d_H = \max(d_1, d_2) \tag{6}$$

Where d_1 and d_2 are:

$$d_1 = \max_{p_1 \in L_1}\left[\min_{p_2 \in L_2}\left[d_E(p_1, p_2)\right]\right]$$
$$d_2 = \max_{p_2 \in L_2}\left[\min_{p_1 \in L_1}\left[d_E(p_2, p_1)\right]\right] \tag{7}$$

Where d_E is Euclidian distance.

Hausdorff distance is successfully used in matching linear features (Yuan and Tao 1999, Chen and Walter 2009). However, it has some limitations when lines to match are not of the same length (e.g. matching data from databases with different scale). In that case, a more adapted solution is Hausdorff semi-distance (d1 or d2) as proposed by Mustière and Devogele (2008). Another modification is median Hausdorff distance that takes median value of all minimal values. Tong et al. (2014) proposed to compute a short-line median Hausdorff distance.

Fréchet distance

Apart from official definition, Fréchet distance has some more informal definitions that describe it very well. One of them is: "The Fréchet distance between two curves in the plane is the minimum length of a leash that allows a dog and its owner to walk along their respective curves, from one end to the other, without backtracking." (Chambers et al., 2010).

More mathematically, Fréchet distance is defined in the following way. Let us considering two polylines f : [0, N]->V and g : [0, M] -> V' and an Euclidian distance d, Fréchet distance is defined as:

$$d_F(f,g) = \min_{\substack{\alpha:[0,1]\to[0,N] \\ \beta:[0,1]\to[0,M]}} \{ \max_{t\in[0,1]}[d(f(\alpha(t)), g(\beta(t)))] \} \quad (8)$$

Where N, M ∈ R are numbers of segments composing respectively polylines f and g, V and V' are vector spaces, and α(t) et β(t) are continuous functions increasing during time, where α(0)=0, β(0)=0, α(1)=N (6) and β(0)=M.

The question which distance has better performance as a similarity measure for matching linear feature is still open. Some authors give an advantage to Fréchet distance in front of Hausdorff distance (Mascret and Devogele, 2006; Bouziani et Pouliot, 2008) stating that it is more adapted for linear polylines similarity determination, while Driemel et al. (2010) consider it more natural distance between curves than Hausdorff distance. However, its computation is significantly more complex and time consuming as demonstrated in the study of Wenk et al., (2006), where few hours where required for matching only one trajectory in a large dataset.

Mean distance

Mean distance between two lines is defined by means of polygon area created between them. As we can see in the equation (4) mean distance between two lines is a ratio of polygon area represented in grey (see Figure 77) and a mean value of total lines' length (McMaster, 1986) (see equitation 9).

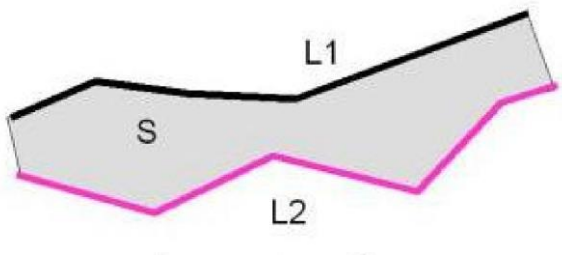


Figure 77 : Mean distance (Olteanu, 2008)

$$d_m = \frac{S}{\frac{L_1 + L_2}{2}} \quad (9)$$

The mean distance is easy to calculate and apply, however, according to (Olteanu 2008) it is more useful in generalisation purposes than data matching since it is based on mean distance between two lines to compare, instead of a maximal like in case of Hausdorff and Fréchet distance.

Orientation

Orientation of two polylines was proved to be a reliable similarity measure for matching two polylines by Olteanu (2008). Orientation measures the degree of collinearity of two polylines, L_1 and L_2 defined as a difference between the orientations of L_1 tangent T_1 and the closest point of L_2, and the L_2 tangent T_2 and the closest point of L_1 (see Figure 78).

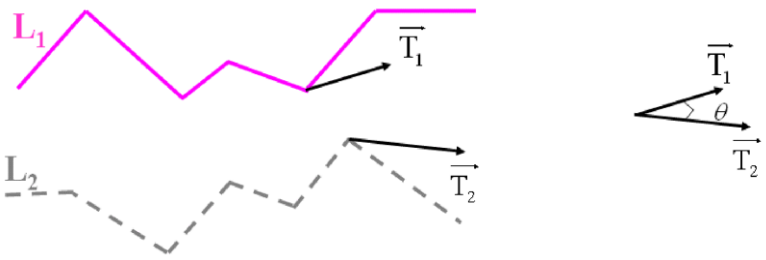


Figure 78 : Orientation of two polylines (Olteanu, 2008)

The process of computation of this similarity measure is defined in the following way. First the closest point of the polyline L_1 to polyline L_2 is found and vice versa. After, tangents T_1 and T_2 of polylines L_1 and L_2 in the closest points are constructed respectively. Finally, the angle Θ is calculated as an angle between tangents T_1 and T_2. The general assumption is: The smaller the angle Θ is, the more similar polylines L_1 and L_2 are.

However, Olteanu (2008) pointed out limitations of this measure in three cases, when polylines have different lengths, are shifted and have high sinuosity.

Epsilon-band tolerance region

The main idea of this measure is to define a buffer around an object to match, in order to find its homologues objects among other objects. The concept of epsilon-band is illustrated in the Figure 79. There are few different approaches relying on this measure. Gabay et al., (1994) consider two objects matched if one is completely within the epsilon-band of another one.

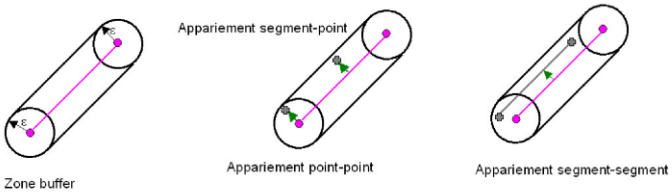


Figure 79 : Epsilon-band implementation in matching (Olteanu, 2008)

On the other hand, Sui et al., (2004) proposed to take into account the length of the object within buffer zone of the other one, i.e. if the part of object within the buffer zone of another object has a certain length, then the two objects are matched.

2.2.2.2 Topological similarity measures

Topology helps in understanding spatial relationships between objects (e.g. intersect, within, touch, etc.). Thus, topology is very convenient to be used in matching geographical databases. For example, Olteanu (2008) rely on the following principle in its matching process “Two geographical objects A and B are matched if the object A has neighbourhood relationships similar to the one of the object B.”

Topological measures have mainly been used in matching networks (Walter and Fritch, 1999; Safra et al., 2006; Mustière and Devogele, 2008, Olteanu-Mustière, 2008). Walter and Fritch (1999) used topology to match objects within one dataset relying on a fact if objects are connected or not. Safra et al., (2006) used topological relationships between objects from two datasets in order to match them. Mustiére and Devogele (2008) proposed an approach named NetMatcher for matching networks. The global approach is that matching roads and matching crossroads are closely related tasks: roads and crossroads are matched by analysing their topological configurations, in order that matched crossroads are connected by matched roads. In addition, they defined a turning criterion to verify consistency of arcs around a node.

Olteanu and Mustière (2008) proposed the neighbourhood criterion based on the assumption that matching a feature depends on the matching of its neighbours. Thus, based on the results of a first matching, two features L1 and L2 are homologous if the neighbours of the features A are matched in a comparable way with the neighbours of the feature B.

Costes (2015) proposed a neighbourhood criterion to match hydrographic networks. It is based on first landmarks matching results and made the assumption that the higher the number of landmarks matched near to two candidates' arcs is the higher the probability that the two candidates are homologous is.

2.2.2.3 Attribute similarity measures

Attribute measures are based on types (e.g. number of lines, name of the road) and values (e.g. highway, street, etc.) of attributes. In case of spatial objects, types of attributes correspond to thematic aspect, whereas values correspond to semantics. Using attributes in measuring similarity between objects to match can thus be related to comparing semantic similarity (Comber et al., 2004) or syntactic (Levenshtein 1965). A difference between semantic and syntactic similarity is significant. Former is based on comparison of meanings of objects, whereas latter is based on comparison of characters composing attributes. For example, the comparison of characters would indicate zero similarity between highway and dual-carriage way as well as between highway and river, whereas semantic information comparison would indicate significantly higher similarity between highway and dual carriage way as two road types, both with two lines than between highway and river.

Among many tools for attribute similarity measuring regarding syntactic aspect, two are very well-known: Levenshtein distance (Levenshtein 1965) and Hamming distance (Hamming, 1950). Former representing minimal number of single-character edits needed to change one word into the other one, whereas latter is applicable only on attribute values with same number of characters.

Semantic similarity measures between two concepts are calculated as a function of length of the path linking the concepts and the position of the concepts in the taxonomy (Meng et al., 2013). For instance, Wu and Palmer measure (WU and Palmer, 1994) considers the position of two concepts c1 and c2 in the taxonomy relatively to the position of the most specific common concept lso (c1,c2). It is based on the assumption that the similarity between two concepts is the function of path length and depth in path-based measures.

2.2.3 Conclusion of the state of the art of data matching

The state of the art presented a review of research done in data matching in last few decades focusing on data matching for linear features. Overall there are many classifications of matching methods according to various points of view. The classifications vary from very global such as according to level of actuation (Xavier et al., 2016), to more specific such as according to type of criteria used (e.g. topological) (Devogele et al., 1996).

Having considered the eligibility of proposed methods for our work we can claim that most of them are not completely adapted.

First, methods relying on topology, either as a main or one of criteria, are not appropriate for our matching problem, since a topologically correct network cannot be built from VGI traces due to their significant topological inconsistencies. These quality issues of VGI traces were particularly elaborated in the Chapter 1 of the thesis.

Second, methods having attribute measures as one of criteria are not eligible for our work since VGI traces have a considerable lack of metadata and values of attributes, which was also presented in detail in the Chapter 1. Not only the lack of attributes is an issue here, but also the fact that VGI traces are very poor regarding the number of attributes describing them (see 1.3.1 in the Chapter 1).

Finally, let us also mention that even our data is GPS data, using map matching solutions is not appropriate match VGI traces to authoritative road network. As already discussed, map matching algorithms are designed to keep a sensor in the road network by default. This means, that all parts of a VGI trace must be matched to the most similar road in the map. Due to that, detection of updates in a road network is not feasible.

Thus, a geometric based method using buffers remained as potentially eligible. However, it is expected that buffer based selection of matching candidates proposed in existing approaches would not perform highly if applied on such VGI dataset. The proposed approaches are using single fixed thresholds. Even in strategies for choosing an optimal buffer size (Lui et al., 2015) when few sizes are tested, the selected one is a single one applied on all features in a dataset or in subsets (smaller parts of the dataset). In case of VGI traces having huge and random variations of quality as proved in Chapter 1, applying a single threshold for all traces in a dataset can cause significant side effects (numerous false positive and false negative matching results) like illustrated in the Figure 80.

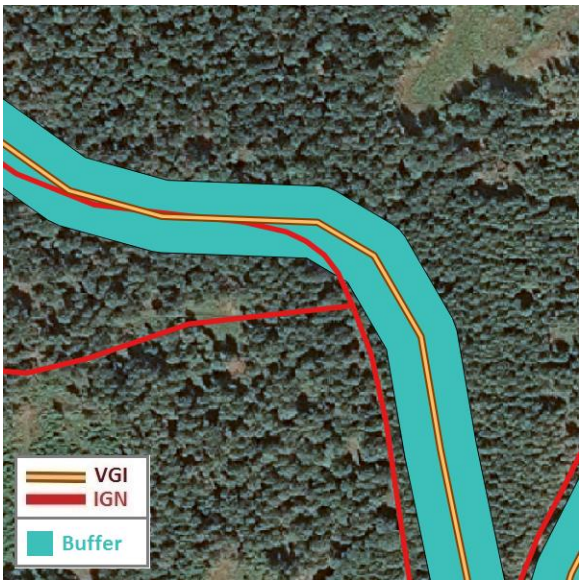


Figure 80 : Disadvantages of fixed buffer size

As we can see in the Figure 80, a fixed size buffer (30m) represented in turquoise has selected all possible IGN candidates of a high accurate part of VGI trace on the left, whereas low accurate part on the right remained with no candidates to match. Therefore, fixed size buffer is not an appropriate solution here.

Therefore, a new matching solution should be proposed, in order to deal with VGI traces matching issues discussed above.

2.3 Data matching approach for detection of missing roads

As mentioned before, data matching has a wide range of use. In our work, data matching is used as a tool for detecting updates in authoritative dataset by identifying differences between two datasets: VGI traces and authoritative dataset to update (i.e. IGN road network).

Cardinalities of matching links used for updates presented in Section 2.2.1.2 are here interpreted in the light of our main goal described in the previous paragraph. Figure 81 illustrates our interpretation of cardinality of matching links between IGN and VGI traces.

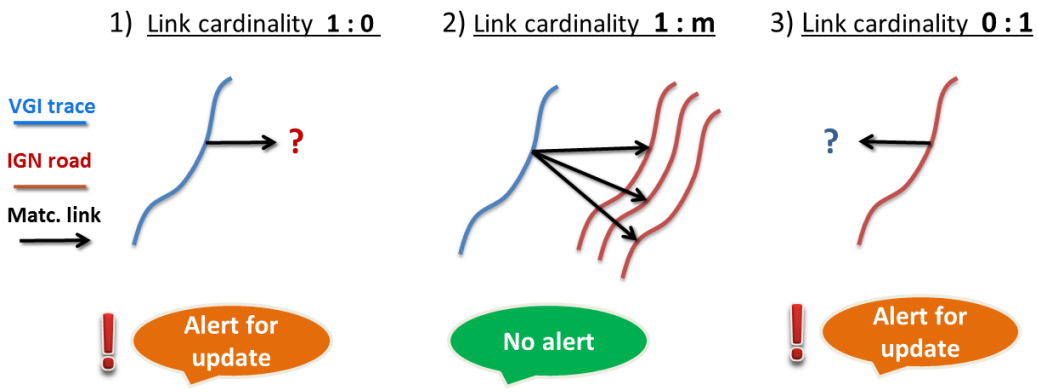


Figure 81 : Possible results of matching authoritative dataset (IGN) to VGI

Three possible results can be distinguished:

1. The cardinality 1: 0 mean that one VGI object does not have a corresponding object in IGN dataset. An alert is generated to highlight that a missing road is found.
2. The cardinality 1: m indicates that a VGI object has at least one corresponding objects in IGN dataset, or VGI objects have corresponding objects in IGN dataset. In this case, an alert is generated confirming that the IGN object still exists.
3. The cardinality 0: 1 means that an IGN object has no corresponding objects in VGI dataset. Two hypotheses may be defined here either that the IGN road does not exist anymore in the reality or the road is not used by the VGI contributors.

We are interested in the first case. Nevertheless, even they are not in the topic of our research work the other two cases are interested for both validation and update goals. Indeed, the second case is interesting in managing authoritative datasets, since a validation that a road represented in an authoritative database still exists in the real world is very important, especially for mountains paths which change frequently. The third case requires more information and better quality of VGI traces like huge amount of VGI traces following the same path, good spatial completeness of VGI dataset. In addition, it requires other external sources for verification like satellite or orthophoto images, Flicker data, itinerary descriptions, etc.

Thus, the hypothesis of our approach is: “Objects in VGI dataset that are not matched to objects from authoritative dataset are considered as potential updates, i.e. candidates for missing roads. “

First, whatever algorithm is used, the automatic matching would not perform highly due to the huge deviations in VGI traces quality. Such deviations of traces already detailed in Chapter 1 (e.g. segments deviating from closest IGN road for 100m) would usually remained unmatched after matching even if they do not correspond to any missing roads.

Second, some parts of traces are out of the road movement due to some behavioural reasons (e.g. exploring forest). After matching, these parts are detected as unmatched. However, neither of them corresponds to any missing roads.

Thus, we decided to divide the process of detection of missing roads in two steps. First, raw data matching process that will be presented in this section, devoted to provide candidates of missing roads. Second, a decision making step that will be presented in the next section, devoted to evaluation of the candidates.

The proposed matching approach is a simple matching solution adapted to VGI traces and their specificities. The main reason for this is that VGI traces are not rich enough regarding data needed for calculation of different similarity measures used in data matching. Further, a simple mostly geometric based solution is supposed to be sufficient, since the quality of VGI traces is successfully improved after applying proposed approach for evaluation of spatial data quality (see 1.4.2 in Chapter 1). Hence, it is expected that after filtering of outliers and secondary human behaviour, matching can be successful (to minimize false negative candidates for missing roads) even without being able to use numerous similarity measures.

Our proposed data matching approach consists of two main steps: selection of candidates to match and filtering of selected candidates as illustrated in Figure 82.

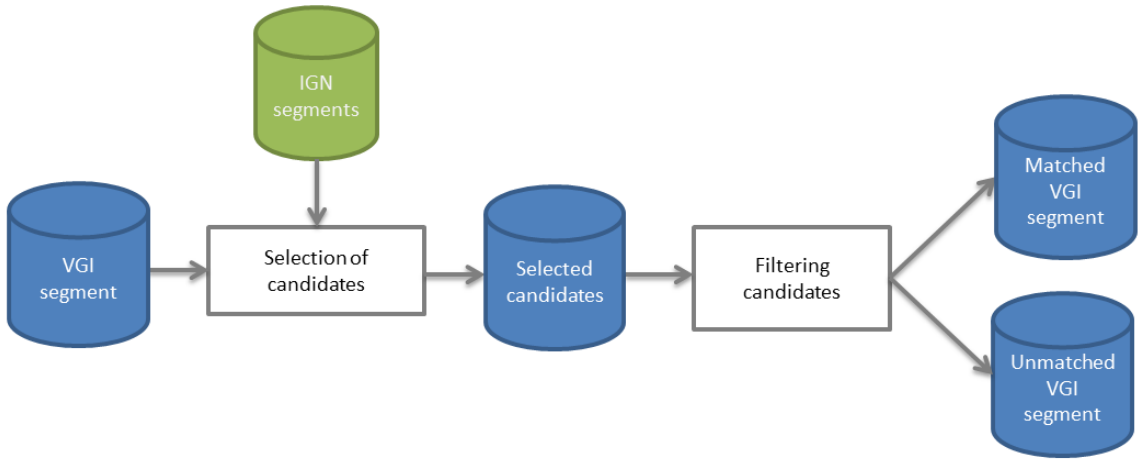


Figure 82 : Proposed data matching approach

In our object based matching approach, an object is a segment composing a VGI trace. As shown in the Figure 82, the selection of candidates starts from VGI segments and considers selection of surrounding IGN road segments. As a result, pre-matched VGI segments are obtained (i.e. selected candidates). Second phase consists in filtering selected candidates by applying geometric based criteria in order to obtain the matched and unmatched segments. Following the hypothesis of the approach, unmatched VGI segments obtained will be considered candidates for missing roads and will be analysed further in the next steps of the global approach, whereas matched VGI segments will not be considered anymore.

2.3.1 Selection of candidates

The first step of the approach is the selection of candidates. For each segment S_{VGIi}, i=1..M, from VGI dataset, we look for close segments in IGN road network, according to a distance criterion modelled by buffers. These segments, noted $\{S_{IGNj}\}$, j=1..N, that intersect the buffer of the segment S_{VGIi} are the candidates for matching with S_{VGIi}. .

As discussed in Section 2.2.3, fixed size buffer solution is not appropriate in our case due to the huge and random variations of VGI traces quality. Thus, we propose an approach based on flexible, quality adapted thresholds. Hence, a greater size is used for VGI low accuracy segments and a smaller size for good accuracy segments, both defined in Section1.4.2.4. The selection strategy is illustrated in Figure 83 where purple buffers are generated for low accuracy segments, whereas green buffers are for good accuracy segments.

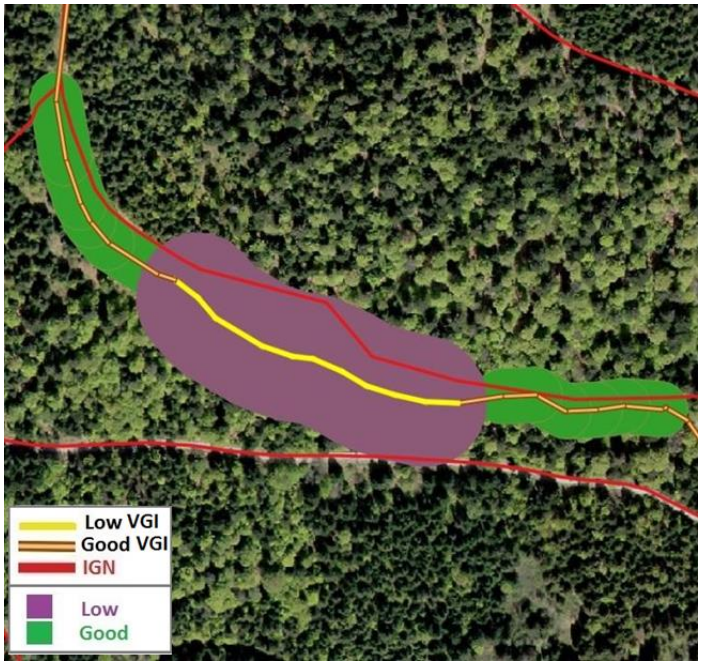


Figure 83 : Quality adapted buffer size strategy

In this way the problem of false positive and false negative unmatched segments (candidates for missing roads) will be reduced, since the main reasons for them are traces accuracy issues.

2.3.2 Filtering of selected candidates

For each VGI segment, S_{VGIi}, i=1...M, the selection phase results in a large number of candidates. However, not all of them are really corresponding to the VGI segment. Thus, the process of filtering is devoted to elimination of candidates that have low similarities with corresponding VGI segment (angle criterion) and taking spatial context (neighbourhood criterion) and authoritative data specification (length criterion) into account.

The criteria are applied one after the other, as follows. First the angle criterion is applied. Second, the neighbourhood criterion is applied the obtained results. By applying angle criterion most probable candidates for matching remained, whereas after applying neighbouring criterion sets of consecutive segments with same matching result are more continuous. Finally, the length criterion is applied after merging the continuous matched and respectively unmatched segments.

Each criterion is presented in detail in next subsections.

2.3.2.1 Direction criterion

The criterion is based on the assumption that two segments are supposed to be matched if they have similar directions. Similarly, they are not supposed to be matched if they are almost perpendicular. The criterion is comparing local orientations between S_{VGIi} and a candidate to matched, S_{IGNj}. The orientation is evaluated through an angular difference θ. If the angle between the two segments is less than a threshold than segments are considering as corresponding, otherwise they should not be matched.

2.3.2.2 Neighbourhood criterion

The criterion is based on the assumption that the matching of an object depends on the matching of its neighbours. To apply this criterion pre-matching results are needed. Thus, for each segment, S_{VGIi}, the algorithm analysing if preceding and succeeding segments have the same matching result.

Following this assumption, two important decisions should be made. Which segments are appropriate for the application of this criterion and to which extent neighbourhood should be taken into account. In our opinion, this criterion should filter 'short segments' with different matching results than their neighbourhood since they could bias final results of trace matching by cutting continuity in matching results.

In terms of neighbourhood that should be considered, we have decided to take two neighbouring segments into account, two preceding and two succeeding (symmetric neighbourhood of four segments), like illustrated in the Figure 84.

Figure 84 : Symmetric neighbourhood of four segments

Figure 84, on the left, illustrates the case when two preceding (A, and B) and succeeding (D, and E) segments are matched, whereas a single segment (C) between them is unmatched. In this case, the neighbourhood criterion allows changing the segment C from unmatched to match. Similarly, Figure 84, on the right, illustrates the case when two preceding (A, and B) and succeeding (D, and E) segments are unmatched, whereas a single segment (C) in between is matched. Following the same strategy like in previous example, the segment C in between is changed from matched to unmatched.

Here there are two intermediate cases, a case of a second and a second to last segment in a trace. In Figure 85 on the left, four segments neighbourhood of the segment (B) in the trace implies one preceding (A) and three succeeding segments (C, D, and E). In case of second to last segment (D) in the Figure 85 on the right, three preceding (A, B, and C) and one succeeding (E) segments should be considered.

Figure 85 : Asymmetric neighbourhood of four segments

From our point of view, these two cases should be treated in the same way like symmetric neighbourhood of four segments. By analysing Figure 17, it seems more likely that those segments in black have the same matching result as their four neighbours, than a different one. Let us mention that first and last segment in a trace will not be considered by this criterion, since they have completely asymmetric neighbourhood. Former has a lack of preceding neighbours while latter has a lack of succeeding neighbours. Hence, it is impossible to apply the same logic like in previous cases. Additionally, their matching results have negligible effects on matching result of other segments in the trace.

2.3.2.3 Length criterion

According to data specification, road types in our focus can exist in the database with certain scale only if they respect a minimal length criterion defined in the specification. Thus, the assumption for this criterion is that for adding a new segment in the dataset, the segment should have a minimal length.

The criterion is defined as follows. First, initial VGI segments are merged according to their matching results (matched or unmatched). Since the detection of missing roads in the authoritative dataset is in our focus, attention will be paid on unmatched aggregated segments of VGI traces as potential candidates for highlights of missing roads. For each unmatched aggregated segment, the algorithm

verifies if the length of the segment is higher than a threshold. If it is the case, then the segment is considered a candidate for a missing road.

Thus, the final result of data matching is the set of unmatched aggregated segments having lengths higher than a threshold.

2.4 Data matching results

In this section, we present the analysis for choosing an optimal buffer size for our matching approach, and visual analyses of matching results. First the presented criteria are formalized, and then the matching approach is applied.

2.4.1 Formalization of criteria

2.4.1.1 Angle criterion

When angle is used as a similarity measure between segments, usually small values are used to indicate that two segments match, such as 15° in the work of Zhang et al., (2005). Since here angle is used in filtering of first matching results, we propose a higher threshold in order to be less restrictive keeping more potential candidates to be evaluated by next criteria. The threshold is empirically sets to 30°. Hence, all VGI segments having an angle wider than 30° with their selected candidates will be qualified as unmatched. Corresponding example is illustrated in Figure 86.

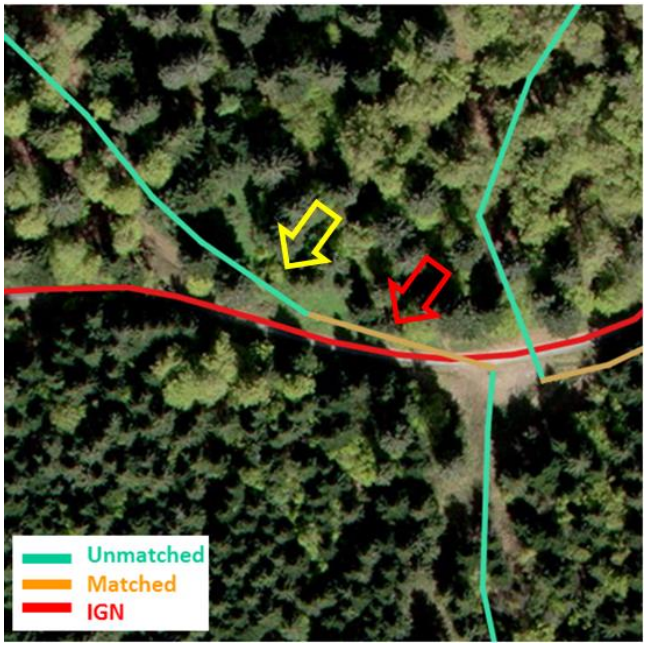


Figure 86 : Angle criterion application result

In this situation, the segment pointed by a red arrow should be matched with IGN road since the angle between them is 24°, whereas the segment pointed by a yellow arrow should not be matched since it composes the angle of 35° with corresponding IGN road.

2.4.1.2 Neighbourhood criterion

To define 'short segments' in case of neighbourhood criterion, we propose a threshold of 100 m. Hence, all segments shorter than 100 m will be considered by this criterion. Segments longer than 100m fulfil BDTopo® specification length criteria to exist as independent paths (BDTopo® Specification 2.2, April 2017), thus it would be difficult to justify their matching result change. Especially in case of unmatched segment, since if its matching result is changed from unmatched to matched, it can cause a direct loss of candidates for missing roads. An example of matching before and after applying neighbourhood criterion is presented in Figure 87.

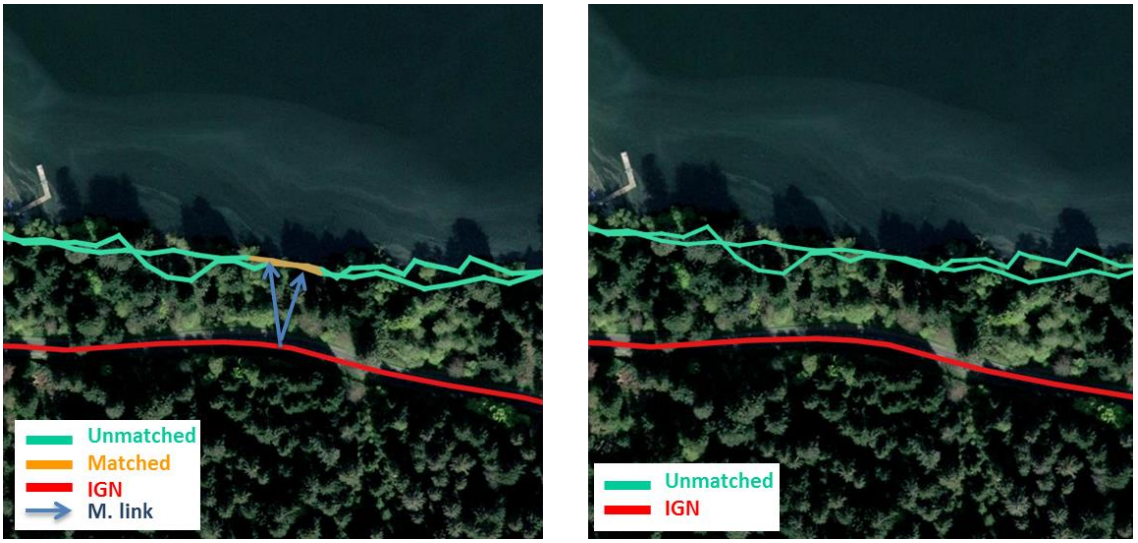


Figure 87 : Neighbourhood criterion application result

It can be seen that all segments of these two traces are unmatched and potentially a missing road, whereas only one segment in both traces is matched to a nearby IGN road. These first matching results are obviously wrong. This is changed after applying the neighbourhood criterion (see Figure 87 on the right).

2.4.1.3 Length criterion

In defining length criterion threshold, we need to consider a specific authoritative data that we use and related data specification. As already presented, our test authoritative data is IGN data contained in BDTopo® database with the last specifications defined in (BDTopo® Specification 2.2, April 2017). According to it, road types in our focus can exist in the BDTopo® database (1:25.000) only if their length is bigger than 100 meters. Thus, in order to meet this criterion, we apply the same threshold on our data. As a result, all unmatched aggregated segments with length below threshold are eliminated from the dataset of candidates for missing roads.

Like illustrated in Figure 88 unmatched aggregated segment presented in green should not be considered anymore as a candidate for missing road since its length is 87m, whereas an aggregated segment presented by violet-blue line in the Figure 88 having length 179m should be considered as missing road candidate.

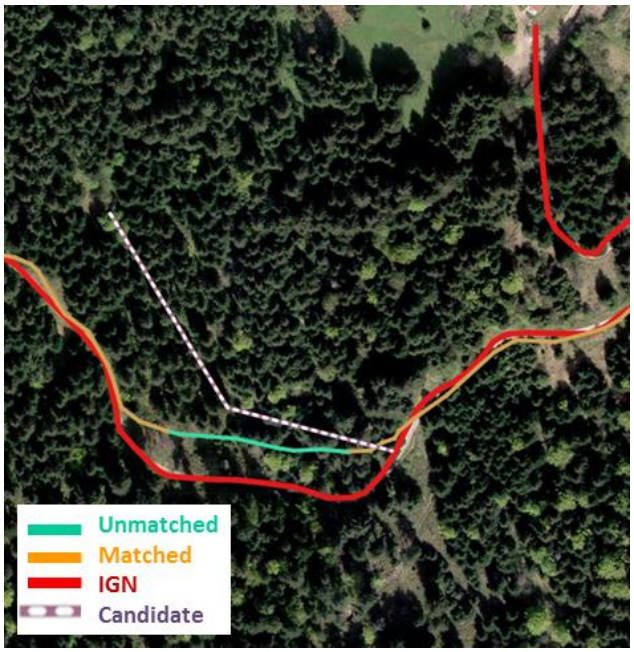


Figure 88 : Length criterion application result

2.4.2 Choosing an optimal buffer size

After choosing two buffers size matching strategy, the question arose which sizes should be used. In order to make most appropriate decision, few sizes have been tested. Since the quality of VGI traces varies significantly, the results of matching are very sensitive on buffer size used. On the one hand, if the size is too small, a lot of segments can be falsely detected as unmatched. On the other hand, if the size is too big, a lot of segments can be falsely detected as matched.

In order to test different buffer sizes and choose the most appropriate with respect to our dataset an interactive data-matching is carried out on a test area in order to define the 'ground truth'. For that purpose, 41 traces are examined. Total length sampled is 920 km which is around 10% of total input data length. The test zone is shown in Figure 89 where traces in purple are test traces.

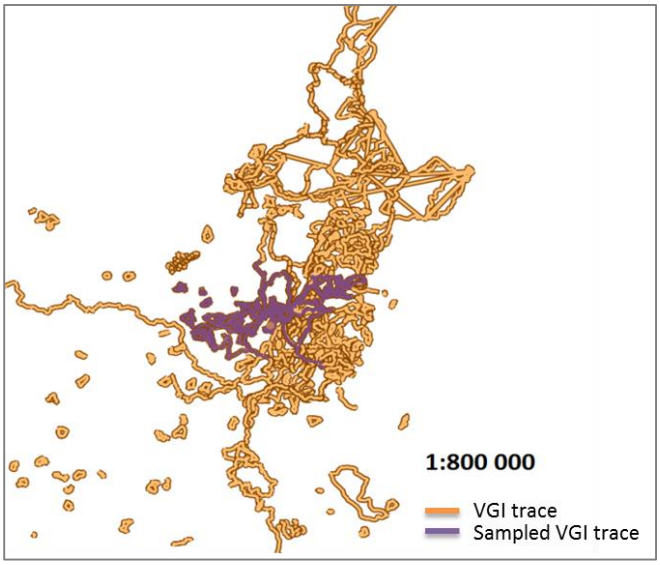


Figure 89 : Test area for validation of matching results

Examination of test traces was done based on most recent orthophoto images and IGN maps supported by expert's knowledge. As a result, 87.5 km of VGI traces were interactively classified as unmatched (9.5%). This is a significant amount of potential updates of BDTopo®. It quantitatively shows two important aspects of our work: a big need for alternative sources in authoritative data updates and a huge potential of VGI data as such data source.

Having applied the approach on the test area, the following results were obtained with different thresholds (see Table 9). Values in the column Buffer represent size of buffers depending on the accuracy, so as first value is a buffer size for low accuracy segments, whereas second value is a buffer size of low accuracy segments.

Table 9 : Results of different buffer sizes used in Data matching

Buffer (m)	R (%)	P (%)	F1
15,15	**85**	**58**	0.69
15,30	**81**	**65**	0.72
20,20	**81**	**70**	0.75
30,30	**72**	**80**	0.76
20,40	**77**	**84**	0.80
30,60	**67**	**87**	0.76
40,40	**65**	**91**	0.76
60,60	**51**	**98**	0.67

Precision and recall of candidates of missing roads (unmatched VGI segments) are calculated based on the length of correctly matched / mismatched segments.

The highest recall (85%) was obtained by fixed buffer size of 15 meters, whereas in the same time this buffer size resulted in the lowest precision (58%). On the other hand, the highest precision of 98% was achieved by biggest buffer size applied in fixed threshold strategy (60m), whereas the recall remained very low (51%). It is quite expected knowing the relation between precision and recall – an increase of precision results in a decrease of recall and vice versa. Figure 90 shows the distribution of precision and recall with respect with the buffer size.

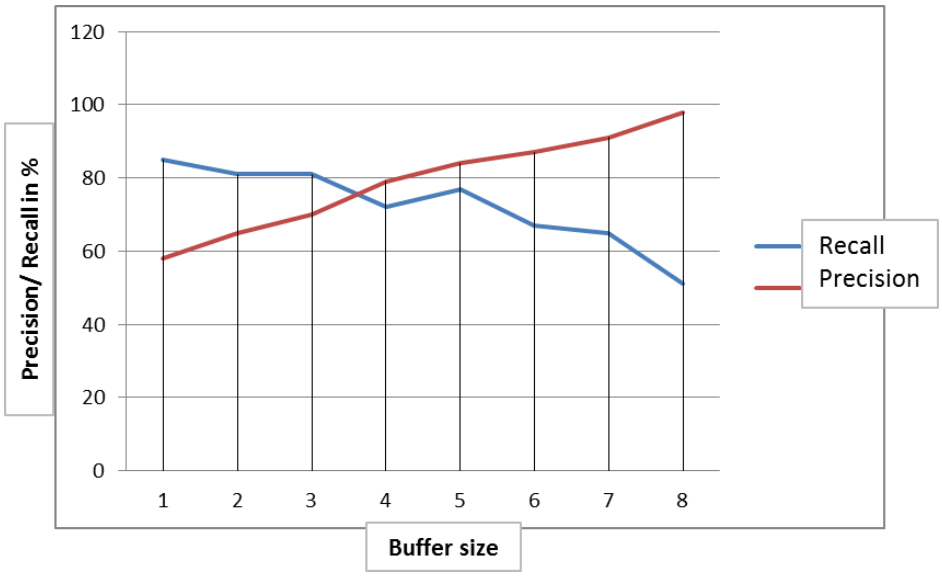


Figure 90 : Precision and recall trends with respect with buffer size

We can observe that by enlarging buffer size the precision is increasing gradually. Overall, the increase is progressive, except in the zone between buffers (30,30) m and (30,60) m where the precision is raising slightly. In the same time, recall has a negative trend as the buffer size is enlarged, except for the buffer size (20,40) m, when it experienced a small rise. These trends are not common when the results of classical matching problem are presented. In that case, precision and recall of matched objects are calculated, thus, precision is decreasing as the buffer size is enlarging, whereas the recall is increasing. However, trend of precision and recall of unmatched objects shown in the Figure 90 are different, since smaller buffer size results in more objects unmatched with lower precision, whereas, bigger buffer size results in less objects that are unmatched with higher precision. For instance, by applying 60m fixed buffer size, mainly VGI segments with a distance bigger than 60m from IGN roads remained unmatched (high precision), whereas the minority of them with smaller distance were matched even if some of them were good candidates for missing roads (low recall).

Even the precision is important for our work, we find recall more important. Indeed, the automated part of the detection of missing roads is supposed to take into account most of possible candidates for missing roads, i.e. to have high recall in order not to miss potential candidates to update. If they are falsely matched to IGN road, they will be finally eliminated from the list of potentially missing roads. On

the other hand, precision can be improved during visual verification by a surveyor of new detected roads from already reduced list of candidates. However, a certain (satisfying) level of precision should be ensured besides good recall in order to minimize visual checking time. Thus, the highest level of recall without a significant drop of precision in the same time is for a buffer size (20,40) m, where the F1 score is the highest (0.8). Therefore, compared to other buffer size scores, this quality adapted buffer size seems to be best compromise solution. We cannot deny that the highest recall is scored by fixed buffer size of 15m, but the precision is too low, only 59%. In this case, almost every second candidate for missing roads would be false candidate for missing road.

In terms of fixed or adapted buffer size comparison, Table 9 shows that according to F1 score, the quality adapted sizes performed better compared to corresponding fixed sizes (e.g. (20,40) m versus (40,40)m or (20,20)m). In addition, it is expected that with the improvement of the results of detection of low accuracy points, the results of data matching will improve as well.

2.4.3 Visual analysis of matching results in the sampling zone

As discussed in previous sub-section, buffer size (20,40) m provided the highest results among other sizes tested. Thus, we selected this size as a buffer size parameter for our matching approach. In this subsection, results of this buffer size in the test are analysed in detail.

Some of examples of efficient matching results in terms of candidates for missing roads detection are presented in Figure 91, where we can see two examples of very precise matching process, by means of which two candidates for missing roads were detected.

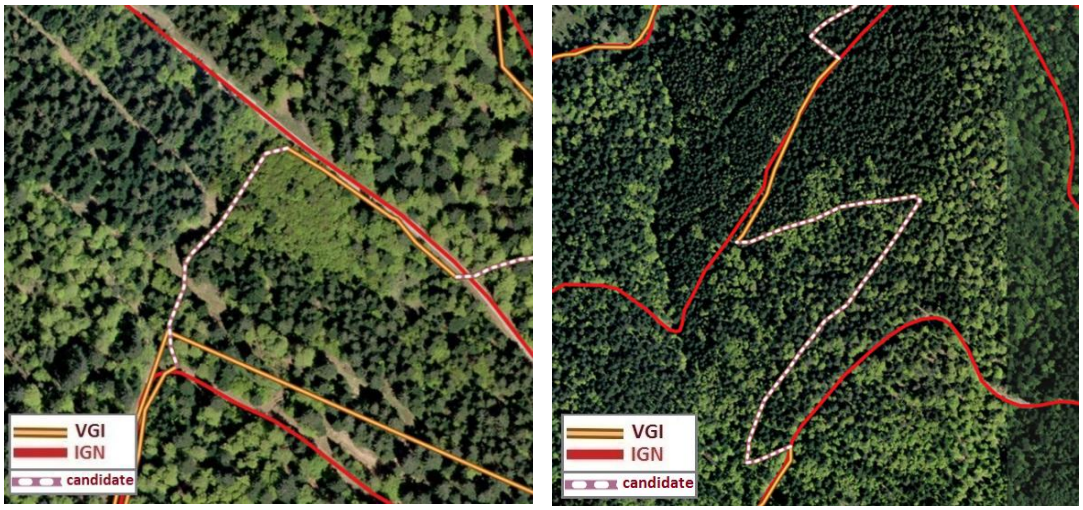


Figure 91 : Good matching results

However, there are some cases where the limits of the approach come to the fore. Most of false positive results, i.e. false candidates for missing roads are caused by variations in VGI traces quality and miss-classified low accurate points. A typical example is illustrated in Figure 92 where a part of VGI trace is

wrongly recognized as a candidate for missing road because the accuracy evaluation falsely considered it with high accuracy, which is corrupting the matching process.

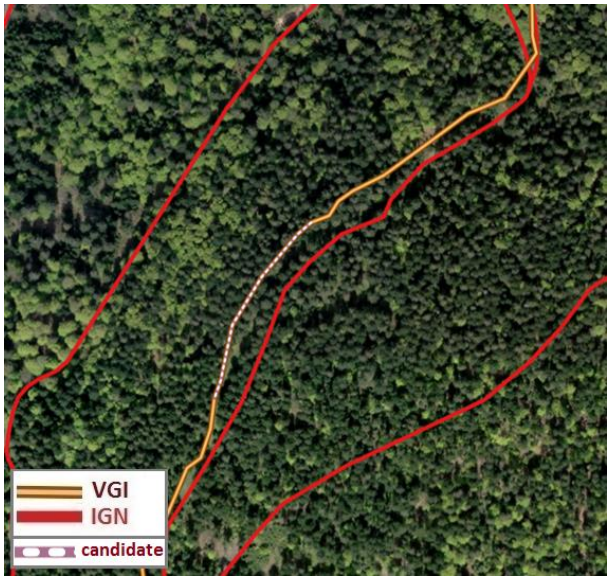


Figure 92 : False positive candidate for missing road

False negative results are also caused by errors in low accuracy point detection, but also due to buffer size design. The example is shown in the Figure 93. In this specific case, points were correctly classified as low accuracy points, thus the buffer of 40m was correctly constructed. However, median distance between IGN road and VGI trace in question is about 15m. Such case, potential missing road and IGN road very close to each other is not very frequent in general, but is not negligible within false negative results.

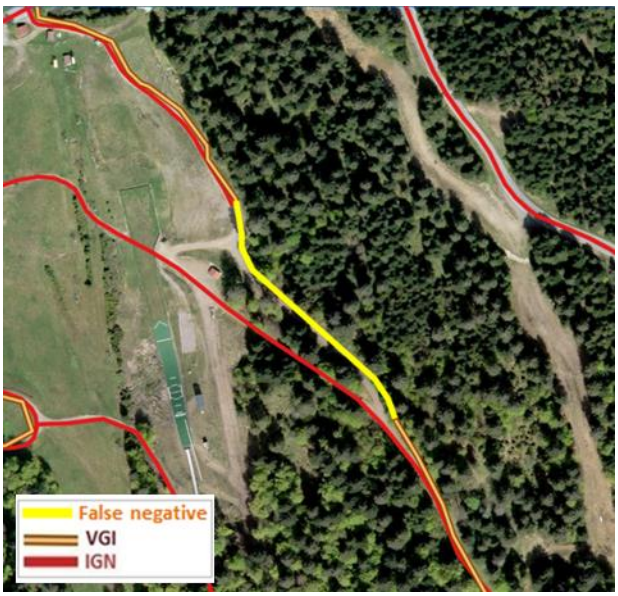


Figure 93 : False negative candidate for missing road

A certain uncertainty of results validation also exists. For some cases manually classified 2.4.2. It is very difficult to make a clear decision if they are missing roads or not. Some of typical cases of validation uncertainty are also presented in Figure 94.

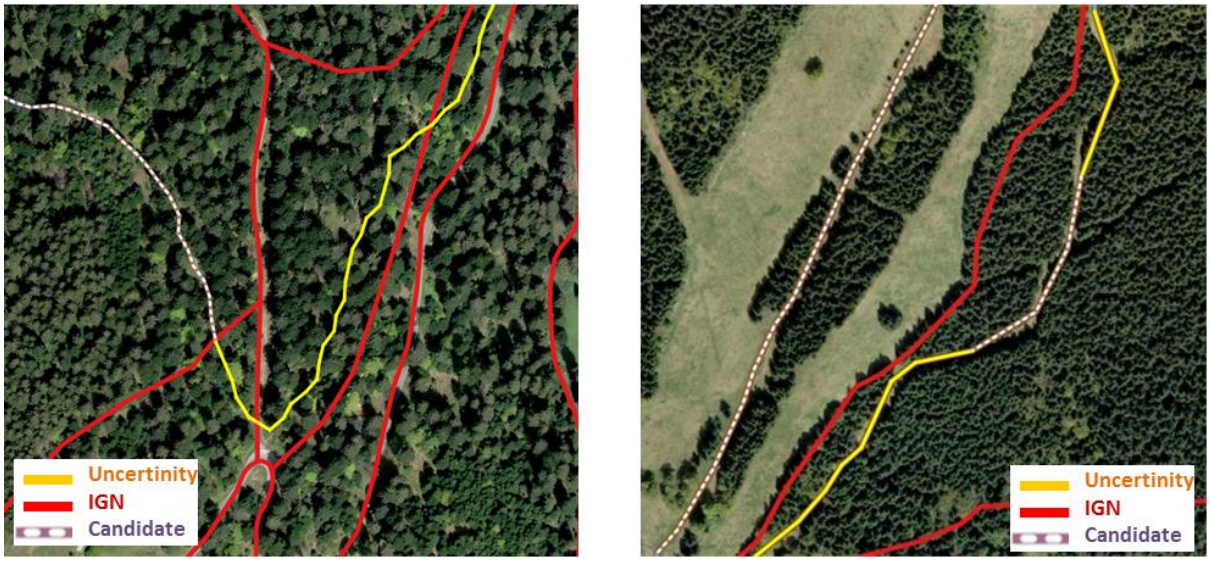


Figure 94 : Validation uncertainty

In both figures it is difficult to decide if segments in yellow are potentially missing roads or just segments with low accuracy. It is also especially challenging to decide which are the points where a missing road begins and ends, such as illustrated in the figure on the right. VGI trace in the right figure is sometimes very close to IGN road (less than 2m) and sometimes very far (more than 40m). On top of that for parts far from IGN roads there is a visual confirmation in orthophoto that a road exists.

The following results show the interest of the filtering phase after the selection phase. The precision and recall during the matching process by buffer (20,40)m are presented in the Table 10.

Table 10 : Matching results in the test area

Matching phase	R (%)	P (%)	F1
Selection	**68**	**68**	0.68
Angle criterion	**74**	**75**	0.74
Neighbourhood criterion	**77**	**80**	0.78
Length criterion	**77**	**84**	0.80

Application of angle criterion improved recall and precision for 6 and 7% respectively. After application of neighbourhood criterion, recall rose for 3% reaching 77%, whereas precision rose for 5% reaching 80%. Finally, after length criterion application, recall remained the same while precision reached 84%.

We can conclude that angle criterion was most successful in filtering matching results. In total, both criteria enhanced recall and precision for 9 and 16% respectively. Thus, F1 score was improved for 0.12.

2.5 Conclusion

In this chapter, we presented the current state of the art of data matching and our approach for matching authoritative road network and VGI traces in order to identify candidates for missing roads in authoritative network. In the state of the art the relevant research in past few decades was presented paying special attention on matching VGI data. The focus was also on exploring the possibilities of employing existing matching methods in detection of missing roads in authoritative road network using VGI traces. Analysing exiting matching methods led us to the conclusion that most of them are not suitable for our purpose mainly due to the lack of metadata of VGI traces that limits application of many existing criteria (e.g. attribute criteria). Buffer based method was identified as a solid basis for solving our matching problem. However, existing buffer based methods, all employing fixed single buffer size, were not adapted to VGI traces matching due to the already mentioned issue 'variations in quality'. Thus, the solid basis identified in buffer matching principle was then extended and modified to fit VGI traces quality issues.

Therefore, the approach we proposed for matching VGI traces and authoritative road network was a buffer based solution with flexible buffer sizes. Sizes (smaller and greater) are adapted to the quality of VGI segments evaluated by the approach for the detection of low accuracy points presented in the Chapter 1. The smaller buffer size is used in matching good accuracy VGI segments to authoritative road network, whereas the greater is used for matching good accuracy VGI segments. The approach was validated in the test zone by comparing matching results of the approach and results obtained by visual checking. Based on the results of the validation, best fitting sizes (20,40m) were selected. They outperformed other sizes tested, especially fixed buffer sizes, which confirmed a need for applying different buffer sizes within a single dataset.

Chapter 3

Decision making

3 Decision making

This section is devoted to making final decision if proposed candidates for missing roads are real missing roads or false alerts for updates. First of all, we discuss the possible interpretations of unmatched aggregated VGI segments after their matching to authoritative road network. Further we propose relevant criteria measuring the degree of confidence on potentially missing roads. Finally, criteria are formalized and applied on candidates for missing roads and the final decision if a missing road should be proposed as an update or not is made.

The data matching process allows identifying potential candidates for missing roads defined as unmatched aggregated segments. Determining which candidates represent the real updates is a difficult task. Indeed, the fact that a part of VGI trace is not matched to IGN road does not necessarily mean that a new update should be highlighted. For example, for the unmatched VGI segments as illustrated in Figure 95 difficult interpretations are possible: i)) GPS error: the deviation of a part of VGI trace from IGN road is caused by GPS error measurements; ii) New road: the deviation of a part of VGI trace is following an existing road in the real world which is not represented in the database; iii) Road modification: the deviation is due to a modification of the existing road. The last two cases are considered as missing roads, thus potential updates.

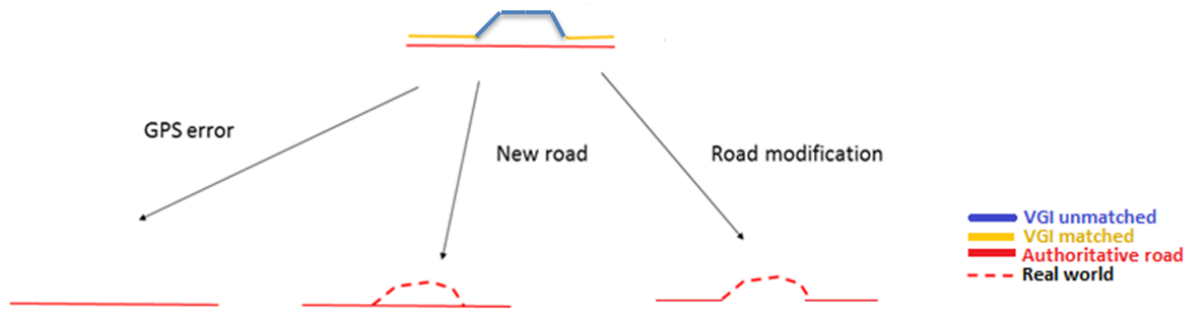


Figure 95 : Unmatched VGI segment – possible interpretations

In order to make a final decision for determining if a candidate for missing road is a real update or not, more information should be involved in decision making, from information regarding VGI traces only, to information regarding authoritative data specification. For this purpose, we propose a multi criteria analysis which consists in combining different criteria in order to define a degree of confidence that measure to which extent we can trust to the proposed missing road.

This chapter is organised as follows. The proposed criteria are first described and modelled. Then, the criteria are combined so that the final scores for qualifying the reliability of each candidate for missing road are obtained. Finally, the results (missing road detected) are presented and validated.

3.1 Definition of criteria for decision making

Taking into account the nature and specificity of VGI traces, we propose four relevant criteria in final selection of potential missing roads: quantity criterion that measures the quantity of candidates for missing roads, quality criterion that takes into account the spatial quality of candidates for missing roads, actuality criterion that uses the actuality of candidates for missing roads and a continuity criterion examining how closely a VGI trace follows the same authoritative road.

The first three criteria are geometric criteria, whereas the last one is an attribute criterion. Criteria are distinct, thus there is no overlapping between them. That should provide more reliable results by avoiding that one aspect of decision making is considered more than once. In order to combine qualitative and quantitative criteria, all scores are normalized between 0 and 1.

3.1.1 Quantity criterion

This criterion refers to the number of VGI traces representing the same missing road. It is an important indicator of existence of a road in the reality. The general philosophy of this indicator is: “The more candidates following the same path exist, the more chances that a road exists in the reality”.

The lack of metadata already mentioned limits us in considering more information than only regarding VGI traces quantity. However, some extra information would be very useful in this context. First, the question if traces are coming from the same contributor is important. Chances that a new road is created are higher if traces indicating that are collected by different persons. Second, time when the traces are collected is important. If majority of traces are collected during a short time period for example on the same day, the clue of new road creation is less reliable than if traces are collected in different periods of time (e.g. different days during a month).

The criterion quantity considers thus a number of traces representing the same missing road. In order to determine that traces are representing the same missing road, we propose a method based on geometric information and using topological and angle criteria. The method is applied as follows.

First, for each candidate for missing road a buffer is created. All other candidates for missing roads that have more than a certain percent of their lengths within the buffer are considered as candidates for the same missing road.

Second, directions between the candidates are computed as proposed by Olteanu (2008). Only the candidates having similar orientations are finally considered as those representing the same missing road.

Due to the weak redundancy of VGI traces in mountainous areas, the criterion C_1 is formalized as follows:

Rule 1: if only one candidate for a missing road exists, then criterion score is equals to 0.25.

Rule 2: If two candidates following the same path exist, then criterion score is 0.5.

Rule 3: If more than two candidates follow the same path, then criterion score is 1.

The principle is simple and can be easily modified in case of higher redundancy between VGI traces by introducing more criterion scores and modifying the redundancies.

3.1.2 Accuracy criterion

This criterion is based on the results of detection of low accuracy GPS points. We start from the general hypothesis that a detected missing road composed of good accuracy points is a more reliable update than one composed of low accuracy points.

An example illustrating why quality of traces is important in final decision making is presented in Figure 96. Even if our data matching approach takes into account the fact that segments of the trace have low accuracy points, there are still situations where low accuracy segments can negatively bias the results of data matching and subsequently candidates for missing roads. For instance, if accuracy of VGI segments is lower than 40m (size of a buffer for low accuracy segments matching), the VGI segments would be falsely recognized as candidates for new roads, like it is shown in Figure 96:

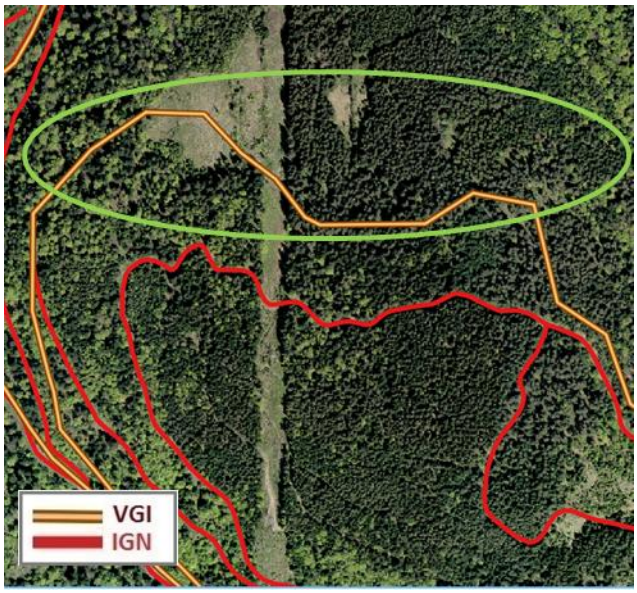


Figure 96: Very low accuracy segments (labelled by green ellipse)

As we can observe in Figure 96 this is especially critical in cases of very low accuracy segments. Very low accuracy segments labelled by a green ellipse in this case deviate from 116m to 204m from the corresponding IGN road. Thus, after data matching these segments were recognized as candidates for a missing road even it is clear that their positions are only a matter of low accuracy. Additionally, good accuracy segments provide more precise results of data matching. Hence, candidates for new roads having good accuracy are more reliable; sometimes having even better geometric quality than IGN roads (see Figure 97).

Figure 97 : VGI traces with better quality than an IGN road

For example, the two VGI traces presented in Figure 97 on the left are actually more accurate than IGN road when compared to most recent othophoto from 2013 used in validation. They are precisely following a path pointed out by a yellow arrow in the Figure 3 on the right, whereas the IGN road is misplaced for more than 15m. One may imagine that the IGN road comes from a source with low accuracy (like an old map), or that the real road has been shifted in reality but this has not been updated in the IGN dataset.

Therefore, presented examples justifies a need for using spatial quality as a criterion in decision making process so that good quality segments should be assigned higher degrees of confidence than low quality ones.

The accuracy criterion score noted C_2 is computed for each candidate for missing road (unmatched aggregated segment) as the percentage of length of good quality segments in its total length.

3.1.3 Actuality criterion

In terms of detection of updates, we consider actuality of traces as a highly important aspect. The assumption for this criterion is as follows: “More recent a trace is, more relevant update will be”. In our context refers to the year in which a trace was recorded. We consider that a higher temporal resolution is useless.

It is important to stress that the criterion is independent of point’s timestamp. If the timestamp for each point is sometime missing, we noticed that there is always a global timestamp for the trace in the GPX file. Indeed, XML schema of GPX files requires a date when a file is created. Usually this is a date of the export of a VGI trace not the data when the trace is recorded. The date is underlined by yellow and can be seen in the Figure 98. Normally, the time discrepancy between the date of record and export are few

days, in extreme situations few weeks. Thus, the information we are looking for here (a year) is not affected by this discrepancy.

```
<?xml version="1.0" encoding="UTF-8"?>
<gpx
 version="1.0"
creator="GPSBabel - http://www.gpsbabel.org"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.topografix.com/GPX/1/0"
xsi:schemaLocation="http://www.topografix.com/GPX/1/0 http://www.topografix.com/GPX/1/0/gpx.xsd">
<time>2010-03-12T09:29:16Z</time>
<bounds minlat="47.985117000" minlon="6.704008000" maxlat="48.000664000" maxlon="6.711577000"/>
<trk>
  <name>track</name>
<trkseg>
<trkpt lat="48.000664000" lon="6.704621000">
  <ele>410.000000</ele>
</trkpt>
<trkpt lat="47.999909000" lon="6.704957000">
  <ele>413.000000</ele>
</trkpt>
<trkpt lat="47.998695000" lon="6.704900000">
  <ele>432.000000</ele>
</trkpt>
<trkpt lat="47.997863000" lon="6.705715000">
  <ele>437.000000</ele>
</trkpt>
<trkpt lat="47.997582000" lon="6.706886000">
  <ele>436.000000</ele>
</trkpt>
```

Figure 98 : An example of GPX file of a VGI trace having a global timestamp

The intermittent nature of mountain path we are trying to update has been already pointed out, as well as difficulties caused by it. Since footpath, bicycle and tractor road are not very perennial in mountainous areas, the date of the collected VGI traces is essentially important. A road that has a trace collected along it in a present year has more chances to still exist, than a road that has a trace collected few years ago. In fact, they can both exist, but a road with a trace collected in the recent past has more possibilities to resist seasonal changes of relief in mountains. The relevance of candidates for new road is designed as illustrated in the Figure 99.

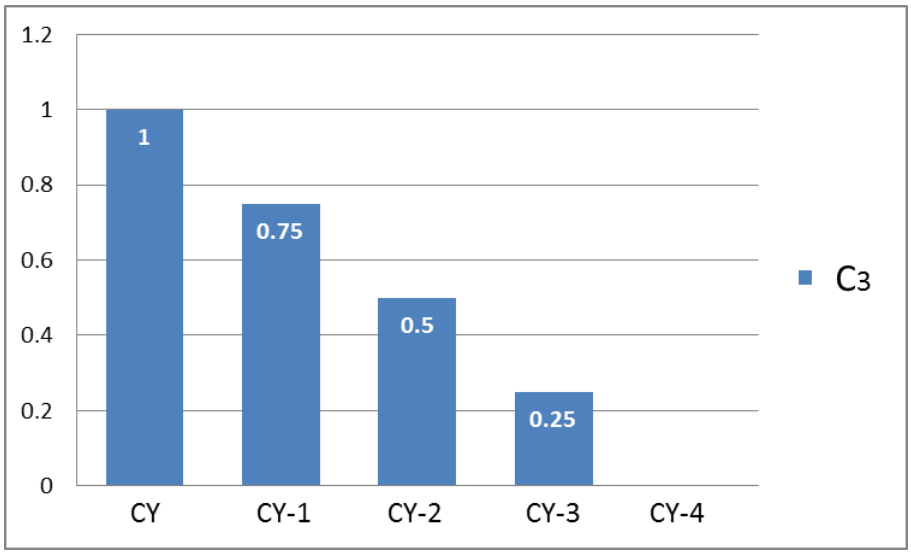


Figure 99: Defining actuality criterion

Maximal score for the criterion C_3 can be assigned only to traces collected in current year (CY in the graph), whereas for each previous year (CY-n, where n ∈ (1-4)) the score is reduced by 0.25 until four years in the past. Traces collected four years ago and before bring 0 relevance to a potential update in decision making.

3.1.4 Continuity criterion

This continuity criterion concerns a very specific case that may occur in matching VGI traces and authoritative datasets. As a result of many factors, such as human behaviour or low accuracy of VGI segments, a case where a VGI trace follows an existing IGN road along most of its length except in some short parts may occur, as illustrated in Figure 100.

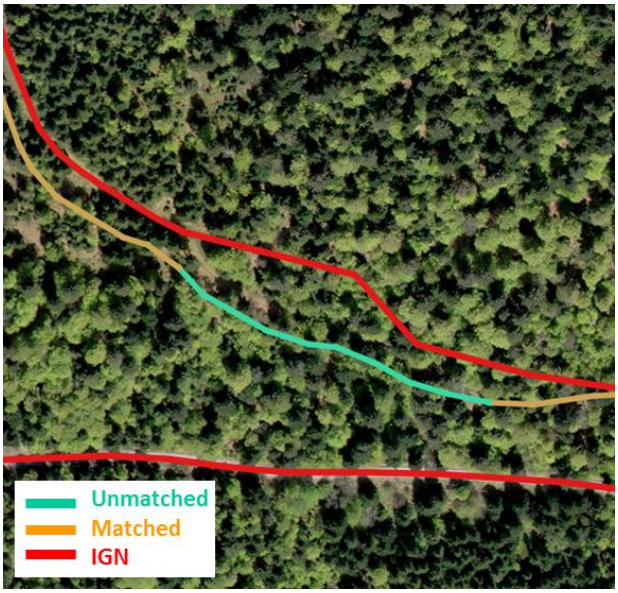


Figure 100 : A VGI trace following an IGN road

If a VGI trace follows an IGN road for a long distance, then it is interrupted for a single part and then follows again, the VGI trace probably corresponds to the IGN road and has fewer chances to really be a missing road. Generally, there are two main reasons why such interruption in following IGN road occurs. The first one is due to significant variations in traces quality or due to human behaviour. We name this case 'candidate for missing road that follows the same authoritative road'. The second one is due to the fact that the unmatched aggregated segment follows a real road which is not represented in the dataset. One of such examples is presented in Figure 101, where we can notice a part of VGI trace not following IGN road for a while, but also a confirmation of a potentially new road with respect to the orthophoto. Thus, this part of the trace should be highlighted as a missing road. We name this case 'candidate for missing road that does not follow the same authoritative road'. Since cases are diametrically different, this criterion is more restrictive than the others. Thus, only two scores are possible: minimal for the candidates for missing road that follow the same authoritative road and maximal for those that do not follow the same authoritative road.

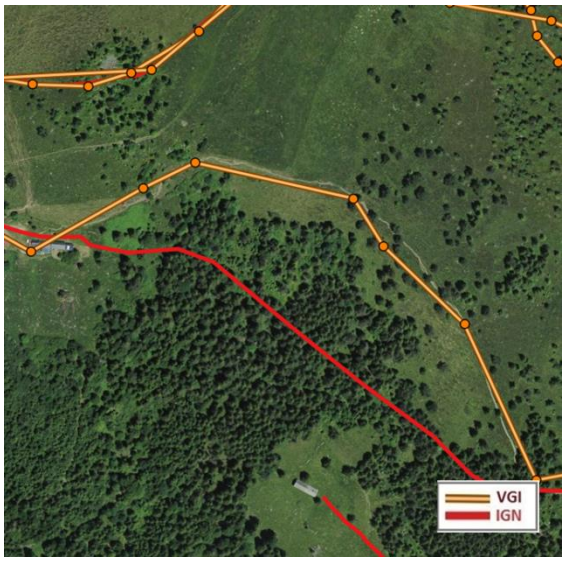


Figure 101 : A case when a part of VGI trace is a missing road even the rest of the trace follows IGN road.

In order to distinguish well these two situations, the criterion is defined as follows.

First, we identify candidates for missing roads (unmatched aggregated segments) whose neighbours are following the same IGN road.

Second, the length of the candidates for missing road and their distance to the nearest IGN road are used to distinguish between real update and deviations due to quality and behaviour. We believe that relatively 'short parts' of a trace not following an IGN road (candidates for missing road), which are in the same time relatively 'close' to it are rather results of human behaviour or variations in quality than a real candidate for a missing road. 'Short parts' are candidates for missing roads having length less than a threshold N. The closeness between the candidate for missing road and the nearest IGN roads is determined by means of Hausdorff semi distance. Thus the score of this criterion named C4 is defined as follows (see equitation 10):

$$\text{If } (L < N \text{ and } H_d < P) \Rightarrow C4=0, \text{ else } C4=1 \qquad (10)$$

Where L is the length of candidate for missing road, N a threshold for determining 'short parts', H_d is the Hausdorff semi distance between candidate for missing road and the nearest IGN road and P a threshold for the closeness between the candidate for missing road and the nearest IGN road . The nearest IGN road is selected by buffer growing method, i.e. buffer is enlarged until the first IGN road is selected. Buffer starts to grow from 20m for good accuracy aggregated segments and from 40m for low accuracy aggregated segments.

3.2 Combination of the criteria

In this section, we describe how the four criteria are combined in order to calculate a degree of confidence for each missing road.

Defining to which extent each criterion influences a decision making may be done through a weighting criteria process. Existing processes to determine weights may be classified in as subjective and objective methods (Roszkowska, 2013). Subjective methods determine the importance of criterion only based on expert's (decision maker) knowledge or preference (Toloie-Eshlaghy 2011, Olteanu-Raimond et al., 2015). On the other hand, objective methods define the weights of criteria through mathematical calculation using objective information in a decision matrix where a decision maker is not involved. They are usually applied when there are many criteria to compare, thus proposed mathematical calculations facilitate the weighting process (Diakoulaki et al., 1995; Chehreghan and Abbaspour, 2017).

In our work, we rely on four criteria in a decision making process. The proposed four criteria are very different in terms of their nature. Thus, it is difficult to determine their relevance on the final decision making without expert's knowledge regarding their influences on final result (degree of confidence of detection of missing road). Therefore, Equal Weight method (EW) seems as a suitable solution since it requires minimal knowledge about priorities of criteria and minimal input of decision maker: "If the decision maker has no information about true weights, then the true weights could be represented as a uniform distribution on the unit n − simplex of weights defined by conditions, where n − simplex of weight is a geometric object" (Roszkowska, 2013).

Thus, the weights for each of presented four criteria calculated by means of EW are 0.25.

The degree of confidence of proposed updates (missing roads) that combines the scores of proposed criteria is calculated as shown in equation 11:

$$S_i = \sum_{j=1}^{n} C_{ij} W_j \qquad (11)$$

where: S_i is a degree of confidence for *i − th* detected missing road, and C_{ij} is the normalized score of *i − th* detected missing road with the respect to *j − th* criterion and W_j is the weight of criteria *j*.

In terms of final decision results, each candidate to update is qualitatively evaluated based on the degree of confidence as follows:

- Strong confidence (if $0.66 < S_i \leq 1$).
  This means that we strongly believe that a proposed missing road should be introduced into authoritative road network
- Medium confidence (if $0.32 < S_i \leq 0.66$).
  This means that we moderately believe that a proposed missing road should be introduced into authoritative road network
- Weak confidence (if $0 < S_i \leq 0.32$).
  This means that we weakly believe that a proposed missing road should be introduced into authoritative road network

Our general recommendation to mapping agencies is that updates with strong confidence can be introduced without an extra time devoted to validation. Updates with medium confidence require visual

checking, whereas updates with weak confidence should not be processed further. Thus, only strong and medium confidence updates should be proposed to mapping agencies.

In case when there are a few candidates for a single missing road (redundancy > 1), the advantage in defining a final update proposed for an authoritative dataset is given to strong confidence candidates. Only them are considered in defining the final update i.e. its degree of confidence and a geometry. The degree of confidence is calculated as a mean of degrees of confidence of strong confidence candidates. Final geometry of the update is then defined by an aggregation of strong confidence candidates. Among many works that exist in this field like Devogele (2002), Petitjean et al. (2011), Etienne et al., (2016), we choose one proposed by Zhang and Sester (2010). Their method is developed for an aggregation of GPS traces, what's more GPS traces with low and average quality. Thus, it fully corresponds to our data. In order to determine the geometry of the final update (missing road centerline), sampling of candidates at certain distances (length of candidates) is done by putting profiles perpendicular to the candidates. The intersections of the profiles with the candidates delivers sampling points for the final missing road centerline.

Final results categorized in these 3 groups of updates will be presented and discussed in the next subsection.

3.3 Highlights of updates – missing roads detected

In this section, we present the final results obtained after combining criteria.

Before the application of the approach for decision making, we need to define the parameters used in modelling presented criteria. First, we propose parameters used in modelling candidates for the same missing road: 80% of the length within the buffer, and the maximum angle between the candidates for same missing road is 30°. Second, the parameters modelling the criteria 'Following the same authoritative road" are set to N=200m and P=30m. All thresholds were empirically defined.

After formalization of criteria, the approach was applied on the entire zone introduced in Section 1.3.1 in the Chapter 1. First, we present the analyses of degree of confidence of missing roads detected, then the visual analyses.

3.3.1 Analyses of degree of confidence

In total 727,1km of missing roads were found in the entire zone. Compared to the total length of VGI traces used, missing roads are composing 7,7%. This is for 2,2% less than in the test zone used for validation of the results defined in Section 2.4.2. It can be explained as a consequence of different spatial distribution of missing roads as well as spatial completeness of VGI traces. Compared to the total length of BDTopo® road network in the entire zone, detected missing roads represent a contribution of 2.4%.

Distribution of detected updates relevance according to proposed classification of update' degree of confidence is shown in Figure 102.

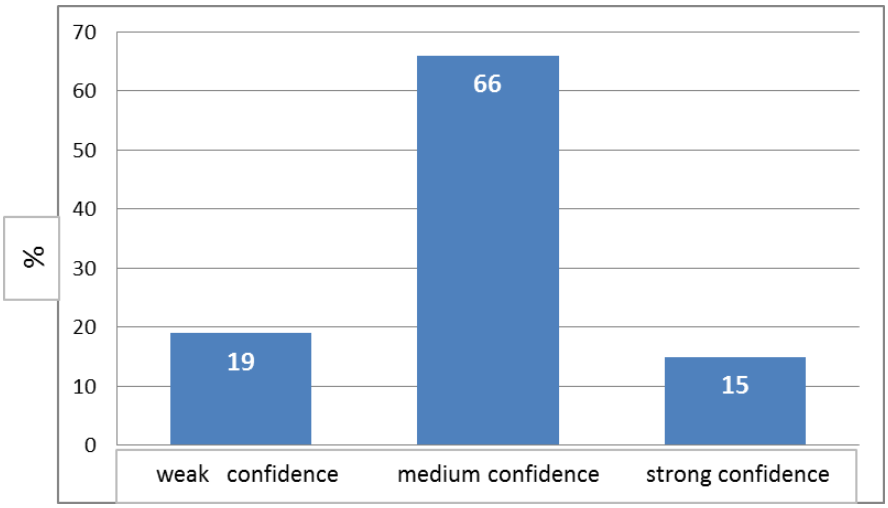


Figure 102 : Distribution of updates degrees of confidence according to presented classification

Overall, the most dominant updates are those with medium degree of confidence, almost two thirds. Together with strong confidence updates they compose 81% of total updates detected. According to our interpretation of proposed degrees of confidence, this will be the amount of proposed updates to authoritative road network. 15% of detected updates will be introduced directly to the road network as those with strong confidence, whereas 66% requires visual checking. It would be more satisfying if the balance between them was better, since the visual checking time would be reduced.

The three categories of confidence give us a global qualification of updates. In order to see if most of updates with medium confidence are closer to weak or strong confidence updates, we computed the distribution of degrees of confidence with a 0.1 graph resolution. Figure 103 illustrates the distribution of degree of confidence where the abscise represents an upper boundary of degree of confidence and the ordinate represents the percentage of missing roads.

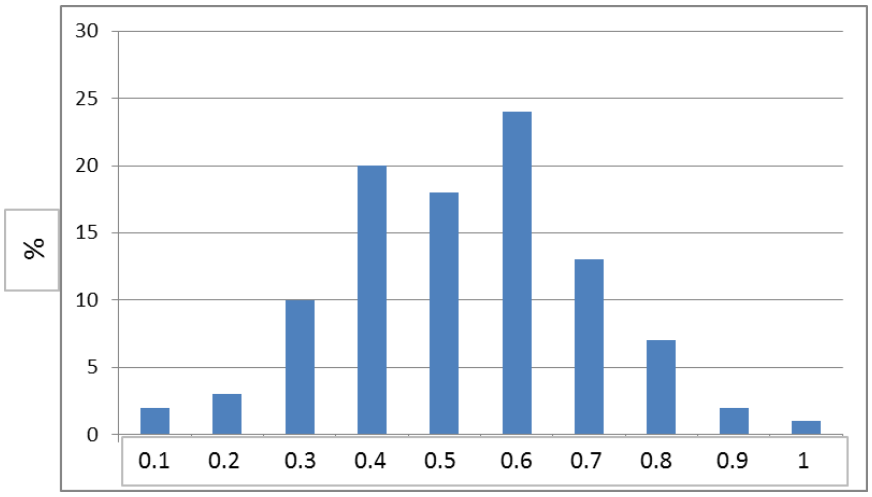


Figure 103 : Distribution of updates degrees of confidence with graph resolution 0.1

It can be seen that more medium confidence updates are closer to strong confidence updates than to the weak ones which is a positive sign of the relevance of proposed updates. The peak of the graph is for the values between 0.5 and 0.6 (24%), thus almost one quart of updates has a medium degree of confidence between 0.5 and 0.6. Relatively small amount of updates with very high degrees of confidence (higher than 0.8) is mainly caused by lower actuality of our test data downloaded two years before current year (in 2015). Thus, all updates have weaker scores for the criterion actuality.

3.3.2 Visual analysis of the final result

Visual analysis of results allows us to classify the detected updates according to two criteria:

- Configuration of missing roads related to existing network
- Interest of introducing the missing road in the network

The analyses are conducted in the test zone defined in Section 2.4.2., where 193 missing roads were detected by the approach.

Configuration of missing roads related to existing network

This aspect takes into account the main characteristics of missing roads regarding their functionality and relations with other roads from the network. Thus we identify here three cases: long independent roads, shortcuts, parallel roads.

First identified type, 'long independent road', is illustrated in Figure 104. This type of identified missing road is characterized by significant length and an independent function from other roads in the network.

Figure 104 : Long independent road

Some real examples of detected long independent roads as missing roads are illustrated in Figure 105.

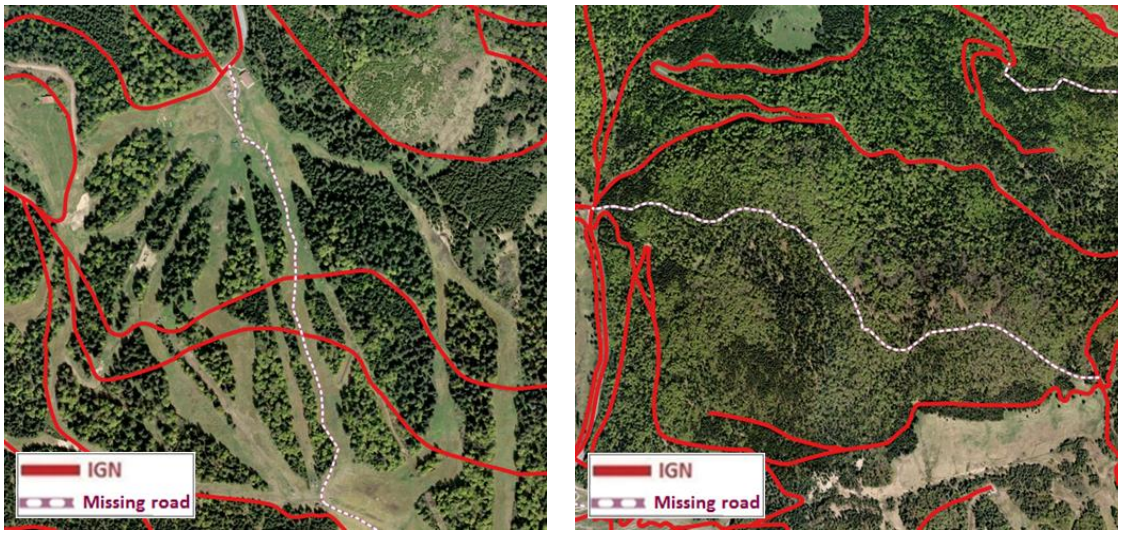


Figure 105 : Real world examples of long independent roads detected by the approach

The total number of such roads identified within 193 missing roads detected in the test zone is 81 (42%) or 51,8km.

The second identified type, 'shortcut', is illustrated in the Figure 106. Shortcuts have complementary function in the road network and usually save the time and distance for the navigation in the network

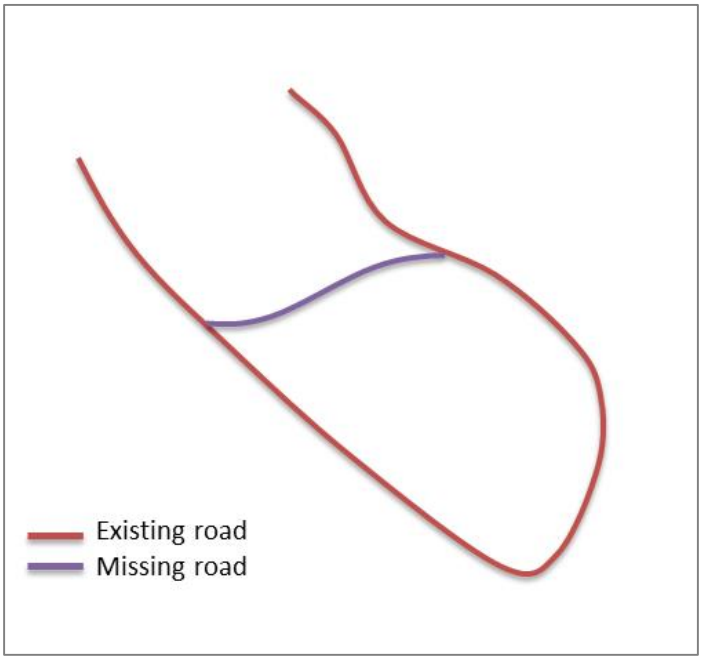


Figure 106 : Shortcut road

A real example of three shortcuts in the small zone is shown in Figure 107.

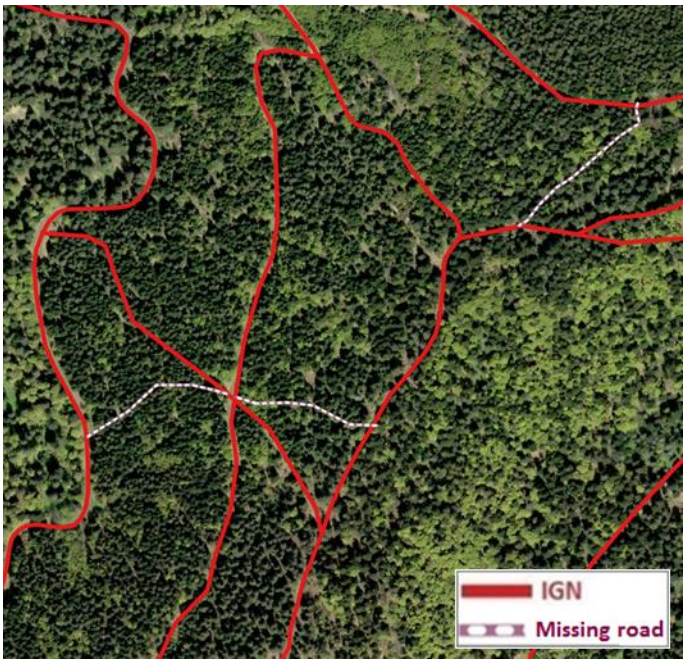


Figure 107 : Real world examples of shortcut roads detected by the approach

The total number of shortcut roads identified within missing roads detected in the test zone is 50 (26%) or 12,1km.

Third identified type, 'parallel road', is illustrated in Figure 108. This category is characterised by a part of VGI trace parallel but far away from an existing IGN road for a long distance, having the same role than existing IGN road for example connecting the same places, A and B.

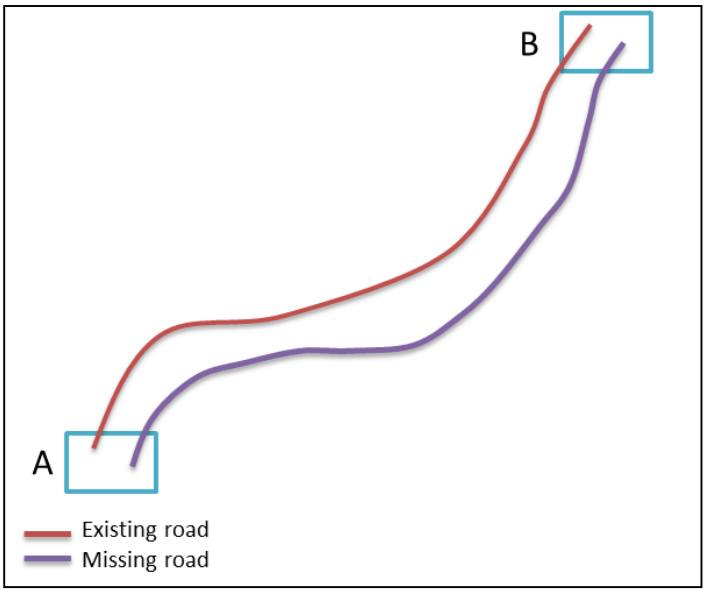


Figure 108 : Parallel roads

Real world examples are presented in Figure 109.

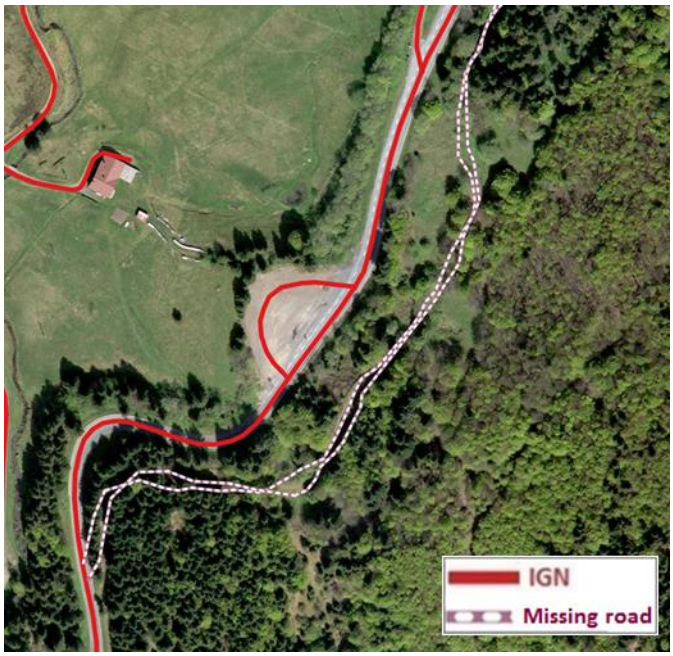


Figure 109 : Examples of long parallel roads detected by the approach

The total number of such roads identified within missing roads detected in the test zone is 62 (32%) or 22,4km.

Overall, the most represented missing roads are long independent roads which have the highest priority for mapping agencies. With slightly more than one quarter, shortcuts are not negligible in missing roads. This can be explained mostly by human behaviour i.e. saving time and distance in their mobility, thus usually taking shortcuts when it is possible. Parallel roads even being presented with almost one third, are with less interests for mapping agencies than previous two types. Most of them are having the same role as existing roads, thus according to some database specifications are not always necessary being considered as obsolete.

Interest of introducing the missing road in the network

This aspect considers the main characteristics of missing roads regarding their relevance for being introduced in the authoritative dataset. Thus, we identify here three cases: missing road that should be introduced, doubtful missing road that should not be introduced, and missing road out of specification.

First case, 'missing roads that should be introduced', are roads that fulfil IGN data specification and have a strong degree of confidence or medium (after visual checking), like roads illustrated in Figure 110.

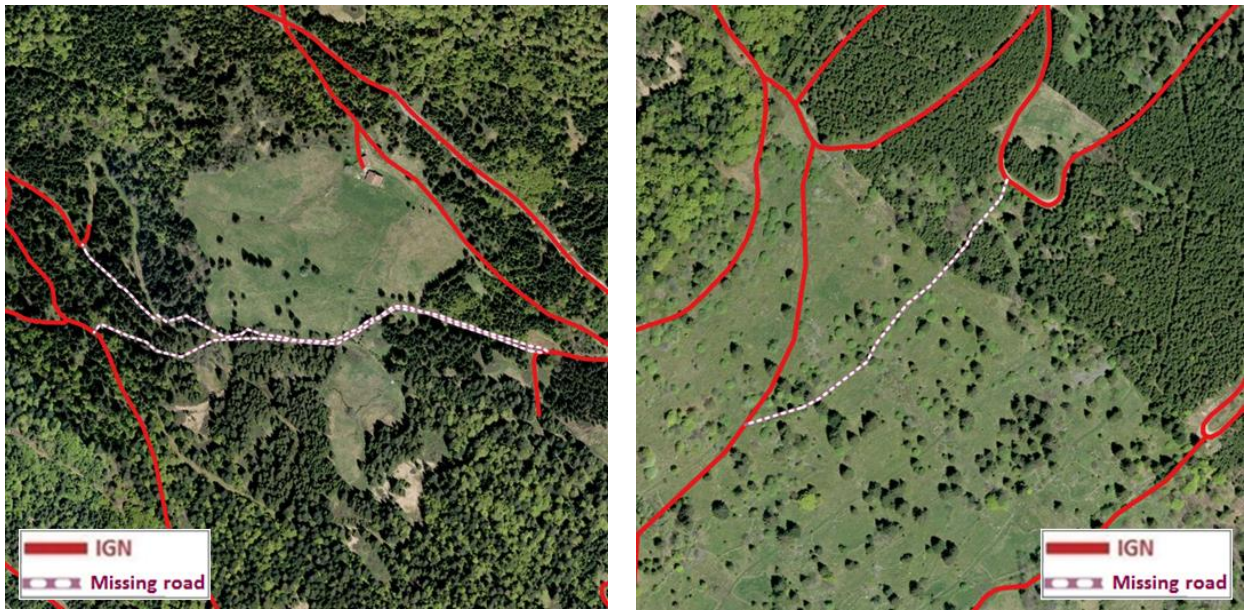


Figure 110 : Roads that are supposed to be introduced to authoritative dataset

Second case, 'doubtful missing roads that should not be introduced', are roads that comply with IGN data specification but have a week degree of confidence like road in Figure 111. They may be used as clues for missing roads, but they may not be introduced without being carefully checked and improved.

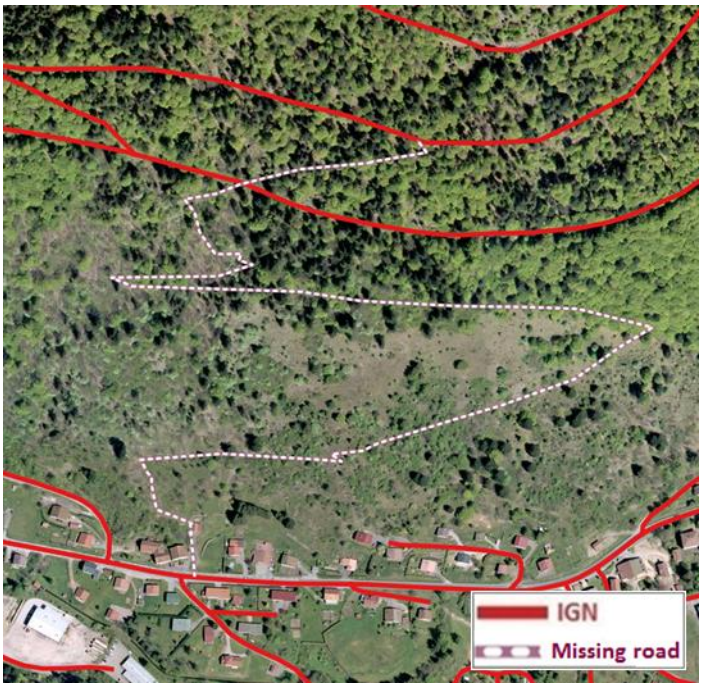


Figure 111 : Roads that are not supposed to be introduced to authoritative dataset

Third case represents missing roads that are out of IGN data specification regardless on their degree of confidence, e.g. roads with dead ends. Such example is illustrated in Figure 112.

Figure 112 : Road out of authoritative (IGN) data specification

Final integration of detected missing roads into the authoritative road network raises some very important questions. The first question is the risk or difficulty of using detected missing road. Sometimes the places where new paths are created could be very dangerous, e.g. close to the steep cliffs. One of the presented examples of detected missing roads (see Figure 105 on the left) is actually a footpath created in the sky course during the periods with no snow. This aspect of new-created roads is not even taken into account by mapping agencies. Thus, this is a point that should be examined more in the future. Second, there is a question of limited access roads, which is usually considered by mapping agencies, but not always. On the one hand, detected missing roads can be created by the people passing through some places illegally. On the other hand, they can be created by the people having an authorization to access some areas, however it does not mean that all the people can do the same using same paths.

Therefore, a final decision if an update is integrated into an authoritative road network or not is on a head of authoritative data production. This is more technical question and a question related to a mapping agency production policy. Since it is not a research question, the validation of the final results of decision making step was not done within the scope of this thesis.

3.4 Conclusion

In this section we presented how a decision is made if a candidate for missing road is supposed to be proposed as an update for missing road or not. In that purpose, four relevant criteria are defined describing quantity, accuracy, and actuality of candidates for missing road and, continuity.

Further, we presented a measure for qualifying missing road regarding its relevance to be proposed as an update. The measure is a degree of confidence of missing road. It is calculated as a result of combining criteria by means of Equal Weight method. Finally, missing roads are classified into three categories of updates: updates with weak confidence, updates with medium confidence and updates with strong confidence.

Most numerous updates are those with medium degree of confidence (66%), followed by low confidence updates (19%) and strong confidence updated (15%). Medium and strong confidence updates compose more than 4/5 of total updates which is a satisfying result. Visual analyses of results allowed us to classify them into two categories: according to configuration update related to existing network and interest of introducing the update in the network.

Conclusion

Conclusion

In this thesis, we proposed a quality based approach for updating authoritative road network by means of GPS traces collected by volunteers during sport activities in mountain areas. Needs for very up-to-date authoritative road network is growing due to the increasing use of navigation technologies and Web 2.0 technology in everyday life. In this research work, we focus on three types of roads: footpath, tractor, and bicycle road. We propose a quality based approach for updating an authoritative dataset composed by three steps: evaluation of spatial data quality, data matching and decision making. The approach is generic owing to its dominant intrinsic component, and thus can be used for other kinds of traces and linear networks. Main accomplishments are presented in the following sections.

Evaluation of spatial data quality

Prior to use of VGI traces in updating authoritative road network, assessing their quality is a first important task. State of the art of spatial data quality presented in Chapter 1 shows that spatial data quality evaluation is a complex task, especially when it comes to recently introduces data sources such as VGI. In addition, a practical study of VGI traces completeness presented in Section 1.3 (Chapter 1) pointed out significant issues of VGI traces in terms of applying existing methods for spatial data quality evaluation. Three main issued are identified: lack of metadata, missing attributes and heterogeneous quality (huge variations in quality).

Thus, the proposed approach has to address spatial data quality evaluation of VGI traces taking into account listed issues as well as the specificities of GPS traces. The solution proposed is composed of three methods for the secondary human behaviour detection, outlier's detection, and low accuracy points detection. Due to our intention to use VGI traces in highlighting updates in authoritative road network, the last is not used in any of methods for evaluation of the traces quality, in order to avoid biased results. Therefore, the approach has a dominant intrinsic component. Regarding the external sources, only DTM and land cover maps are used and only in one of the methods, the method for detection of low accuracy points. In addition, machine learning techniques for learning from real examples are employed.

The approach for evaluation of VGI traces quality has two main contributions:

First, methods based on machine learning techniques to detect outliers and assess the spatial accuracy of points composing a GPS trace. Former is fully intrinsic, whereas latter has some extrinsic components. Both methods are designed to model and detect points with different characteristics from regular points of a VGI trace, such as outliers and low accuracy points by learning from the real examples. Especially important contributions of both methods are the point quality indicators, since the prior knowledge regarding outliers and low accuracy points detection in VGI traces was very poor. Within developing the indicators, fifteen internal metric and geometric characteristics of traces (e.g. direction change or mean distance between consecutive points) as well as external factors (e.g. canopy cover) were tested on their influences on traces quality. Five of them are proved to be those associated with outlier GPS points. The indicator AngleMean is found most dominant and efficient in the detection of outliers since it is

employed in all five generated rules by machine learning classification. Owing to proposed indicators, outliers are modelled and detected with a success of 79%, both precision and recall. Regarding extrinsic indicators recognised as influencing on points accuracy, CanopyCover was identified as most dominant appearing in three of seven rules for detection of low accuracy points. Additionally, the important contribution of the method for detection of low accuracy points is an estimation of GPS point's accuracy without comparing to referential data or using metadata containing GPS accuracy information such as PDOP, number of visible satellites, precision of GPS device used.

Thus the main advantages of those two methods are that they are not dependent on: Transportation mode of traces. The method can be applied on unclassified traces according to the transportation mode. It is frequently the first step of most of methods for outlier points and stops in human mobility analysis. Data redundancies. Traces are analysed independently, one by one, thus no redundancies in traces following the same path is required. This is important in the context of VGI traces collected in mountainous areas, where there are few redundancies. Missing attributes. The methods can be applied on traces having missing attributes, since detection of secondary human behaviour is a geometric method, and detection of outliers has 3 of 5 rules also geometric based, thus not dependent on attributes.

Second, a method based on intrinsic characteristics of a trace allowing to detect secondary human behaviour for improving the smoothness of GPS traces and thus, to obtain a GPS trace having a similar representation as for topographic paths. The method is very robust and can be applied even on traces having no or missing timestamps, compared to the most of the methods for stops detection that require time component. It has two advantages in common with methods for outliers and low accuracy points detection. The method is not dependent on transportation mode classification and missing attributes like similar methods presented in the Chapter 1. It is very simple and generic solution for detection of anomalies in human mobility behaviour that performed highly in our test zone reaching the precision of 98% and the recall of 93%.

Overall, we presented a novel approach for assessing GPS points accuracy and improving the smoothness of VGI traces that overcomes their high heterogeneity due to the fact that traces are coming from different and unknown contributors, are collected by different GPS devices during different activities on the one hand, and which is independent from the usage of referential data, on the other hand.

Data matching

Data matching is a necessary step for identifying the differences between a source data set (VGI traces) and a dataset to update (authoritative IGN road network), and thus identifying potential candidates for updates.

A review of literature presented in Section 2.2 showed that most of the existing matching methods are not adapted to handle VGI traces specificities. In addition, the lack of limited application of many of

existing approaches by disabling the use of some important criteria and similarity measures that are widely used in data matching, like attribute criterion or topological criteria.

For that purpose, we proposed a growing buffer based solution depending on points accuracy. In this way, searching for the matching candidates of low accuracy parts of VGI traces is differentiated from the process used for good accuracy parts, supposing that higher tolerance (greater buffer size) should be applied in case when selecting potential matching candidates of low quality parts of the traces.

The main strength of the approach is that it can be applied on traces with high variations of accuracy and precision, and that it is not dependent on attributes or topology. This is a convenient solution for VGI traces or other data collected without protocols. This simple, but very general approach, is well adapted in our context, where matching is only a preliminary step for searching clues of updates.

Decision making

The final step of the global approach is the decision making. It consists in evaluation a candidate for an update, in other words, to which extent we can trust to the proposed update. For that purpose, an indicator named ‘degree of confidence’ is proposed by combining four different criteria.

Proposed updates were classified according to the degree of confidence as: weak, medium and strong confidence updates. Owing to this classification a strategy regarding the integration of updates into an authoritative road network is proposed. Thus, updates having strong degree of confidence are eligible to be introduced directly into an authoritative dataset, whereas those having a medium degree of confidence need to pass a visual checking (verification) by the surveyors. Finally, updates with a weak degree of confidence are not supposed to be considered further, since their reliability is very unsatisfying. This is an important contribution of decision making step, since the handling of detected updates after they are detected is one of the essential questions of the approach. In this way, the time dedicated to the integration of updates is reduced by automatically dealing with strong and weak confidence updates.

Decision making is an important step because it allow to take into consideration other aspects such as authoritative data specification or spatial relations between a candidate to update and existing road network. However, the final decision if a detected update will be integrated into an authoritative road network is on the head of authoritative data production.

Overall, this thesis contributed to the analyses of digital trajectories collected by navigation technology and internet users. For successful analyses, handling and assessing data quality is very important. Besides presented contributions in this field, there is still a lot to be done, since this type of data is recently introduced and continuously evolving.

Perspectives and future work

In terms of data source used (VGI traces), further efforts regarding establishing protocols of traces collecting and sharing could be useful. High heterogeneity of traces, lack of metadata and missing attributes are found as very limiting factors for wider and more successful use of VGI traces in updating formal datasets. Therefore, an extra metadata could be useful such as: information about a type of GPS device used, PDOP (if applicable), and original elevation data. If available, they could be used as additional traces' quality indicators. Employed together with existing indicators, they would improve the performance of the approach for quality evaluation, since they are very relevant for GPS data quality evaluation according to presented state of the art of GPS data quality, especially PDOP claimed as most important smartphone GPS quality measure. Having more updates with a strong degree of confidence would subsequently reduce time needed for visual checking of updates.

Perspectives regarding the evaluation of VGI traces quality are as follows. New indicators for detection of outliers and low accuracy points can be proposed such as radar data can be as an indicator measuring the polarization HV (horizontal transmitting, vertical receiving). Our assumption is that knowing both radar data and GPS are using the same wavelength, there could be a correlation between HV and the accuracy of GPS. Detection of low accuracy points can be more flexible by allowing less sharp division of GPS points on good and low accurate. To do that, a probabilistic approach could be employed to assess the accuracy in a more detailed classification than good versus low accuracy classification, but in the scale from 0-1. Perspectives regarding data matching approach can be seen in two directions. First, more attributes available and better attribute completeness of VGI traces would allow employing wide range of attribute criteria and similarity measures, e.g. semantic similarity. This would make a data matching approach more robust and successful. Second, the improvement of evaluation of VGI traces quality, especially detection of low accuracy points would subsequently improve the performance of data matching approach proposed since it mostly relies on the result of detection of low accuracy points (buffer size selection).

Perspectives regarding the decision making are lying in introducing new criteria based on different data sources. Some new data sources could be potentially used in a decision making. First, Flickr data[8] (geo-tagged images) can be considered. Having been matched to missing roads, the images could be used either as discriminating criterion (to confirm or reject an update) or combined with the proposed decision making criteria. Second, areal images such as orthophoto can be used. Agnostic detection of roads is difficult, thus it may be possible to use our approach to focus on potential updates, then to use areal images to search for clues of the updates in the images. Here, image processing would be applied so that detected changes in the landscape are interpreted as the clues of missing roads. Finally, results of image processing (change detection) can be combined together with the results of the proposed decision making approach. Thirdly, textual descriptions of itineraries can be used and combined with the results of the presented approach so that extra information about the update is obtained. In that way, a confirmation of updates could be obtained by relaying on someone's testimony that he/she used an

[8] www.flickr.com

identified missing road in its itinerary. For that purpose, natural language processing tools could be employed.

Appendix

Appendix

Data sources

As presented in 1.3.1 VGI traces used in this work are obtained from French web-sites specialized for sharing GPS traces between hikers and mountain bikers. Four web-sites used as a data sources are as follows:

- RandoGPS (http://www.randogps.net/)
  RandoGPS is a French website dedicated for storing and presenting hiking (randonnée) GPS traces recorded by crowd. Data is organized and accessible according to administrative division of France on the level of departments. Traces are available only for French territory. In total 158 traces are coming from this source.

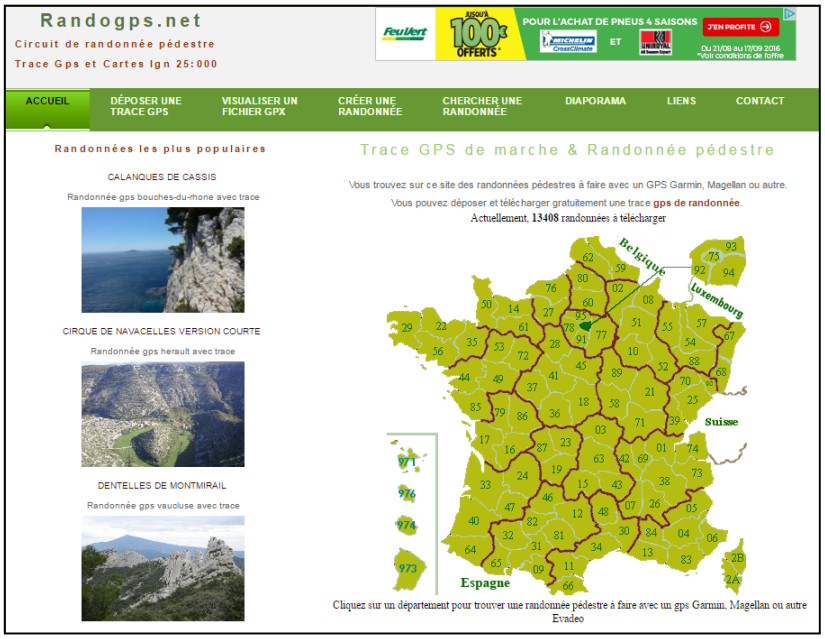


Figure 113 : Randogps website look

- TracesGPScom (http://www.tracegps.com/)
  TracesGPScom is a French website established for storing and presenting VGI GPS traces from sport activities such as: running, hiking, walking and cycling. Data is organized and accessible according to administrative division of France on the level of regions. Traces are available only for French territory. The majority of traces were obtained from this web-site – 163.

Figure 114 : TraceGPS website look

- VTTour (http://www.vttour.fr/)
  VTTour is a French website dedicated for storing and presenting VGI GPS traces collected by crowd while cycling. Data is presented according to time of upload. However, an advanced search according to departments is possible. Besides traces from French territory which are a significant majority, website contains traces from other countries like Spain, Italy, Canada, etc. In total 45 traces has this web-site as a source.

Figure 115 : VTTour website look

- VisuGPX (http://www.visugpx.com/)
  VisuGPX is a French website established for storing and presenting VGI GPS traces from obtained in: running, hiking, walking, cycling and moto-cycling. Data is presented according to time of upload. However, an advanced search according to postal codes i.e. cities. Apart from traces from French territory that are a significant majority, website contains traces from other countries like Spain, Italy, Canada, etc.
  The number of traces coming from this source is 60.

Figure 116 : Visu GPX website look

The traces are available in different formats from website to website, like: GPX, KML, CSV, GDB, GARMIN TCX, etc. An only format provided by all websites is GPX. That is why, we have decided to download and use traces stored as GPX files. For that purpose, a JAVA script for automatic downloads of traces from listed websites and storing them to postGIS database were created and applied. The script is also assigning the id of a trace to all points composing the trace, as well as a name of data source, i.e. name of website.

References

References

Abdalla, R., 2016. Introduction to geospatial information and communication technology (GeoICT).
URL http://www.worldcat.org/isbn/9783319336039

Al-Bakri, M., Fairbairn, D., 2010. Assessing the accuracy of 'crowdsourced' data and its integration with official data sets. In: Accuracy 2010 Symposium. pp. 317–320.

Al-Bakri, M., Fairbairn, D., Aug. 2012. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. International Journal of Geographical Information Science 26 (8), 1437–1456.
URL http://dx.doi.org/10.1080/13658816.2011.636012

Allen, A. M., Maansson, J., Jarnemo, A., Bunnefeld, N., 2014. The impacts of landscape structure on the winter movements and habitat selection of female red deer. European Journal of Wildlife Research 60 (3), 411–421.
URL http://dx.doi.org/10.1007/s10344-014-0797-0

Alvares, L. A., Fernandes, J. A., Macedo, D., Bogorny, V., Moelans, B., Kuijpers, B., Vaisman, A., 2007. A model for enriching trajectories with semantic geographical information. In: 15th Annual ACM International Symposium on Advances in GIS.

Amarintrarak, N., Runapongsa Saikeaw, K., Tongsima, S., Wiwatwattana, N., 2009. SAXM: Semi-automatic XML schema mapping. In: Proceedings of the 24th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). pp. 44–47.

Antoniou, V., Morley, J., Haklay, M., 2009. The role of user generated spatial content in mapping agencies. In: Proceedings of GISRUK conference.

Antoniou, V., Skopeliti, A., Aug. 2015. Measures and indicators of VGI quality: AN OVERVIEW. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W5, 345–351.
URL http://dx.doi.org/10.5194/isprsannals-ii-3-w5-345-2015

Arsanjani, J. J., Zipf, A., Mooney, P., Helbich, M., 2015. OpenStreetMap in GIScience : experiences, research, and applications.
URL http://www.worldcat.org/isbn/9783319142791

Ballatore, A., Zipf, A., 2015. A conceptual quality framework for volunteered geographic information. In: Fabrikant, S. I., Raubal, M., Bertolotto, M., Davies, C., Freundschuh, S., Bell, S. (Eds.), Spatial Information Theory. Vol. 9368 of Lecture Notes in Computer Science. Springer International Publishing, pp. 89–107.
URL http://dx.doi.org/10.1007/978-3-319-23374-1_5

Barron, C., Neis, P., Zipf, A., Dec. 2014. A comprehensive framework for intrinsic OpenStreetMap quality analysis. Transactions in GIS 18 (6), 877–895.
URL http://dx.doi.org/10.1111/tgis.12073

Batista, G. E. A. P. A., Prati, R. C., Monard, M. C., Jun. 2004. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6 (1), 20–29.
URL http://dx.doi.org/10.1145/1007730.1007735

Ben-Moshe, B., Elkin, E., Levi, H., Weissman, A., 2011. Improving accuracy of GNSS devices in urban canyons. In: CCCG.

Bergman, C., Oksanen, J., Dec. 2016. Conflation of OpenStreetMap and mobile sports tracking data for automatic bicycle routing. Transactions in GIS 20 (6), 848–868.
URL http://dx.doi.org/10.1111/tgis.12192

Bishr, M., Kuhn, W., 2007. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 365–387.
URL http://dx.doi.org/10.1007/978-3-540-72385-1_22

Blunck, H., Kjærgaard, M., Toftegaard, T., 2011. Sensing and classifying impairments of GPS reception on mobile devices. In: Lyons, K., Hightower, J., Huang, E. (Eds.), Pervasive Computing. Vol. 6696 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, Ch. 22, pp. 350–367.
URL http://dx.doi.org/10.1007/978-3-642-21726-5_22

Bordogna, G., Kliment, T., Frigerio, L., Brivio, P., Crema, A., Stroppiana, D., Boschetti, M., Sterlacchini, S., May 2016. A spatial data infrastructure integrating multisource heterogeneous geospatial data and time series: A study case in agriculture. ISPRS International Journal of Geo-Information 5 (5), 73+.
URL http://dx.doi.org/10.3390/ijgi5050073

Bouziani, M., Pouliot, J., Mar. 2008. Optimisation de la mise à jour des bases de données géospatiales proposition d'une procédure automatisée d'appariement géométrique d'objets linéaires. Revue internationale de géomatique 18 (1), 113–137.
URL http://dx.doi.org/10.3166/geo.18.113-137

Bruns, A., 2008. Blogs, Wikipedia, Second life, and Beyond : from production to produsage. Peter Lang.
URL http://www.worldcat.org/isbn/9780820488660

Buard, E., Feb. 2011. Pratiques spatiales des populations animales: analyses par les trajectoires. In: Actes de Dixièmes Rencontres de Théo Quant.

Buard, E., 2013. Dynamiques des interactions espèces espace, mise en relation des pratiques de déplacement des populations d'herbivores et de l'évolution de l'occupation du sol dans le parc de hwange (zimbabwe). Ph.D. thesis, Université Paris 1.

Cain, J. W., Krausman, P. R., Jansen, B. D., Morgart, J. R., Sep. 2005. Influence of topography and GPS fix interval on GPS collar performance. Wildlife Society Bulletin 33 (3), 926–934.
URL http://dx.doi.org/10.2193/0091-7648(2005)33\%5B926:iotagf\%5D2.0.co;2

Cannavo', F., Mattia, M., Rossi, M., Palano, M., Bruno, V., May 2010. A new algorithm for automatic Outlier Detection in GPS Time Series. In: EGU General Assembly Conference Abstracts. Vol. 12 of EGU General Assembly Conference Abstracts. p. 5027.

Chambers, E. W., Colin de Verdière, E., Erickson, J., Lazard, S., Lazarus, F., Thite, S., Apr. 2010. Homotopic fréchet distance between curves or, walking your dog in the woods in polynomial time. Computational Geometry 43 (3), 295–311.
URL http://dx.doi.org/10.1016/j.comgeo.2009.02.008

Chawla, N., 2005. Data mining for imbalanced datasets: An overview. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, pp. 853–867.
URL http://dx.doi.org/10.1007/0-387-25465-x_40

Chehreghan, A., Ali Abbaspour, R., May 2017. An assessment of spatial similarity degree between polylines on multi-scale, multi-source maps. Geocarto International 32 (5), 471–487.
URL http://dx.doi.org/10.1080/10106049.2016.1155659

Chen, H., Walter, V., 2009. Quality inspection and quality improvement of large spatial datasets. In: Proceedings of the GSDI 11 World Conference.

Chen, Y., Gong, J., Pan, J., Chen, X., Ke, Y., Jan. 2005. Conflation technology using in spatial data integration on the internet. pp. 56–63.
URL http://dx.doi.org/10.1117/12.572920

Chen, C., Shahabi, C., Knoblock, C. A., 2004. Utilizing road network data for automatic identification of road intersections from high resolution color orthoimagery. In: In Proceedings of the Second Workshop on Spatio-Temporal Database Management(STDBM'04), colocated with VLDB. pp. 1477–1480.

Ciepluch, B., Mooney, P., Winstanley, A. C., Apr. 2011. Building generic quality indicators for OpenStreetMap. In: 19th annual GIS Research UK (GISRUK).
URL http://eprints.maynoothuniversity.ie/2483/

Cohen, W. W., 1995. Fast effective rule induction. In: In Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, pp. 115–123.

Comber, A., Fisher, P., Wadsworth, R., Oct. 2004. Integrating land-cover data with different ontologies: identifying change from inconsistency. International Journal of Geographical Information Science 18 (7), 691–708.
URL http://dx.doi.org/10.1080/13658810410001705316

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G., Aug. 2013. Using control data to determine the reliability of volunteered geographic information about land cover. International Journal of Applied Earth Observation and Geoinformation 23, 37–48.
URL http://dx.doi.org/10.1016/j.jag.2012.11.002

Costes, B., Jan. 2015. Appariement hiérarchique de réseaux hydrographiques imparfaits. Revue Internationale de Géomatique 25 (1), 53–74.
URL http://dx.doi.org/10.3166/rig.25.53-74

Géodésie et Nivellement, 2010. Descriptifs quasi-geoides et grilles de conversion altimetrique sur la france metropolitaine. Tech. rep.

DeCesare, N. J., Squires, J. R., Kolbe, J. A., Sep. 2005. Effect of forest canopy on GPS-based movement data. Wildlife Society Bulletin 33 (3), 935–941.
URL http://dx.doi.org/10.2193/0091-7648(2005)33\%5B935:eofcog\%5D2.0.co;2

Dempster, A., 1967. Upper and lower probabilities induced by multivalued mapping. pp. 325–339.

D'Eon, R. G., Delparte, D., Mar. 2005. Effects of radio-collar position and orientation on GPS radio-collar performance, and the implications of PDOP in data screening. Journal of Applied Ecology 42 (2), 383–388.
URL http://dx.doi.org/10.1111/j.1365-2664.2005.01010.x

Devogele, T., 2002. A new merging process for data integration based on the discrete fréchet distance. In: Richardson, D., van Oosterom, P. (Eds.), Advances in Spatial Data Handling. Springer Berlin Heidelberg, pp. 167–181.
URL http://dx.doi.org/10.1007/978-3-642-56094-1_13

Devogele, T., Trevisan, J., Raynal, L., 1996. Building a multi-scale database with scale-transaction relationships. In: Spatial Data Handling (SDH). Taylor & Francis, pp. 337–351.

Diakoulaki, D., Mavrotas, G., Papayannakis, L., Aug. 1995. Determining objective weights in multiple criteria problems: The critic method. Computers & Operations Research 22 (7), 763–770.
URL http://dx.doi.org/10.1016/0305-0548(94)00059-h

Domingos, P., 1999. MetaCost: A general method for making classifiers Cost-Sensitive. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. Vol. 155-164.

Driemel, A., Har-Peled, S., Wenk, C., Feb. 2012. Approximating the fréchet distance for realistic curves in near linear time. Discrete & Computational Geometry 48 (1), 94–127.
URL http://dx.doi.org/10.1007/s00454-012-9402-z

Duran, A., Earleywine, M., Apr. 2012. GPS data filtration method for drive cycle analysis applications.
URL http://dx.doi.org/10.4271/2012-01-0743

Eliasson, M., 2014. A kalman filter approach to reduce position error for pedestrian applications in areas of bad GPS reception. Master's thesis, Universitet Umea.

Estellés-Arolas, E., González-Ladrón-de Guevara, F., Apr. 2012. Towards an integrated crowdsourcing definition. Journal of Information Science 38 (2), 189–200.
URL http://dx.doi.org/10.1177/0165551512437638

Ester, M., peter Kriegel, H., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231.

Etienne, L., 2011. Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction a la détection de comportements inhabituels. application au trafic maritime. Ph.D. thesis, Universite de Bretagne occidentale.

Etienne, L., Devogele, T., Buchin, M., McArdle, G., May 2016. Trajectory box plot: a new pattern to summarize movements. International Journal of Geographical Information Science 30 (5), 835–853.
URL http://dx.doi.org/10.1080/13658816.2015.1081205

Fan, H., Yang, B., Zipf, A., Rousell, A., Apr. 2016. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. International Journal of Geographical Information Science 30 (4), 748–764.
URL http://dx.doi.org/10.1080/13658816.2015.1100732

Fan, H., Zipf, A., Fu, Q., Neis, P., Apr. 2014. Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science 28 (4), 700–719.
URL http://dx.doi.org/10.1080/13658816.2013.867495

Fawcett, T., Provost, F., 1996. Combining data mining and machine learning from effective user profiling. AAAI.

Field, A., 2009. Discovering statistics using spss : (and sex and drugs and rock 'n' roll'). SAGE Publications.
URL http://www.worldcat.org/isbn/9781847879073

Field, A. P., 2000. Discovering statistics using SPSS for Windows : advanced techniques for the beginner. Sage Publications.
URL http://www.worldcat.org/isbn/0761957553

Flanagin, A., Metzger, M., Aug. 2008. The credibility of volunteered geographic information. GeoJournal 72 (3-4), 137–148.
URL http://dx.doi.org/10.1007/s10708-008-9188-y

Fonte, C. C., Bastin, L., See, L., Foody, G., Lupia, F., 2015. Usability of VGI for validation of land cover maps. International Journal of Geographical Information Science 29 (7), 1269–1291.
URL http://dx.doi.org/10.1080/13658816.2015.1018266

Frair, J. L., Fieberg, J., Hebblewhite, M., Cagnacci, F., DeCesare, N. J., Pedrotti, L., 2010. Resolving issues of imprecise and habitat-biased locations in ecological analyses using GPS telemetry data. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1550), 2187–2200.
URL http://rstb.royalsocietypublishing.org/content/365/1550/2187

Gabay, Y., Doytsher, Y., Oct. 1994. Automatic adjustment of line maps. In: Proceedings of the GIS/LIS Conference. pp. 333–341.

Galán, C. O. n., Rodríguez-Pérez, J. R., Torres, J. M., Nieto, P. J. G., 2011. Analysis of the influence of forest environments on the accuracy of GPS measurements by using genetic algorithms. Mathematical and Computer Modelling 54 (7–8), 1829–1834, mathematical models of addictive behaviour, medicine & engineering.
URL http://www.sciencedirect.com/science/article/pii/S0895717710005698

George, D., Mallery, P., 2003. SPSS for Windows step by step : a simple guide and reference, 11.0 update. Allyn and Bacon.
URL http://www.worldcat.org/isbn/0205375529

Girres, J.-F., Touya, G., Aug. 2010. Quality assessment of the french OpenStreetMap dataset. Transactions in GIS 14 (4), 435–459.
URL http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x

Gomez-Gil, J., Ruiz-Gonzalez, R., Alonso-Garcia, S., Gomez-Gil, F., Nov. 2013. A kalman filter implementation for precision improvement in Low-Cost GPS positioning of tractors. Sensors 13 (11), 15307–15323.
URL http://dx.doi.org/10.3390/s131115307

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221.
URL http://dx.doi.org/10.1007/s10708-007-9111-y

Goodchild, M. F., Hunter, G. J., Apr. 1997. A simple positional accuracy measure for linear features. International Journal of Geographical Information Science 11 (3), 299–306.
URL http://dx.doi.org/10.1080/136588197242419

Goodchild, M. F., Li, L., 2010. Automatically and accurately matching objects in geospatialdatasets. In: Proceedings of Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science. pp. 98–103.

Gravetter, F. J., Wallnau, L. B., 2014. Essentials of statistics for the behavioral sciences. Wadsworth Cengage Learning.
URL http://www.worldcat.org/isbn/9781133956570

Hagenauer, J., Helbich, M., Jun. 2012. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. International Journal of Geographical Information Science 26 (6), 963–982.
URL http://dx.doi.org/10.1080/13658816.2011.619501

Haklay, M., Aug. 2010a. How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Environment and Planning B: Planning and Design 37 (4), 682–703.
URL http://dx.doi.org/10.1068/b35097

Haklay, M., 2010b. How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Environment and Planning B 37: 682-703.
URL http://epb.sagepub.com/cgi/content/long/37/4/682

Hamming, R., 1950. Error detecting and error correcting codes, technical report 2. Bell System Technical Journal.

Hashemi, P., Ali Abbaspour, R., 2015. Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept. In: Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M. (Eds.), OpenStreetMap in GIScience. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, pp. 19–36.
URL http://dx.doi.org/10.1007/978-3-319-14280-7_2

Hausdorff, F., 1914. Grundzüge der Mengenlehre. Chelsea Pub. Co.
URL http://www.worldcat.org/isbn/9780828400619

Helbich, M., Amelunxen, C., Neis, P., 2012. Comparative spatial analysis of positional accuracy of OpenStreetMap and proprietary geodata. Angewandte Geoinformatik, 24–33.

Hong-Minh, T., Smith, D., Jul. 2007. Hierarchical approach for datatype matching in XML schemas. In: 24th British national conference on databases. Berlin/Heidelberg: Springer-Verlag, pp. 120–129.

I. G. N., 2015. BD topo version 2.1 descriptif de contenu. Tech. rep.

Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., Smoreda, Z., 2013. Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Vandenbroucke, D., Bucher, B., Crompvoets, J. (Eds.), Geographic Information Science at the Heart of Europe. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, pp. 247–265.
URL http://dx.doi.org/10.1007/978-3-319-00615-4_14

ISO 19157:2013, geographic information – data quality. Tech. rep.

Janeau, G., Adrados, C., Joachim, J., Gendner, J.-P., Pépin, D., 2004. Performance of differential GPS collars in temperate mountain forest. Comptes Rendus Biologies 327 (12), 1143–1149.
URL http://www.sciencedirect.com/science/article/pii/S1631069104001696

Japkowicz, N., 2000. Learning from imbalanced data sets: A comparison of various strategies. Tech. rep., AAAI.

Jiang, Z., Sugita, M., Kitahara, M., Takatsuki, S., Goto, T., Yoshida, Y., 2008. Effects of habitat feature, antenna position, movement, and fix interval on GPS radio collar performance in mount fuji, central japan. Ecological Research 23 (3), 581–588.
URL http://dx.doi.org/10.1007/s11284-007-0412-x

Kaplan, E. D., Hegarty, C., Nov. 2006. Understanding GPS : principles and applications, second edition. Hardcover.
URL http://www.worldcat.org/isbn/1580538940

Kessler, C., Trame, J., Kauppinen, T., Sep. 2011. Tracking editing processes in volunteered geographic information: The case of OpenStreetMap. In: proceedings of Workshop on Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data (IOPE 2011), Conference on Spatial Information Theory: COSIT'11. Belfast, Maine, USA.

Kieler, B., Huang, W., Haunert, J.-H., Jiang, J., 2009. Matching river datasets of different scales. In: Sester, M., Bernard, L., Paelke, V. (Eds.), Advances in GIScience. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 135–154.
URL http://dx.doi.org/10.1007/978-3-642-00318-9_7

Knight, N. L., Wang, J., Oct. 2009. A comparison of outlier detection procedures and robust estimation methods in GPS positioning. Journal of Navigation 62 (4), 699–709.
URL https://www.cambridge.org/core/article/a-comparison-of-outlier-detection-procedures-and-robust-estimation-methods-in-gps-positioning/8C40F30ED6ACF28B3B80FD119BF59D36

Koukoletsos, T., Haklay, M., Ellul, C., Aug. 2012. Assessing data completeness of VGI through an automated matching procedure for linear data. Transactions in GIS 16 (4), 477–498.
URL http://dx.doi.org/10.1111/j.1467-9671.2012.01304.x

Krumm, J., Davies, N., Narayanaswami, C., Oct. 2008. User-Generated content. IEEE Pervasive Computing 7 (4), 10–11.
URL http://dx.doi.org/10.1109/mprv.2008.85

Kubat, M., Holte, R. C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Machine Learning 30 (2), 195–215.
URL http://dx.doi.org/10.1023/A:1007452223027

Kumar, S., 2000. Data integration and accuracy issues in digital topographic databases. In: International Archives of Photogrammetry and Remote Sensing. Vol. 33.

Leibovici, D. G., Evans, B., Hodges, C., Wiemann, S., Meek, S., Rosser, J., Jackson, M., Aug. 2015. On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W5, 195–202.
URL http://dx.doi.org/10.5194/isprsannals-ii-3-w5-195-2015

Levenshtein, V., 1965. Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR, 845–848.

Lewis, D. D., Catlett, J., 1994. Heterogeneous uncertainty sampling for supervised learning. In: Cohen, W. W., Hirsh, H. (Eds.), Machine Learning Proceedings 1994. Morgan Kaufmann, San Francisco (CA), pp. 148–156.
URL http://www.sciencedirect.com/science/article/pii/B978155860335650026X

Lewis, J. S., Rachlow, J. L., Garton, E. O., Vierling, L. A., Mar. 2007. Effects of habitat on GPS collar performance: using data screening to reduce location error. Journal of Applied Ecology 44 (3), 663–671.
URL http://dx.doi.org/10.1111/j.1365-2664.2007.01286.x

Lin, D., Jul. 1997. An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning. pp. 296–304.

Liu, C., Xiong, L., Hu, X., Shan, J., Jul. 2015. A progressive buffering method for road map update using OpenStreetMap data. ISPRS International Journal of Geo-Information 4 (3), 1246–1264.
URL http://dx.doi.org/10.3390/ijgi4031246

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., 2009. Map-matching for low-sampling-rate GPS trajectories. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09. GIS '09. ACM Press, New York, NY, USA, pp. 352–361.
URL http://dx.doi.org/10.1145/1653771.1653820

Mascret, A., Devogele, T., 2006. Intégration de données Terre/Mer basée sur une approche paysagère - SAGEO 2006. In: Colloque International de Géomatique et d'Analyse Spatiale.

McKenzie, G., Janowicz, K., Adams, B., Mar. 2014. A weighted multi-attribute method for matching user-generated points of interest. Cartography and Geographic Information Science 41 (2), 125–137.
URL http://dx.doi.org/10.1080/15230406.2014.880327

McMaster, R., 1986. A statistical analysis of mathematical measures for linear simplification. The American Cartographer 23.

Milo, T., Zohar, S., Aug. 1998. Using schema matching to simplify heterogeneous datatranslation. In: 24th International Conference on Very Large Databases. pp. 122–123.

Mooney, P., Corcoran, P., Mar. 2012. Characteristics of heavily edited objects in OpenStreetMap. Future Internet 4 (1), 285–305.
URL http://dx.doi.org/10.3390/fi4010285

Mustière, S., Devogele, T., Dec. 2008. Matching networks with different levels of detail. GeoInformatica 12 (4), 435–453.
URL http://dx.doi.org/10.1007/s10707-007-0040-1

Neis, P., Zielstra, D., Zipf, A., Dec. 2011. The street network evolution of crowdsourced maps: OpenStreetMap in germany 2007–2011. Future Internet 4 (1), 1–21.
URL http://dx.doi.org/10.3390/fi4010001

Neis, P., Zielstra, D., Zipf, A., Jun. 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. Future Internet 5 (2), 282–300.
URL http://dx.doi.org/10.3390/fi5020282

Newson, P., Krumm, J., 2009. Hidden markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09. GIS '09. ACM Press, New York, NY, USA, pp. 336–343.
URL http://dx.doi.org/10.1145/1653771.1653818

Nutanong, S., Jacox, E. H., Samet, H., May 2011. An incremental hausdorff distance calculation algorithm. Proceedings of the VLDB Endowment 4 (8), 506–517.
URL http://dx.doi.org/10.14778/2002974.2002978

Olteanu, A. M., 2008. Fusion de connaissances imparfaites pour l'appariement de données géographiques : proposition d'une approche s'appuyant sur la théorie des fonctions de croyance. Ph.D. thesis, Université Paris-Est.

Olteanu-Raimond, A.-M., Couronné, T., Fen-Chong, J., Smoreda, Z., 2011. Modélisation des trajectoires spatio-temporelles issues des traces numériques de téléphones portables. le paris des visiteurs, qu'en disent les téléphones mobiles?

Olteanu Raimond, A.-M., Mustière, S., 2008. Data matching – a matter of belief. In: Ruas, A., Gold, C. (Eds.), Headway in Spatial Data Handling. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 501–519.
URL http://dx.doi.org/10.1007/978-3-540-68566-1_29

Olteanu-Raimond, A.-M., Mustière, S., Ruas, A., Jun. 2015. Knowledge formalization for vector data matching using belief theory. Journal of Spatial Information Science (10).
URL http://dx.doi.org/10.5311/josis.2015.10.194

Ordoñez, C., Martínez, J., Rodríguez-Pérez, J. R., Reyes, A., 2011. Detection of outliers in GPS measurements by using Functional-Data analysis. Journal of Surveying Engineering 137 (4), 150–155.
URL http://dx.doi.org/10.1061/(ASCE)SU.1943-5428.0000056

Palma, A. T., Bogorny, V., Kuijpers, B., Alvares, L. O., 2008. A clustering-based approach for discovering interesting places in trajectories. Annual ACM on Applied computing.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., Brunk, C., 1994. Reducing misclassification costs. In: Cohen, W. W., Hirsh, H. (Eds.), Machine Learning Proceedings 1994. Morgan Kaufmann, San Francisco (CA), pp. 217–225.
URL http://www.sciencedirect.com/science/article/pii/B9781558603356500349

Pendyala, R. M., 2002. Development of GIS-based conflation tools for data integration and matching. Tech. rep., Florida Department of Transportation.

Petitjean, F., Ketterlin, A., Gançarski, P., Mar. 2011. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition 44 (3), 678–693.
URL http://dx.doi.org/10.1016/j.patcog.2010.09.013

Ranacher, P., Brunauer, R., Trutschnig, W., der Spek, S. V., Reich, S., 2016. Why GPS makes distances bigger than they are. International Journal of Geographical Information Science 30 (2), 316–333.
URL http://dx.doi.org/10.1080/13658816.2015.1086924

Rocha, J. A. M. R., Oliveira, G., Alvares, L. O., Bogorny, V., 2010. Times V.C., DB-SMoT: A direction-based spatio-temporal clustering method. In: IEEE International Conference.

Rosen, B., Saalfeld, A., 1985. Match criteria for automatic alignment. In: Proceedings of 7th International Symposium on Computer-Assisted Cartography. pp. 456–462.

Roszkowska, E., 2013. Rank ordering criteria weighting methods - a comparative overview. Optimum. Studia Ekonomiczne (5), 14–33.

Royston, P., 1992. Approximating the Shapiro-Wilk w-test for non-normality. Statistics and Computing 2 (3), 117–119.
URL http://dx.doi.org/10.1007/BF01891203

Royston, P., 1995. Remark AS R94: A remark on algorithm AS 181: The W-test for normality. Journal of the Royal Statistical Society. Series C (Applied Statistics) 44 (4), 547–551.
URL http://dx.doi.org/10.2307/2986146

Ruiz, J. J., Ariza, F. J., Ureña, M. A., Blázquez, E. B., Sep. 2011. Digital map conflation: a review of the process and a proposal for classification. International Journal of Geographical Information Science 25 (9), 1439–1466.
URL http://dx.doi.org/10.1080/13658816.2010.519707

Ruiz-Lendínez, J. J., Ariza-López, F. J., Ureña Cámara, M. A., Jul. 2016. A point-based methodology for the automatic positional accuracy assessment of geospatial databases. Survey Review 48 (349), 269–277.
URL http://dx.doi.org/10.1179/1752270615y.0000000030

Saalfeld, A., Jan. 1988. Conflation automated map compilation. International Journal of Geographical Information Systems 2 (3), 217–228.
URL http://dx.doi.org/10.1080/02693798808927897

Safra, E., Kanza, Y., Sagiv, Y., Beeri, C., Doytsher, Y., Jan. 2010. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. International Journal of Geographical Information Science 24 (1), 69–106.
URL http://dx.doi.org/10.1080/13658810802275560

Safra, E., Kanza, Y., Sagiv, Y., Doytsher, Y., 2006. Efficient integration of road maps. In: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems - GIS '06. ACM Press, pp. 59+.
URL http://dx.doi.org/10.1145/1183471.1183483

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., Rutzinger, M., Apr. 2016a. Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. ISPRS International Journal of Geo-Information 5 (5), 55+.
URL http://dx.doi.org/10.3390/ijgi5050055

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., Rutzinger, M., 2016b. Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. ISPRS International Journal of Geo-Information 5 (5), 55.
URL http://www.mdpi.com/2220-9964/5/5/55

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., Haklay, M. M., 0. A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science 0 (0), 1–29.
URL http://dx.doi.org/10.1080/13658816.2016.1189556

Shafer, G., 1976. A mathematical theory of evidence. Princeton University Press.

Shapiro, S. S., Wilk, M. B., Dec. 1965. An analysis of variance test for normality (complete samples). Biometrika 52 (3-4), 591–611.
URL http://dx.doi.org/10.1093/biomet/52.3-4.591

Shi, W., 2012. Advances in geo-spatial information science. CRC Press.
URL http://www.worldcat.org/isbn/9780203125786

Smart, P. D., Quinn, J. A., Jones, C. B., Mar. 2011. City model enrichment. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 223–234.
URL http://dx.doi.org/10.1016/j.isprsjprs.2010.12.004

Song, W., Haithcoat, T. L., Keller, J. M., Jan. 2006. A snake-based approach for TIGER road data conflation. Cartography and Geographic Information Science 33 (4), 287–298.
URL http://dx.doi.org/10.1559/152304006779500669

Song, W., Keller, J. M., Haithcoat, T. L., Davis, C. H., Feb. 2011. Relaxation-Based point feature matching for vector map conflation. Transactions in GIS 15 (1), 43–60.
URL http://dx.doi.org/10.1111/j.1467-9671.2010.01243.x

Sui, H., Li, D., Gong, J., 2004. Automatic feature-level change detection (FLCD)for road network. In: Proceedings of the 20th ISPRS Congress.

Tao, C., Yuan, S., 1999. Development of conflation components. In: Proceedings of Geoinformatics. pp. 1–13.

Thangaraj, M., And, C. R. V., Mar. 2013. Article: Performance study on rule-based classification techniques across multiple database relations. International Journal of Applied Information Systems 5 (4), 1–7, published by Foundation of Computer Science, New York, USA.

Thierry, B., Chaix, B., Kestens, Y., Mar. 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. International Journal of Health Geographics 12 (1), 14+.
URL http://dx.doi.org/10.1186/1476-072x-12-14

Toloie-Eshlaghy, A., Homayonfar, M., Aghaziarati, M., Arbabiun, P., 2011. A subjective weighting method based on group decision making for ranking and measuring criteria values. Australian Journal of Basic and Applied Sciences (5(12)).

Tong, X., Liang, D., Jin, Y., Apr. 2014. A linear road object matching method for conflation based on optimization and logistic regression. International Journal of Geographical Information Science 28 (4), 824–846.
URL http://dx.doi.org/10.1080/13658816.2013.876501

Tong, X., Shi, W., Deng, S., Sep. 2009. A probability-based multi-measure feature matching method in map conflation. International Journal of Remote Sensing 30 (20), 5453–5472.
URL http://dx.doi.org/10.1080/01431160903130986

Touya, G., Brando-Escobar, C., Jun. 2013. Detecting Level-of-Detail inconsistencies in volunteered geographic information data sets. Cartographica: The International Journal for Geographic Information and Geovisualization 48 (2), 134–143.
URL http://dx.doi.org/10.3138/carto.48.2.1836

Trochim, W., Donnelly, J. P., 2006. The Research Methods Knowledge Base. Cengage Learning.
URL https://books.google.fr/books?id=097mAAAACAAJ

Turner, A., 2006. Introduction to neogeography.
URL http://www.worldcat.org/isbn/0596529953

Uitermark, H., Vogels, A., van Oosterom, P., 1999. Semantic and Geometric Aspects of Integrating Road Networks. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 177–188.
URL https://doi.org/10.1007/10703121_15

van Exel, M. Dias, E. Fruijtier, S., 2010. The Impact of Crowdsourcing on Spatial Data Quality Indicators. In Proceedings of GiScience 2011 (September 2010).

van Wijngaarden, F., van Putten, J., van Oosterom, P., Uitermark, H., 1997. Map integration—update propagation in a multi-source environment. In: Proceedings of the fifth ACM international workshop on Advances in geographic information systems - GIS '97. ACM Press, pp. 71–76.
URL http://dx.doi.org/10.1145/267825.267844

van Winden, K., Biljecki, F., van der Spek, S., Oct. 2016. Automatic update of road attributes by mining GPS tracks. Transactions in GIS 20 (5), 664–683.
URL http://dx.doi.org/10.1111/tgis.12186

Volz, S., 2005. Data-Driven matching of geospatial schemas. In: Cohn, A., Mark, D. (Eds.), Spatial Information Theory. Vol. 3693 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 115–132.
URL http://dx.doi.org/10.1007/11556114_8

Walter, V., 1997. Zuordnung von raumbezogenen daten - am beispiel der datenmodelle ATKIS undGDF. Ph.D. thesis, Universitat Stuttgart.

Walter, V., Fritsch, D., Jul. 1999. Matching spatial data sets: a statistical approach. International Journal of Geographical Information Science 13 (5), 445–473.
URL http://dx.doi.org/10.1080/136588199241157

Wenk, C., Salas, R., Pfoser, D., 2006. Addressing the need for Map-Matching speed: Localizing globalb Curve-Matching algorithms. In: 18th International Conference on Scientific and Statistical Database Management (SSDBM'06). IEEE, pp. 379–388.
URL http://dx.doi.org/10.1109/ssdbm.2006.11

Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: Proceedings of the 32nd Meeting of Association of Computational Linguistics. pp. 33–138.

Xavier, E. M. A., Ariza-López, F. J., Ureña Cámara, M. A., Aug. 2016. A survey of measures and methods for matching geospatial vector datasets. ACM Computing Surveys 49 (2), 1–34.
URL http://dx.doi.org/10.1145/2963147

Xie, P., Petovello, M. G., Feb. 2015. Measuring GNSS multipath distributions in urban canyon environments. IEEE Transactions on Instrumentation and Measurement 64 (2), 366–377.
URL http://dx.doi.org/10.1109/tim.2014.2342452

Yan, Y., Feng, C.-C., Wang, Y.-C., 2016. Utilizing fuzzy set theory to assure the quality of volunteered geographic information. GeoJournal, 1–16.
URL http://dx.doi.org/10.1007/s10708-016-9699-x

Yang, B., Zhang, Y., Luan, X., Feb. 2013. A probabilistic relaxation approach for matching road networks. International Journal of Geographical Information Science 27 (2), 319–338.
URL http://dx.doi.org/10.1080/13658816.2012.683486

Ying, S., Li, L., Gao, Y. R., Min, Y., 2011. Probabilistic matching of map objects in multi-scale space. In: Proceedings of the 25th International Cartographic Conference.

Yuan, S., Tao, C., 1999. Development of conflation components. In: Proceedings of Geoinformatics. pp. 1–13.

Zandbergen, P. A., Barbeau, S. J., Jun. 2011. Positional accuracy of assisted GPS data from High-Sensitivity GPS-enabled mobile phones. Journal of Navigation 64 (03), 381–399.
URL http://dx.doi.org/10.1017/s0373463311000051

Zhang, M., Meng, L., Sep. 2007. An iterative road-matching approach for the integration of postal data. Computers, Environment and Urban Systems 31 (5), 597–615.
URL http://dx.doi.org/10.1016/j.compenvurbsys.2007.08.008

Zhang, M., Wei, S., Meng, L., Jul. 2005. A generic matching algorithm for line networks of different resolutions. In: 8th ICA WORKSHOP on Generalisation and Multiple Representation.

Zhang, X., Ai, T., Stoter, J., Zhao, X., Jun. 2014. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. ISPRS Journal of Photogrammetry and Remote Sensing 92, 147–163.
URL http://dx.doi.org/10.1016/j.isprsjprs.2014.03.010

Zhang, X., Zhao, X., Molenaar, M., Stoter, J., Kraak, M. J., Ai, T., 2012. Pattern classification approaches to matching building polygons. In: XXII ISPRS Congress. pp. 19–24.

Zhixian, Y., Parent, C., Spaccapietra, S., Chakraborty, D., 2010. A hybrid model and computing platform for Spatio-Semantic trajectories. In: 7th Extended Semantic Web Conference.

Zielstra, D., Hochmair, H., Dec. 2011. Comparative study of pedestrian accessibility to transit stations using free and proprietary network data. Transportation Research Record: Journal of the Transportation Research Board 2217, 145–152.
URL http://dx.doi.org/10.3141/2217-18

Zimmermann, M., Kirste, T., Spiliopoulou, M., 2009. Finding stops in Error-Prone trajectories of moving objects with Time-Based clustering. In: Tavangarian, D., Kirste, T., Timmermann, D., Lucke, U., Versick, D. (Eds.), Intelligent Interactive Assistance and Mobile Multimedia Computing. Vol. 53 of Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 275–286.
URL http://dx.doi.org/10.1007/978-3-642-10263-9_24