# A REGRESSION MODEL OF SPATIAL ACCURACY PREDICTION FOR OPENSTREETMAP BUILDINGS

I. Maidaneh Abdi [1,2,*], A. Le Guilcher[1], A-M. Olteanu-Raimond[1]

[1]Univ.Paris-Est, LASTIG MEIG, IGN, ENSG, F-94160 Saint-Mandé, France -
(abdi.maidaneh, arnaud.le-guilcher, ana-maria.raimond)@ign.fr

[2]ITU-I, Djibouti University , Djibouti - ibrahim_maidaneh_abdi@univ.edu.dj

**Commission IV, WG IV/4**

**KEY WORDS:** Regression, Intrinsic and Extrinsic Quality, Spatial Accuracy, OpenStreetMap, Reference Data, Multi-criteria Data Matching, Spatial Data, Belief Theory.

## ABSTRACT:

Data quality assessment of OpenStreetMap (OSM) data can be carried out by comparing them with a reference spatial data (e.g authoritative data). However, in case of a lack of reference data, the spatial accuracy is unknown. The aim of this work is therefore to propose a framework to infer relative spatial accuracy of OSM data by using machine learning methods. Our approach is based on the hypothesis that there is a relationship between extrinsic and intrinsic quality measures. Thus, starting from a multi-criteria data matching, the process seeks to establish a statistical relationship between measures of extrinsic quality of OSM (i.e. obtained by comparison with reference spatial data) and the measures of intrinsic quality of OSM (i.e. OSM features themselves) in order to estimate extrinsic quality on an unevaluated OSM dataset. The approach was applied on OSM buildings. On our dataset, the resulting regression model predicts the values on the extrinsic quality indicators with 30% less variance than an uninformed predictor.

## 1. INTRODUCTION

Spatial Data quality is necessary for researchers and practitioner in Geographic Information Science (GIS) (Devillers et al., 2007). The issues of quality impact all fields using geographic information such as safety operation, data integration. Relying on poor quality can be misleading for decision making process (e.g implantation of new commercial center or entail positional error for building a dam). Thus, the assessment of spatial accuracy becomes crucial and it is part of the quality concept covering the entire process from acquisition to diffusion of geographic information (Devillers et al., 2007).

Originally, the spatial quality was described as the conformity of a product with some standards of spatial data quality using a threshold acceptability. Thus, the accuracy of data refers to the degree of closeness between the measurement of the quantity and the accepted true value of that quantity. Spatial Data quality assessment is the process of comparing data to their accepted true values, according to fixed specifications. International Organization for Standardization (ISO) metadata, defined the following quality measures: completeness, consistency, positional accuracy, temporal accuracy, and thematic accuracy [1]. When information on the specifications is missing [2], a so called reference database is used to assess the quality of a dataset. In all cases, spatial data quality is assessed by comparison and requires both an external database and specifications. This type of spatial quality evaluation is named extrinsic quality and uses a methodology which involves all parameters mentioned by the ISO as measures of extrinsic quality data.

With the advent of crowdsourcing practice defined in Goodchild (2007), researchers have started to explore how these resources from Volunteered Geographic Information (VGI) could be enhanced to fit different scientific and societal needs. One of the most successful VGI projects is OpenStreetMap (Neis et al., 2013). A lot of research effort was put to evaluate the quality of OSM data. For example, OSM data quality is evaluated by comparing them to authoritative data (Girres and Touya, 2010), (Haklay, 2010). The work of Siebritz (2014) advised to evaluate the conformity of the OSM data with the existing topographic data of the Mapping National Agency (South Africa) through a threshold of acceptability in order to retain only those areas whose OSM data respect this threshold. These areas mark new changes in the city and guide the planning of areas for the collection of topographic data. There has been other essential works on the evaluation of OSM data through the parameter of positional accuracy, such as that of Goodchild and Hunter (1997) and Girres and Touya (2010). For the evaluation of the semantic accuracy, we note the work provided by Haklay (2010), while among those who have studied the evaluation of completeness, we refer to the work of several authors (Ather, 2009; Kounadi, 2009; Brando and Bucher, 2010; Fan et al., 2014).

ISO standards mention some intrinsic indicators to express general characteristics of the data such as purpose, usage, and lineage. The purpose describes the intended usage of the dataset while usage indicates in which application the dataset has been utilized. The lineage describes the history of a dataset compilation, acquisition and derivation to its form at the time of use (Van Oort and Bregt, 2005). These intrinsic indicators are used to assess the authoritative data.

However, the intrinsic indicators defined by the ISO cannot be applied to OSM data because their mapping process is different

[*] Corresponding author
[1] http://www.isotc211.org/
[2] https://www.infostore.saiglobal.com/preview/is/en/2013/i.s.eniso19157-2013.pdf?sku=1700851/

https://doi.org/10.5194/isprs-annals-V-4-2020-39-2020 | © Authors 2020. CC BY 4.0 License.

from that of the authoritative data for these causes: lack of specifications, use of different tools for data capture and variety of contributor's profiles. Some intrinsic proxies, such as indicators based on on contributor expertise and historic data, can help the evaluation of data quality. In this case, there is a type of evaluation quality aiming to define what could be called intrinsic quality (Antoniou and Skopeliti, 2015).

In the case of VGI, quality indicators may concern OSM contributors, such as trust, reputation, experience, credibility, local knowledge and reliability (Flanagin and Metzger, 2008), (Van Exel et al., 2010). For example, reputation is assessed by the history of past interactions between contributors (Senaratne et al., 2017; Barron et al., 2014). Furthermore, knowing that the VGI are captured through a participatory processes (Goodchild, 2007), research works explore the history of edits to provide some way to evaluate data quality based on the history of contribution (Barron et al., 2014). Based on OSM data history, Truong et al. (2019) proposed to classify contributors as pioneers, moderators and truthful contributors by analyzing the interactions between them, over time. In the work of Hashemi and Abbaspour (2015), the authors employed a concept of similarity in misrepresentations to detect topological incoherence based on contributions history. Others research tried to assess data quality based only on geometrical content in order to measure intrinsic qualities like topological consistency by detecting thematic incoherence (using spatial context) or by analysing geometrical consistency regarding the resolution of data (Touya and Brando-Escobar, 2013).

Despite the above research, there are still concerns about issues raised by the OSM data quality assessment. Senaratne et al. (2017) indicate that these contributor indicators are most often subjective and difficult to formalize. Morever, we add that it is not easy to access historical data, given the variability of data from one database to another. Besides, work based solely on geometrical content is not yet sufficient because it detects errors locally and it is hardly possible to provide a continuous and quantitative measure of positional or geometrical accuracy.

In a context of lack of the reference data, Our overall objective is to infer the quality of OSM data from intrinsic indicators.

To do this, our contribution is based on two research axis: the matching of OSM data and the establishment of a multiple regression model. First, data matching is the cornerstone of this work because our approach is intended to be precise and robust in terms of matching. Our knowledge model is based on The Belief Theory. As defined in Olteanu-Raimond et al. (2015), this method allows to take into account cases where knowledge could be missing, uncertain or imprecise. This means that the possibility of abstaining from matching is admitted. The data matching based on Belief Theory, allows to merge knowledge from several matching criteria to decide to choose the best homologous feature supported by all the matching criteria or to go for a solution of undecided.

In this paper, we extend the method proposed in Olteanu-Raimond et al. (2015) by modeling and defining criteria able to measure the similarities between polygons. For matching polygons, we propose to use geometric and position criteria. For the geometric criteria, we consider the *angular distance* and the *radial distance*, while for the position criteria, we have chosen the *surface distance* and the *Hausdorff distance*. These distances are computed between a feature from a reference database and a group of candidate features from the OpenStreetMap database. At this stage of data matching, the added value of this work compared to previous work such as in Olteanu-Raimond et al. (2015), appears in two levels. Firstly, matching based on belief theory has been tested for the first time on surface shaped data in this work. Secondly, the definition of matching criteria and their thresholds for establishing belief functions is also a novelty in the work on Belief Theory matching.

Second one can conceive a method providing a prediction for the extrinsic quality of data using machine learning. One approach is based on in depth learning methods that use several layers to select certain descriptors in order to progressively reconstruct the output of the starting layer as similar as possible. For example, Xu et al. (2017) used an auto-encoder network to reconstruct the best form of a variable through anomaly detection on that variable. This method uses on the one hand OSM and reference data, and on the other hand an input image to detect the footprint of buildings. However our work aims at providing an intrinsic estimator for the extrinsic quality by using only the current version of OSM and reference data. The approach consists in finding a regression linking extrinsic quality indicators with intrinsic indicators. Our research assumption is that intrinsic indicators characterizing geometric features allow the estimation of extrinsic indicators.

Ultimately, in order to reach the objectives mentioned above, three different aspects are studied:

- define the extrinsic indicators by matching VGI data with authoritative data;
- identify and define appropriate intrinsic indicators;
- define a robust method to estimate extrinsic indicators from intrinsic data quality indicators.

Two statistical methods are tested: multiple standard regression and lasso regression. The paper is structured as follows: section 2 describes the proposed methodology. Section 3 details the results while providing an analysis of these results and suggesting potential improvements.

## 2. METHODOLOGY

### 2.1 Approach

Our approach is composed by four steps (see figure 1). The first step consists in finding out a set of indicators which describe the geometrical and positional accuracy from the scientific literature. After that, we implement a computation process to define four distances between two features through each measure. Theses distances are considered as *extrinsic indicators*.

In a second step, we define a *multi-criteria* matching algorithm which is able to define homologous features between OSM and reference data.

Once the homologous features are identified, the third step consists in defining intrinsic indicators based on properties which reflect the quality of an feature. These are the indicators that are supposed to correlate with the extrinsic indicators (extrinsic quality).

The final step consists of applying statistical techniques to identify statistical correlations between extrinsic indicators and intrinsic indicators in order to predict a relative geometrical (e.g shape) and positional accuracy for each feature.

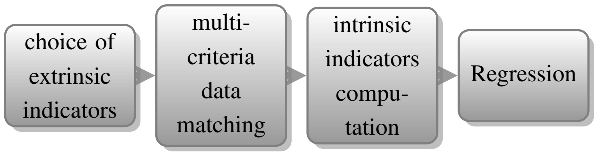
choice of extrinsic indicators
multi-criteria data matching
intrinsic indicators computation
Regression

Figure 1: Steps of our proposed approach (left to right).

## 2.2 Choice of extrinsic indicators

With a view of selecting extrinsic indicators, it is specified that our study focuses on buildings represented by surface features. A surface feature is the flat surface formed by connecting the points constituting the contour of a polygon.

According to Bel Hadj Ali (2001), quality control of a geographical feature should be based not only on the deviation of its position from its counterpart in the ground truth or the reference dataset, but also on the deviation of its shape. Thus, in order to be able to reflect the spatial quality of an OSM dataset with respect to a reference database, we must establish measurement indicators relating to the position and shape of a surface feature. Therefore, we propose the surface distance and the Hausdorff distance for measuring the deviation of position, and the angular and the polygonal (or radial) distances for measuring the shape deviations.

The surface distance is the ratio of the area of the symmetric difference of the two features and the area of their union. If the two features are completely disjointed, their surface distance is thus 1. If there is a perfect equality between the two features, it is equal to 0.

The Hausdorff distance is between the two full surfacic entities and not only the polylines (Bel Hadj Ali, 2001).

In the work of Bel Hadj Ali (2001), the angular function is defined on any vertex of the polygon, as being the angle between the tangent on this vertex and the curvilinear abscissa normalized by the perimeter of the polygon. The angular function is expressed in radian on curvilinear abscissa values between 0 and 1, which gives the appearance of a piecewise continuous function on the edges. An angular function is invariant by translation, rotation and homothethy but depends on the point of origin. An angular distance between two polygons is the integral of the differences of their two angular functions.

The radial distance between two polygons is the integral of the differences of their two polygonal functions or signatures. The latter is defined as a function measuring for any vertex of the polygon, the Euclidean distance from that vertex to the center of mass of the polygon for a curvilinear abscissa value normalized to the perimeter of the polygon. A radial function is invariant by translation and rotation but depends on the point of origin (Bel Hadj Ali, 2001).

To overcome the problem of the origin point (i.e. the phase shift between two radial/angular functions for the same polygon), the shift that minimizes angular (or radial) distance is chosen. The resulting angular (respectively radial) distance is invariant (geometrically) to a similarity (respectively to a rigid transformation).

For the calculation of the radial signature, significant errors can be made if the calculation is made only at the vertices of a polygon, implying that it is necessary to over-sample in order to obtain an accurate representation. Ali and Vauglin (2000) proposes to sample at least twice the number of vertices of the polygon. Thus, we have sampled all OSM and reference polygons using 100 points to give the same resolution for all polygons and a high fidelity of representation of signatures. The examples below (see figures 2 and 3) illustrate the computation distances (angular and radial) distances.

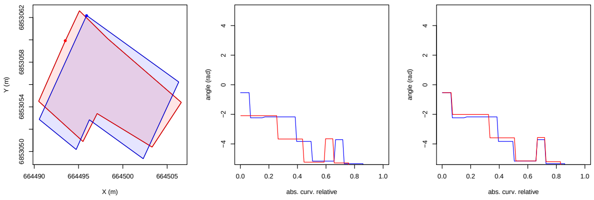

Figure 2: Angular distance computation: input buildings (left), raw signatures (center) and registered signatures (right). On this example, the optimal rotation is $\Delta\theta = 25°$ and $d_a = 27°$.

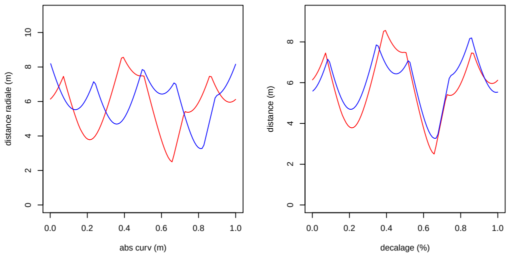

Figure 3: Radial distance computation: raw signatures (left) and registered signatures (right). On this example, the optimal shift is 15 % and $d_r = 0.78$ m.

## 2.3 Multi-criteria data matching

For each feature from the reference dataset we are looking from candidates into OSM dataset. Then, each couple (featureRef, candidateOSM) is compared by using different criteria. The multi-criteria data matching approach consists in defining different matching criteria. For each matching criterion, three belief functions are defined for each of the three hypotheses, namely the hypothesis *candidate is the homologous feature* ($appCi$), the hypothesis *candidate is not the homologous feature* ($-appCi$) and the ignorance hypothesis *I don't know if the candidate is the homologous feature* ($\Theta$). This expresses well the belief that we grant a candidate through these hypotheses and is materialized by what is called a *mass of belief.*

The definition of belief functions is used to configure the data matching algorithm in such a way as to make it converge towards the most plausible or credible decision by minimizing the conflicts that occur when two criteria support two distinct candidates in the same way. Thus, in order to define thresholds for the establishment of belief function, we had to conduct an empirical study by observing the distribution of values of the matching criteria.

At first sight, we observe a great similarity of the values on the geometric criteria of the candidate features. For example, for small values on angular (respectively radial) distance, several candidates can be supported by the hypothesis that each of the candidates is the homologous feature. This would lead to the appearance of strong conflicts. To reduce the conflict, the result of the geometric criteria must be corroborated with the values taken from the position indicators. This leads us to assign a rather strong mass of belief to the ignorance hypothesis of doubt about these geometric criteria and to leave to the position criteria the choice of deciding among the candidates, the one that seems the most probable.

Furthermore, when the values on the angular distance (respectively radial distance) are high enough, we doubt that the candidate is the homologous feature. This still favors the ignorance hypothesis and weakly the hypothesis of non-matching on the geometrical criteria.

This phase of modeling the knowledge and defining the thresholds is named *the initialization of belief masses*. For each indicator, we define a matching criterion. Figure 4 illustrates the belief functions for the matching criteria based on angular (top) and Hausdorff (bottom) distances. Note that the matching criterion based on the radial distance (respectively surface distance ) follows similar belief functions as angular distance (respectively Hausdorff distance) except the thresholds which are slightly different (see Table 1). The sum of the belief masses bust be equal to 1. The following table summarizes the set of thresholds involved in the equations of the belief functions.

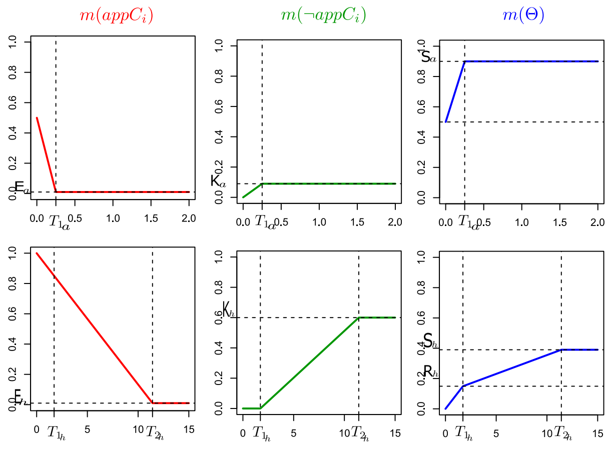

Figure 4: Belief function: masses of belief for each hypothesis ($appCi$, $\neg appCi$, and $\Theta$), for angular distance $da$ (top) and Hausdorff distance $dh$ (bottom) by their thresholds ($T_1$ and $T_2$) and their parameter values ($E_a$, $K_a$,$S_a$ for $da$, and $E_h$ and $K_h$,$S_h$,$R_h$ for $dh$)

| settings | $T_1$ | $T_2$ | $E$ | $K$ | $S$ |
|---|---|---|---|---|---|
| da | 0.25 | $--$ | 0.01 | $0.1 - T_1$ | 0.9 |
| dr | 0.7 | $--$ | 0.01 | $0.1 - T_1$ | 0.9 |
| dh | 1.72 | 11.42 | 0.01 | $1 - E - S$ | 0.6 |
| ds | 0.5 | 0.6 | 0.01 | $1 - E - S$ | 0.6 |

Table 1: Settings of belief functions of angular distance($da$), radial distance ($dr$), haussdorf distance ($dh$) and surface distance ($ds$) as the criteria of the data matching

## 2.4 Intrinsic indicators

Once the data matching has been carried out, we proceed to define intrinsic indicators that could be correlated to geometric and positional accuracy. The objective is to define intrinsic indicators for polygons that can succeed in indicating the quality of the capturing of geographical features. This requires analyzing the source of error in the data, then formalizing intrinsic indicators as precursors of a possible geometric and positional error.

Thus, Batton-Hubert et al. (2019) identified three most common sources of imperfection in VGI: measuring instruments, a lack of experience and knowledge of the data, or an act of vandalism. These imperfections affect positional and geometric accuracy. Without explicitly characterizing these imperfections, we take them into account to define following intrinsic indicators:

- *rectangular* ($rec$): this indicator measures the proximity of a polygon to its smallest surrounding rectangle (SSR), and it is computed as the ratio of the area of the polygon to the area of the SSR. The indicator takes the value 1 when the area of the polygon is perfectly equal to that of the SSR and tends towards 0 when it is too small to that of the SSR. This tells us about a rectangular shape of a building.
- *mean-lengths* ($lme$): this indicator is defined as the average length of sides of building.
- *max-lengths* ($lmx$): this indicator measures how many times the length of the longest segment of the polygon is longer than the average length of the segments of the polygon. It is calculated as the ratio of the length of the longest segment of the polygon to the average length of the segments of the polygon. It takes values are greater than or equal to 1. This may indicate a possible input error on a given segment.
- *min-lengths* ($lmn$): this indicator measures how many times the shortest segment of the polygon is shorter than the average length of the segments of the polygon. It is calculated as the ratio of the length of the shortest segment of the polygon to the average length of the segments of the polygon. It takes values less than or equal to 1.
- *outlier* ($out$): this indicator measures the degree of distance of a summit from the others to consider it as an aberrant point. For each vertex, an average distance is calculated which separates it from the other vertices. The vertex with the greatest average distance is selected. This value will be divided by the average length of the segments of the polygon. The value is greater than 1, the more it is suspected that the vertex in question is an outlier.
- *compacity* ($cpc$): this indicator measures the compactness of the polygon according to the principle that the circle is the figure whose area is maximal for a given perimeter. It is the ratio between the area of the polygon under study and the area of the circle having the same perimeter as the polygon. The values range from 0 to 1, so that a value close to 0 (low compactness) reflects an extended shape, and a value close to 1 (high compactness) reflects a compact or circular shape.
- *convexity* ($cvx$): this indicator measures the degree to which the shape of the polygon resembles a cubic shape rather than a hollow or bumpy shape. It is the quotient of the surface area of the polygon by its convex envelope. The values range from 0 (slightly convex) to 1 (perfectly convex).
- *elongation* ($elg$): this indicator measures the degree to which the shape of the polygon resembles a square. It is the ratio of width to length of the smallest surrounding rectangle (SSR). It tends towards 1 when the shape tends to be square.

https://doi.org/10.5194/isprs-annals-V-4-2020-39-2020 | © Authors 2020. CC BY 4.0 License.

- *q-reconstruct* ($qrc$): this indicator measures the quality of the reconstruction of the polygon using some of the vertices of the polygon. A reconstruction threshold of 80% of the polygon shape is set and the proportion of the number of vertices that reconstruct (for this threshold) to the total number of vertices of the polygon is calculated. A proportion close to 1 implies a large number of vertices and reflects a redundant shape that may express a poor quality of polygon capture.
- *right-angle* ($ragl$): this indicator measures the number of approximately right angles and still tells us about the regular shape of a polygon.
- *perimeter* ($per$): it is defined as the sum of the lengths of the sides of the polygon.
- *area* ($are$): it is defined as the area of the polygon
- *orientation* ($ori$): this indicator measures the overall orientation of the SSR of the polygon.
- *granularity* ($grn$): this indicator measures the granularity of a polygon. It is computed as the quotient of the number of vertices on the perimeter.

Subsequently, these indicators are calculated for each building from the OSM database and will be considered as explanatory variables for the regression model to be proposed.

### 2.5 Regression method

Machine learning can provide some solutions when the accuracy of data is reduced by many measurement errors whose it is difficult to identify the source and when intrinsic indicators can not be exhaustive but sufficient to study a relation with extrinsic indicators. We seek a generic method that could link intrinsic indicators with extrinsic indicators. Our work chose to use a multiple regression model which is the basic case of machine learning. We use as explanatory variables of the model, the set of intrinsic indicators defined above, while the extrinsic indicators derived from the data matching are considered as model-dependent variables. For each matched building, we want to associate a list of values from the explanatory variables and a value relative to each of the dependent variables such as radial distance, angular distance, Hausdorff distance and surface distance.

To achieve a balance between adjustment and parsimony, two methods to determine the optimal regression model for information loss: AIC (Akakie Information Criterion) method and BIC (Bayesian Information Criterion) method seems to be suited for our needs. In the study done on Yang (2005), the comparison made between AIC and BIC uses, suggests using AIC method. When there is a need to predict, the AIC's model should be used but when the goal of learning is to explain, the best model is obtained using a BIC model. AIC gives an effective model while BIC model retrieves the true model among others.

Thus, in this work we used the model obtained by the AIC method, since we believe that this model should form the basis of the expected regression model and should be considered first.

To further penalize the model obtained by using AIC, in order to retain only the most important variables, we apply the LASSO method regression. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The LASSO procedure encourages simple, sparse models (i.e. models with fewer parameters). The goal of LASSO regression is to obtain the subset of predictors that minimizes prediction error for a dependent variable. The LASSO does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model. Variables with non-zero regression coefficients variables are most strongly associated with the response variable. Therefore, when a regression model is carried out it can be helpful to do a LASSO regression in order to predict how many variables your model should contain. This secures that your model is not overly complex and prevents the model from over-fitting which can result in a biased and inefficient model. The result is a regression model containing a strict set of explanatory variables with the advantage of being more interpretable. The LASSO method has been used many times in the field of geographic information. This is the case of the study Inoue et al. (2018) on the geographical segmentation of the real estate market. Using a LASSO method of generalized merging, the author seeks to extract the most important variables among the regional parameters of a price model.

In order to ensure the best quality of the regression model obtained using the LASSO method, we adopted a cross-validation approach that consists of training on the dependent variables on one part of the sample and then estimating the performance of the model with the other part of the sample. We used a *2-fold cross-validation* by repeating the operation 100 times in a bootstrap function. This gives us 100 values of the explained proportion of variance called the *R-squared of the regression.* We analyze the distribution of the explained variance values to characterize the proportion of the explained variance within a confidence interval. This would correspond to the ability of the regression model to estimate a numerical value of the extrinsic quality on a given building. In other words, it refers to an estimate of the expected relative error on the shape or position of a building within a confidence interval in which this predicted variable is expected to fluctuate. The greater the rate of variance explained, the more accurate the prediction and the narrower the confidence interval should be.

## 3. APPLICATION

In this section, we present the results of the entire process described in the previous section applied on two polygon datasets (OSM data and authoritative data) representing buildings.

### 3.1 Presentation of data

In a OSM qualification process, the methodology is applied to two sets of data representing the buildings: on one hand the reference data coming from the IGN's BDTopo building theme with less than 1 m of accuracy and representing more than $20m^2$ area buildings, on other hand the buildings data represented by ways in OSM.

This work was conduced on a study area in "Val-de-Marne department" (East of Paris, France, 94).The study area is a window of $29.77km^2$. In this area, we loaded 29171 buildings in the reference dataset and 38582 buildings in the OSM dataset.

### 3.2 Results of data matching

First, for each feature A, belonging to BDTopo, the data matching algorithm looks for candidates in the OSM dataset before choosing the best one according to the approach detailed in section 2. The threshold to select the candidates is empirically set 30 m. Then the distances (i.e. angular, radial surface, and Hausdorff distances) between A and the selected OSM candidates are computed in order to initialize the masses of belief for each criterion. At the end of data matching, each feature from BDTopo is classified into one of three categories: matched, non-matched, and undecided. The results obtained on the test area are depicted in Table 2.

| type of matching | $Number$ |
|---|---|
| matched | 22989 |
| non-matched | 1143 |
| undecided | 5020 |

Table 2: Results of matching of BDTopo dataset

Figure 5 illustrates the three categories of the data matching results.

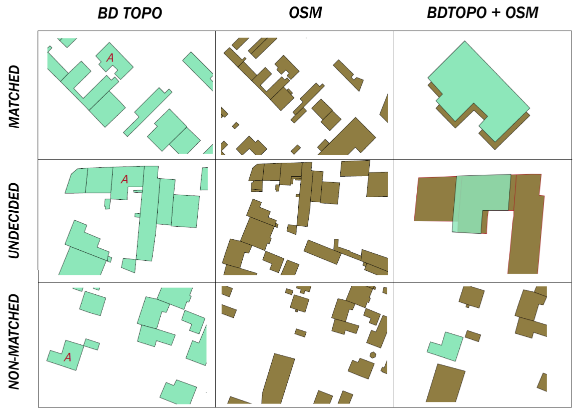

Figure 5: Three cases of data matching: feature ***A*** in BDTopo data (left), candidates OSM data (center), and the result of matching (right), respectively for the cases of matched (top), undecided (middle) and non-matched (right).

To validate the matched features, we would like to compute the proportion ($p$) of *correctly matched faeture*. We know that the true value of $\hat{p}$ is in the interval $\hat{p} \pm \varepsilon(p, \alpha, N)$ with the error margin $\varepsilon$ given by the following equation 1:

$$\varepsilon(p, \alpha, N) = z_{\alpha/2}\sqrt{\frac{p(1-p)}{N}} \tag{1}$$

where $N$ is the number of features and $p$ is the estimated value of the proportion based on these $N$ data points, $\alpha$ is the risk (complementary of the confidence level), $z_{\alpha/2}$ represents the number of standard deviations from which we will deviate around zero considering that the estimator follows a normal distribution.

On a first sample of size *100*, we obtain a rough value of the estimator equal to 0.9759 whose a priori margin error is less than or equal to 10%. Subsequently, we try to calculate the necessary size of a representative sample which would reduce the value of the margin of error with a confidence level at 95% (risk $\alpha$ = 5%). Therefore solving equation (1), we obtain a value $N$ = *174* (with $z_{\alpha/2}$ = 1.96 and $p$ = *0.9759-0.1=0.8759*, $\varepsilon$ = 0.1). By adding 74 pairs of matched features to the sample again, the final check returns an estimator's value equal to 0.9885. To provide a robust validation on the estimator, the value of the margin of error is calculated using equation (1). Thus, we obtain a more accurate estimate of the proportion of correctly matched features equal to 0.9885 over a margin of error $\alpha = 1.5\%$ with a 95% confidence level. The accuracy of the matching results is now equal to $98.85\ (+/-1.5)$.

The verification and validation phases were carried out on a plug-in developed in the GEoxygene platform that illustrates the BDTopo feature with its candidates around it, mentioning for each possible match the values of the criteria [3].

### 3.3 Results of regression

Knowing our aim to infer the dependent variables (extrinsic indicators) from the explanatory variables (intrinsic indicators), we are looking for a regression model that highlights the existence of a significant correlation between an extrinsic indicator and a group of intrinsic indicators. The significance of the correlation of an explanatory variable is reached when the *p-value* is less than 0.05%.

First, a standard multiple regression is performed for each dependent variable with the 14 explanatory variables defined in section 2.4.

Secondly, by a descending process on the calculation of the AIC value, we proceed to remove the non-significant explanatory variables until we obtain a parsimonious model, relative to the smallest AIC value.

For example, in the case of angular distance (dependent variable), Table 3 shows that 12 variables are significant with a p-value of about $10^{-16}$, which justifies the significance of a correlation. With this standard linear regression model, we obtained a share of 31.8% of explained variance relative to the total variance of the dependent variable.

| Coeffs. $a$ | $\hat{a}$ | $\sigma_a$ | $\mathbb{P}[\geqslant \lvert t \rvert]$ |
|---|---|---|---|
| Intercept | $9.57.10^{-2}$ | $1.72.10^{-2}$ | $2.98.10^{-8}$ |
| $a_{rec}$ | $-1.06.10^{-1}$ | $1.73.10^{-2}$ | $7.20.10^{-10}$ |
| $a_{lmx}$ | $5.89.10^{-3}$ | $1.70.10^{-3}$ | $5.36.10^{-4}$ |
| $a_{lme}$ | $-9.33.10^{-3}$ | $4.15.10^{-4}$ | $***$ |
| $a_{lmn}$ | $-3.53.10^{-2}$ | $4.06.10^{-3}$ | $***$ |
| $a_{out}$ | $1.64.10^{-2}$ | $1.67.10^{-3}$ | $***$ |
| $a_{cpc}$ | $-1.88.10^{-1}$ | $1.52.10^{-2}$ | $***$ |
| $a_{cvx}$ | $2.01.10^{-1}$ | $3.02.10^{-2}$ | $2.55.10^{-11}$ |
| $a_{elg}$ | $1.12.10^{-1}$ | $6.16.10^{-3}$ | $***$ |
| $a_{ori}$ | $4.20.10^{-3}$ | $8.67.10^{-4}$ | $1.27.10^{-6}$ |
| $a_{grn}$ | $2.46.10^{-1}$ | $3.05.10^{-2}$ | $8.04.10^{-16}$ |
| $a_{per}$ | $1.48.10^{-3}$ | $8.70.10^{-5}$ | $***$ |
| $a_{are}$ | $-5.40.10^{-5}$ | $9.31.10^{-6}$ | $6.69.10^{-9}$ |

Table 3: Estimated values of coefficients ($a$) with standard deviations $\sigma_a$ and p-value ($\mathbb{P}[\geqslant \lvert t \rvert]$). The symbol $***$ means that p-value is below $2.10^{-16}$.

Finally, to refine the regression model from the AIC method, we develop another form of regression using the LASSO method,

[3] https://github.com/mdvandamme/VisuValideMultiCriteriaMatching/

which should provide a parsimonious model, but with far fewer explanatory variables, with practically the same share of variance explained, thus reducing the regression function to only a few explanatory variables.

With the LASSO regression, we have computed the distribution of values of the explained variance rate for 1000 bootstrap versions using cross-validation with 70% of the data used for training and 30% for validation, on a sample of 19519 matched OSM features. For the case of the angular distance, the mean value is 27.46% and the lower confidence interval ("L95") value is 25.43%. We can therefore affirm that the proportion of variance explained by the regression is greater than 25.43% with a 95% confidence interval. For the radial distance, the explained variance rate is 21% ("L95") while for the Hausdorff distance it is 11% ("L95"). The lowest rate is recorded for the surface distance, which has only 4% ("L95") variance explained by the regression.

In relation to the previous results obtained with the LASSO regression method, we retain the 5 most important variables identified by the LASSO model. They are stated as follows for:

- angular distance:*granularity*, *outlier*, *elongation*, *convexity* and *min-lengths*.
- radial distance:*rectangular*, *outlier*, *compacity*, *perimeter* and *mean-lengths*.
- Hausdorff distance:*granularity*, *elongation*, *size*, *shape* and *max-lengths*
- surfacic distance:*granularity*, *compacity*, *right-angle*, *size* and *shape*.

Figure 6 illustrates the most important variables given by the LASSO regression for the angular distance. Indeed, the smaller the value of the norm L1, the greater the regularization (penalization of the explanatory variables). As long as the value of L1 norm is equal to '0', the model remains empty and as we increase the value of L1 norm (by decreasing the regularization), we witness the progressive appearance of the explanatory variables in the model because their coefficients differ from the value '0'. These variables stand out from '0' in order of importance. We can see on the graph the first 5 important variables ($grn$,$out$,$elg$,$cvx$,$lmn$) that determine the LASSO regression model for the case of angular distance.

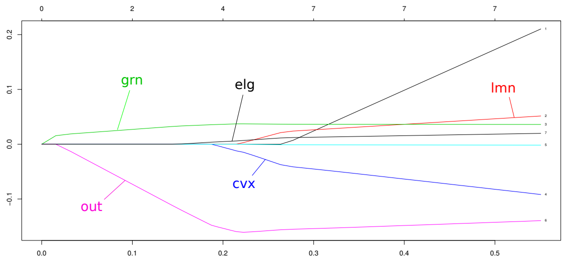

Figure 6: LASSO regression method for angular distance: the first five important variable:grn(granularity), out (outlier), elg(elongation), cvx (convexity), lmn (min-lenghts)

Then, the final regression model could be expressed through a prediction function for extrinsic quality $Y$ from intrinsic indicators $X = (x_{grn}, x_{out}, x_{elg}, x_{cvx}, x_{lmn}) \in \mathbb{R}^5$ in the form of the following equation (e.g. for the case of angular distance):

$$\widehat{Y}(X) = 0.352 + 0.397 \times x_{grn} + 0.04 \times x_{out} + 0.058 \times x_{elg}$$

$$-0.281 \times x_{cvx} - 0.062 \times x_{lmn}$$

Thus, for a building with the following values respectively on the 5 explanatory variables (e.g. for the case of angular distance): $X = (x_{grn} = 5.3 * 10^{-3}, x_{out} = 2.032, x_{elg} = 0.665, x_{cvx} = 0.8807, x_{lmn} = 0.229)$, the model gives a predicted value equal to $0.2207$ varying in this confidence interval: $[0.015; 0.44]$.

### 3.4 Analysis of results and discussion

In order to be able to conclude on the validity of our regression model and its results, we test a number of assumptions known as the multiple linear regression assumptions, namely the normality assumption, the homoscedasticity assumption and the non-multicolinearity assumption.

To test the hypothesis of normality, which assumes that our dependent variable is normally distributed, we draw up the normal QQ diagram. The Normal QQ, or quantile-quantile diagram, is a graphical tool that helps us assess whether a data set is likely to come from a theoretical distribution such as a normal distribution. A Q-Q diagram is a scatterplot created by plotting two sets of quantiles relative to each other, with the observed quantiles on the y-axis and their theoretical Normal quantiles on the x-axis. If the two sets of quantiles come from the same distribution, we should see the points form a roughly straight line.

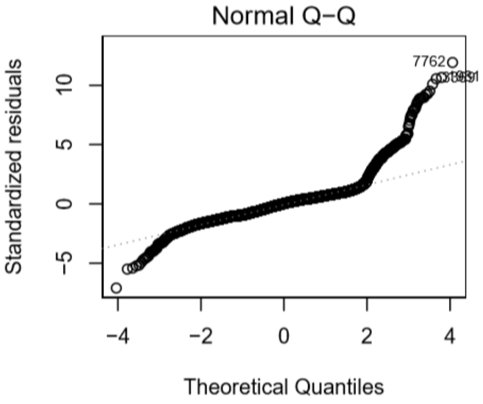

Figure 7: Normal Q-Q

In Figure 7, we try to compare the standardized residuals observed quantiles with their theoretical quantiles. We observe that most of the clouds of the points fall on the line except at the two ends, where they form light tails. Setting aside the light tails at the ends, we assert that the regression model resembles a Gaussian distribution. This confirms the assumption of normality.

Homoscedasticity is observed when the dispersion of the residuals is homogeneous over the entire spectrum of values of the explanatory variables. This is a desirable property since if the residuals do correspond to measurement hazards, there is no reason for the dispersion of the residuals to change with the values of the predictor. To do this, the estimated residuals $\widehat{e_i} = Y_i - \widehat{Y_i}$ (Residuals) are represented as a function of the fitted values, with $\widehat{Y}_i$ a fitted value and $Y_i$ observed value.The estimation technique used assumes that the estimated residuals have a constant variance (not dependent on i).

Thus, in Figure 8, where we represent the estimated residuals as a function of the fitted values, we can see that the residuals disperse randomly independently of the fitted values. This shows that the variance of the residuals is homogeneous and constant. This shows the variance of the residuals to be homogeneous and constant, and it is inferred that the homoscedasticity hypothesis is verified. In addition, it is also visually observed that the residuals do not show any particular organization. This could confirm the linearity hypothesis.

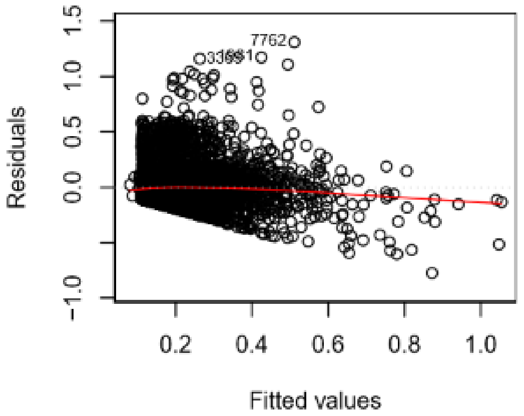

Figure 8: Residuals vs Fitted values

To strengthen the homoscedasticity assumption, the Breusch-Pagan test is used to determine the nature of the variance of the error term (residuals): if the variance is constant, then we have homoscedasticity. On the contrary, if the variance varies, then we have heteroscedasticity. By taking homoscedasticity as the null hypothesis H0, it is enough to check whether the p-value is less than 5%.Thus, the application of the test gives as a result a p-value lower than $10^{-16}$. This confirms the hypothesis of homoscedasticity (Zaman, 2000).

On the other hand, we were interested to see if the residuals were not self-correlated. The Durbin-Watson test is a statistical test designed to test autocorrelation of residuals in a linear regression model. The estimation technique used assumes that the residuals are uncorrelated so that the Durbin-Watson (DW) statistic must be close to the value 2. At the end of the Durbin-Watson test, the DW= 2.0135 is calculated. This confirms the uncorrelation of the residuals of our regression model and therefore the hypothesis of independence of the residuals (Draper and Smith, 1998).

At the end of the verification of the multiple linear regression hypothesis, we tried to verify the hypothesis of non-multicolinearity. Strictly speaking, we speak of perfect multicollinearity when one of the explanatory variables in a model is a linear combination of one or more other explanatory variables introduced into the same model. The absence of perfect multicollinearity is one of the conditions required to be able to estimate a linear model. In non-statistical terms, collinearity occurs when two or more variables measure the ”same thing”. The most traditional approach is to examine the Variance Inflation Factor (VIF) (Stine, 1995). VIFs estimate how much the variance of a coefficient is ”increased” due to a linear relationship with other predictors. For example, an VIF of 1.8 tells us that the variance of this particular coefficient is 80% higher than the variance that would be observed if this factor was completely uncorrelated with other predictors.

However, there is no consensus on the VIF value beyond which multicollinearity should be considered to exist. Some authors suggest to look in more detail at variables with an VIF above 2.5. For our case, in the regression model studied (case of with angular distance), the following VIF values were obtained for the explanatory variables $X = (x_{grn}, x_{out}, x_{elg}, x_{cvx}, x_{lmn})$, respectively the following VIF values:$VIF = (vif_{grn} = 1.046491, vif_{out} = 1.507921, vif_{elg} = 1.387913, vif_{cvx} = 1.566782, vif_{lmn} = 1.538345)$. This confirms the assumption of non-multicollinearity.

## 4. CONCLUSION

The accuracy of spatial data is mostly assessed by computing extrinsic indicators which compare a spatial dataset (e.g. OSM data) with a reference dataset (e.g. authoritative spatial data). Nevertheless, there are cases where reference data is not available (for example countries not having a National Mapping Agency). In these cases, assessing the spatial data quality of a VGI data such OSM become an issue. In this work, we propose an approach that allows to derive extrinsic indicators from intrinsic indicators. By using a robust data matching results, the model was able to establish a regression model estimated for four extrinsic quality indicators by using a panel of 14 intrinsic indicators. This provides an estimate of a possible relative geometric and positional accuracy of a building. Although the results are considerable but modest, they have shown that there is a signal to detect and predict extrinsic quality. More research is needed to improve our results and future research directions for improvement have to be investigated.

A first way to improve our work is to take into account the spatial context based on a notion of neighbourhood for which the values of an extrinsic indicator can be self-correlated. The neighborhood can be generated based on urban considerations, such as alignment with the road, regularity, similarity and proximity of buildings. Then, for this spatial structure, buildings inside it could be aggregated and the inference of the extrinsic indicator from intrinsic indicators can be estimated at this new scale rather that at individual scale.

Second based on the work of Xu et al. (2017), another way to improve the proposed approach is to consider the implementation of a classification algorithm to classify buildings into two classes : satisfactory quality, and unsatisfactory quality, which would be characterized by thresholds in the shape and position criteria. On the buildings deemed of satisfactory quality, a final classification model will be studied which will decide whether a building is good or poor quality on the basis of a classifier constructed by the random forest.

In order to extend our work more widely, it would be appropriate to study how the regression formula built with examples in one area performs on a different area (rural, coastal, montainous, or from another country), and also apply the proposed framework on linear or punctual data.

## References

Ali, A. B. H. and Vauglin, F., 2000. Assessing positional and shape accuracy of polygons in vector gis. In: *Procedings 4th In-*

*ternational Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, pp. 9–12.

Antoniou, V. and Skopeliti, A., 2015. Measures and indicators of vgi quality: An overview. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences* 2, pp. 345.

Ather, A., 2009. A quality analysis of openstreetmap data. *ME Thesis, University College London, London, UK*.

Barron, C., Neis, P. and Zipf, A., 2014. A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS* 18(6), pp. 877–895.

Batton-Hubert, M., Desjardin, E. and Pinet, F., 2019. *L'imperfection des données géographiques 1: Bases théoriques*. ISTE Group.

Bel Hadj Ali, A., 2001. Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification. In: *ECSQARU'01 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems, 19-21 September, Toulouse (France)*.

Brando, C. and Bucher, B., 2010. Quality in user generated spatial content: A matter of specifications. In: *Proceedings of the 13th AGILE international conference on geographic information science*, Springer Verlag: Guimar aes, Portugal, pp. 11–14.

Devillers, R., Bédard, Y., Jeansoulin, R. and Moulin, B., 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science* 21(3), pp. 261–282.

Draper, N. R. and Smith, H., 1998. Serial correlation in the residuals and the durbin–watson test. *Applied Regression Analysis* pp. 179–203.

Fan, H., Zipf, A., Fu, Q. and Neis, P., 2014. Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science* 28(4), pp. 700–719.

Flanagin, A. J. and Metzger, M. J., 2008. The credibility of volunteered geographic information. *GeoJournal* 72(3-4), pp. 137–148.

Girres, J.-F. and Touya, G., 2010. Quality assessment of the french openstreetmap dataset. *Transactions in GIS* 14(4), pp. 435–459.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4), pp. 211–221.

Goodchild, M. F. and Hunter, G. J., 1997. A simple positional accuracy measure for linear features. *International journal of geographical information science* 11(3), pp. 299–306.

Haklay, M., 2010. How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design* 37(4), pp. 682–703.

Hashemi, P. and Abbaspour, R. A., 2015. Assessment of logical consistency in openstreetmap based on the spatial similarity concept. In: *Openstreetmap in GIScience: experiences, research, applications*, Springer, pp. 19–36.

Inoue, R., Ishiyama, R. and Sugiura, A., 2018. Identification of Geographical Segmentation of the Rental Apartment Market in the Tokyo Metropolitan Area (Short Paper). In: S. Winter, A. Griffin and M. Sester (eds), 10th International Conference on Geographic Information Science (GIScience 2018), Leibniz International Proceedings in Informatics (LIPIcs), Vol. 114, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 32:1–32:6.

Kounadi, O., 2009. Assessing the quality of openstreetmap data. *Msc geographical information science, University College of London Department of Civil, Environmental And Geomatic Engineering*.

Neis, P., Zielstra, D. and Zipf, A., 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future internet* 5(2), pp. 282–300.

Olteanu-Raimond, A.-M., Mustiere, S. and Ruas, A., 2015. Knowledge formalization for vector data matching using belief theory. *Journal of Spatial Information Science* 2015(10), pp. 21–46.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C. and Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science* 31(1), pp. 139–167.

Siebritz, L.-A., 2014. Assessing the accuracy of openstreetmap data in south africa for the purpose of integrating it with authoritative data. Doctoral dissertation, University of Cape Town.

Stine, R. A., 1995. Graphical interpretation of variance inflation factors. *The American Statistician* 49(1), pp. 53–56.

Touya, G. and Brando-Escobar, C., 2013. Detecting level-of-detail inconsistencies in volunteered geographic information data sets. *Cartographica: The International Journal for Geographic Information and Geovisualization* 48(2), pp. 134–143.

Truong, Q. T., De Runz, C. and Touya, G., 2019. Analysis of collaboration networks in openstreetmap through weighted social multigraph mining. *International Journal of Geographical Information Science* 33(8), pp. 1651–1682.

Van Exel, M., Dias, E. and Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. In: *Proceedings of the GIScience 2010 Doctoral Colloquium, Zurich, Switzerland*, Vol. XXVII-B1, pp. 14–17.

Van Oort, P. and Bregt, A., 2005. Do users ignore spatial data quality? a decision-theoretic perspective. *Risk Analysis: An International Journal* 25(6), pp. 1599–1610.

Xu, Y., Chen, Z., Xie, Z. and Wu, L., 2017. Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science* 31(10), pp. 1929–1951.

Yang, Y., 2005. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika* 92(4), pp. 937–950.

Zaman, A., 2000. Inconsistency of the breusch-pagan test. *Journal of Economic and Social Research* 2(1), pp. 1–11.