Quality assessment of the OSM data from the mapping party of Baeza (Spain)

Francisco Javier Ariza-López, José Luis García-Balboa,
Virtudes Alba-Fernández, José Rodríguez-Avi and Manuel Ureña-Cámara

[1] Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría, Universidad de Jaén, Spain

[2] Departamento de Estadística e Investigación Operativa, Universidad de Jaén, Spain

Abstract

OpenStreetMap (OSM) is the best known example of VGI. The main objective of this study is to perform a quality assessment of the geographic database (GD) from a specific mapping party (MP), which is a festive and collaborative process. A cluster sampling strategy is applied. The quality elements considered are completeness and thematic accuracy. Results indicate a high presence of omissions which reduces the overall quality.

Keywords: quality assessment, OSM, mapping party

1. Introduction

Despite the large amount of mapping parties (MPs) that have taken place, there are few studies focusing on this phenomenon. This social event was introduced in 2006 (Amelunxen, 2010). This phenomenon is a public call open to all types of public (not limited by age, gender, education, knowledge, etc). However, it is limited in time to some dates and concentrates the effort in an area of special interest, usually a city or any part thereof.

The objectives of an MP are diverse, but related (Perkins and Dodge, 2008; Haklay and Weber, 2008): fill gaps in coverage, galvanize the mapping community, data collection exercises, create and annotate content for localized areas, introduce new users and contributors, creating and fostering local OMS user groups, creating a vibrant social community around the project.

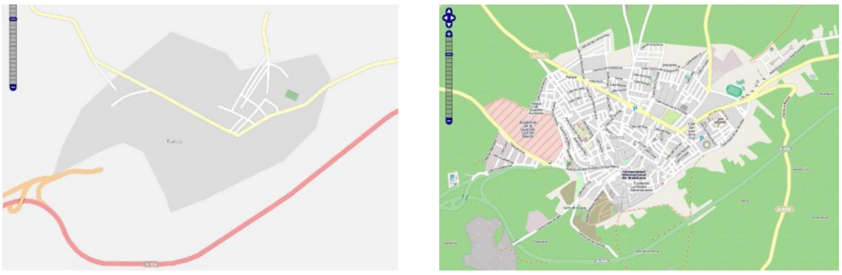

Figure 1: The town of Baeza in OSM before (left) and after (right) MPB.

Also Haklay *et al.* (2010) demonstrates that VGI must be approached as heterogeneous datasets that should be evaluated locally and not globally. As an example, in this paper we focus on assessing the quality of a data set from a MP celebrated in Baeza (Spain) (MPB) (see Figure 1). This work is novel because: a) it is the first evaluation of the quality of a MP, b) the reference is the real world (most studies on the quality of OSM do not perform a ground-truth control) and c) we use the concepts of metaquality proposed in ISO 19157 for reporting the quality of results.

2. Identification of components to describe quality

Following ISO 19113, the analysis is centered on the data quality elements completeness and thematic accuracy. The data quality subelements are commission, omission and classification correctness. The scope is the whole geographic database (GD) of MPB. Other data quality elements fall outside the scope of this study. The main goal of organizers of the MPB (provincial government of Jaén, Spain) was to incorporate all the items from a list of feature types by the volunteers. The organizers had previously added the whole street network and had stated they were not much concerned about positional accuracy or logical consistency of the items but for completeness and thematic accuracy.

From ISO/TS 19138 measure 7 (rate of missing items) is selected for omission, measure 3 (rate of excess items) for commission and measure 61 (misclassification rate) for classification correctness. Although the errors are counted, the results are to be presented as quality Q, that is to say as features without defects. This allows a more natural understanding of the final results. Therefore three new measures (Table 1) have been defined, which have been identified as 1001, 1002 and 1003.

Table 1: Definition of new data quality measures (following ISO/DIS 19138).

Line	Component	Id 1001	Id 1002	Id 1003
1	Name	Rate of correctness in omissions	Rate of correctness in commissions	Rate of correctness in classification
2	Alias	Omission correctness	Commission correctness	Classification correctness
3	Data quality element	Completeness	Completeness	Thematic accuracy
4	Data quality subelement	Omission	Commission	Classification correctness
5	Data quality basic measure	Correct items rate	Correct items rate	Correct items rate
6	Definition	100-complement of measure 7 of ISO/DIS 19138: number of missing items in the dataset in relation to the number of items that should have been present.	100-complement of measure 3 of ISO/DIS 19138: number of excess items in the dataset in relation to the number of items that should have been present.	100-complement of measure 61 of ISO/DIS 19138: number of incorrectly classified features in relation to the number of features that are supposed to be there.
7	Description	--	--	--
8	Parameter	--	--	--
9	Data quality value type	percentage	percentage	percentage
10	Data quality value structure	--	--	--
11	Source reference	--	--	--
12	Example	--	--	--
13	Identifier	1001	1002	1003

2. Sampling design

Here a statistical sampling is essential because a full inspection is not viable. Two different strategies can be considered:

- Centered on the type of geometry (point, line and area), which supposes the design of three different simple random samplings (SRS). But the omission of features cannot be assessed with an SRS because of the lack of exhaustive knowledge of the population.
- Centered on a global view, applying a cluster sampling (CS) over the framework defined by the city blocks and open spaces from the city of Baeza. The CS simplifies the fieldwork and it considers that each of the clusters has the same heterogeneity as the population.

The CS was selected in order to assess the omission of features. In CS, the population is grouped into *N* clusters. Here by cluster we mean any of the city blocks and open spaces from the city of Baeza.

Let us denote by M_i the population size of *i*th cluster, and $M = \sum_{i=1}^{N} M_i$ the population size. As the sizes of the clusters are not homogeneous, the 'type-ratio' estimator is suggested. Let $\hat{P}_R$ be the type-ratio estimator of the population proportion associated with variable *A* as in Equation (1):

$$\hat{P}_R = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i} \qquad (1)$$

where $A_i = \sum_{j=1}^{M_i} A_{ij}$ and A_{ij} is the value of *A* (1 or 0) for the element *j* belonging to the cluster *i* in the sample. An unbiased variance estimator is (Equation 2):

$$\hat{V}\left(\hat{P}_R\right) = \frac{N(N-n)}{n\left(n-1\right)M^2}\left(\sum_{i=1}^{n} A_i^2 + \hat{P}_R^2 \sum_{i=1}^{n} M_i^2 - 2\times\hat{P}_R \sum_{i=1}^{n} M_i A_i\right) \qquad (2)$$

and a confidence interval at level $100(1-\alpha)\%$ is obtained as (Equation 3):

$$\left(\hat{P}_R \pm Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{P}_R)}\right) \qquad (3)$$

In order to know an estimation of the value of *N*, an official product was used. By this way the size of the population is *N = 400.*

Due to the lack of awareness about the population and about the errors in the GD, a pilot scheme needs to be introduced. Therefore a cluster is selected randomly for each zone of MPB, which results in 20 clusters. In addition 7 open spaces are selected by an SRS. Figure 2 shows the spatial distribution of the 27 clusters.

Each cluster of this sample has to be inspected, and therefore the final sample of features to be inspected is 100% of the features from the real world which are inside any of these clusters. Fieldwork is essential since the GD contains detailed urban data, which is in many cases difficult to find in other databases. It is performed by going round all the streets of the cluster and considering two alternatives: (a) using a printout of the GD inside the cluster and comparing it with the real world, checking omissions, excess features and misclassifications, and (b) using a blank map with the boundary of the cluster and writing an independent version of the cluster, with all the features of the real world. Both methodologies were performed independently by two different people.

After the inspection, a total of 995 features of the real world had been inspected, which allowed us to perform an estimation of the population of features of *M = 14740*, from a mean amount of features per cluster of 36.85. With this estimation and from the analysis of the results, the pilot scheme is finally considered of enough statistical reliability, so no more clusters are considered for this exploratory study.

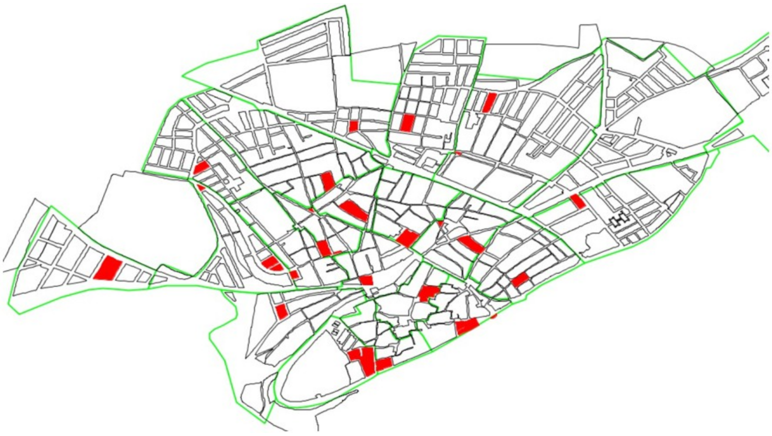

Figure 2: Sample distribution of the clusters in the CS. Red areas (27) represent the clusters of the sample. Green lines represent the zones of the MPB. Black lines represent the city blocks of the town.

3. Results

Table 2 shows a global estimation of the quality of 45.63% with a precision near 12%. Regarding the type of error, and therefore the quality measure (1001, 1002 and 1003) for each quality subelement, Table 3 shows the estimated qualities and the corresponding IC at level 95%. The quality estimation is very poor in the omission of features, lower than the 50%, with a precision ε near 12%. On the other hand, the quality result is near 100% in relation to the commissions and classification of features, with the precision ε lower than 2%.

Table 2: Global estimation of quality and the confidence interval in the CS

Sampling	Sample size	Estimated quality (%)	ε (%)	CI 95% (%)
CS	995	45.63	11.69	(33.93, 57.32)

Table 3: Estimation of quality and the confidence interval for each quality measure in the CS

Quality element	Quality subelement	Scope	Quality measure	Quality result (%)	ε (%)	IC 95% (%)
Completeness	Omission	GD of	1001	47.84	11.59	(36.25, 59.43)

	Commission	MPB	1002	98.79	0.97	(97.82, 99.77)
Thematic accuracy	Classification correctness		1003	98.99	1.72	(97.28, 100.00)

3. Conclusions

In this paper we focus on assessing the quality of a data set from a specific MP, held in the town of Baeza (Spain) (MPB). An important conclusion is that the GD is not complete because the omission correctness is low (48%), which significantly reduces the global quality. Nevertheless the data included in the GD shows high levels of quality, with high commission correctness (99%) and high classification correctness (98%).

Our experience indicates that the goal of the organizers was too ambitious, therefore the list of feature types so much extensive for this MP. The volunteers could not have enough time to include all items from the universe of discourse.

As a final conclusion it can be indicated that the GD analysed is a multiple interpretation of the real world, with a high presence of omissions. Also, the GD is heterogeneous because of the fieldwork and office work of the volunteers, who apply different criteria apart from the effort by the organizers of MPB to avoid this problem. The GD can be considered as a ‘hobby’ product, but not a professional database, which could be used by NMAs. Nevertheless we consider that the data capture with the OSM data model is useful for spreading the use of the geographic information and geomatics, and the Web 2.0. The MP social event has not only a ludic and informative interest, but also an advertising interest from potential sponsors.

The CS appears to be the most favourable sampling scheme to this type of assessment. The advantages are the following: it includes all quality elements and quality subelements (i.e. omissions), and the field work requires a minor effort when checking clusters instead of independent features, therefore the efficiency is higher and the fraction of the population which is checked is higher.

Another novelty in this study is the information about the metaquality, specifically the element confidence, as proposed by the future IS ISO 19157. The precision of the sampling has been specified for each quality result. The confidence interval is too wide for the omission results. This means that the GD was worse than expected and therefore the sample size should have been larger.

References

Amelunxen, C. (2010), “An Approach to gecoding based on volunteered Spatial Data”. In: Zipf, A. *et al*. (eds.). *Geoinformatik 2010*. Die Welt im Netz. Kiel.

Haklay, M., Weber, P. (2008), “OpenStreetMap: user-generated street maps”. *Pervasive Computing, IEEE*, Vol. 7(4):12-16.

Haklay, M., Basiouka, S., Antoniou, V., Ather, A. (2010), “How many volunteers does it take to map an area well? the validity of linus’ law to volunteered geographic information”. *The Cartographic Journal* , Vol. 47(4): 315 – 322.

Napolitano, M., Mooney, P. (2012), “MVP OSM. A tool to identify areas of high quality contributor activity in OpenStreetMap”. *The Bulletin of the Society of Cartographers*, Summer 2012.

Perkins, C., Dodge, M. (2008). “The potential of user-generated cartography: a case study of the OpenStreetMap Project and MapChester mapping Party”. *North West Geography*, Vol. 8(1):19-32.