Cuadernos de Investigación Geográfica *Geographical Research Letters*	2023	Nº 49	pp. 191-210	EISSN 1697-9540

http://doi.org/10.18172/cig.5280

EVALUATION OF THE QUALITY OF THE VOLUNTARY GEOGRAPHIC INFORMATION FOR THE ROAD NETWORK IN BOGOTÁ D.C.

LUIS A. NIÑO BELTRÁN [1], AQUILES E. DARGHAN CONTRERAS*[1], LIBIA D. CANGREJO ALJURE[2], EDWIN F. GRISALES CAMARGO[2]

[1] Universidad Nacional de Colombia. Facultad de Ciencias Agrarias. Bogotá, Colombia.

[2] Universidad Nacional de Colombia. Facultad de Ingeniería. Bogotá, Colombia.

ABSTRACT. The production of Voluntary Geographic Information has been growing considerably and continues to be an active area of research. However, the lack of knowledge about the quality of information generated on a voluntary and participatory basis raises challenges and questions about its use. In the review carried out for the Colombian case, no studies related to the subject were identified; consequently, this study is presented on the evaluation of the quality of this type of information on the road network of Bogotá with respect to completeness, positional accuracy and thematic accuracy. This evaluation was carried out by means of a semi-automatic process that uses a mobile buffer and the centroid of the roads to make the corresponding comparisons between two data sources. The results found reveal that the method used allowed to compare up to 85.0% of the data, and that the OpenStreetMap mesh has a completeness of 85.4%, over the entire area of Bogotá. A positional accuracy of 3.98 m and a thematic accuracy related to the percentage of error in the attributes: Road hierarchy, direction of flow and road naming of 35.8%, 15.0% and 34.6% respectively. The quality evaluated through completeness, positional and thematic accuracy in synergistic terms is deficient with respect to the minimum quality levels established in the standard data model, however, the evaluation for each of the attributes shows an acceptable quality in terms of completeness and thematic accuracy.

Evaluación de la calidad de la Información Geográfica Voluntaria de la red vial de Bogotá D.C.

RESUMEN. El aumento en la producción de Información Geográfica Voluntaria (VGI) ha venido creciendo considerablemente y se han realizado diversos estudios al respecto. Sin embargo, el desconocimiento de la calidad de la información generada en forma voluntaria y participativa, plantea retos y cuestionamientos sobre el uso de este tipo de información. En la revisión efectuada para el caso colombiano no se identificaron estudios relacionados con el tema; en consecuencia, se presenta este estudio sobre la evaluación de la calidad VGI de la malla vial de Bogotá respecto a la completitud, a la exactitud posicional y a la exactitud temática. Esta evaluación se realizó por medio de un proceso semiautomático que usa un buffer móvil y el centroide de las vías para realizar las comparaciones correspondientes entre dos fuentes de datos. Los resultados encontrados revelan que el método empleado permitió comparar hasta el 85,0% de los datos, además se calculó que la malla OSM (OpenStreetMap) tiene una completitud del 85,4%, sobre toda el área de Bogotá. Una exactitud posicional de 3,98 m y una exactitud temática relacionada al porcentaje de error en los atributos: Jerarquía vial, Dirección de flujo y Nombramiento de las vías de 35,8%, 15,0% y 34,6% respectivamente. La calidad VGI evaluada a través de la completitud, la exactitud posicional y la exactitud temática es considerada conjuntamente como deficiente, Sin embargo, evaluada la calidad separadamente a través de las medias indicadas, se concluyó que los datos VGI gozan de una completitud aceptable, una exactitud posicional óptima y una exactitud temática deficiente.

Key words: OpenStreetMap, road network, regular expressions, positional accuracy, thematic accuracy.

Palabras clave: OpenStreetMap, red de carreteras, expresiones regulares, precisión posicional, precisión temática.

Received: 2 December 2021

Accepted: 1 June 2022

[*]**Corresponding author:** Aquiles E. Darghan Contreras, Universidad Nacional de Colombia, Facultad de Ciencias Agrarias. Bogotá, Colombia. E-mail: aqedarghanco@unal.edu.co

1. Introduction

With the appearance of new and improved ways of acquiring, sharing and updating information through *online* platforms and devices (WEB 2.0), the number of contributors who can create, store and edit geographic data has increased (Hudson *et al.,* 2009; Darwish and Lakhtaria, 2011). This facilitates obtaining information at low cost and is faster than traditional methods. However, these advantages are accompanied by certain problems related to quality (Esmaili *et al.,* 2013; Moreri *et al*., 2018). This dynamic is known under the term "Volunteered Geographic Information" (VGI) (Janelle and Goodchild, 2011), where the data is provided by citizens who act as sensors on the world that surrounds them, generating geographic information (Goodchild and Li, 2012).

One of the most successful VGI projects currently in force is OpenStreetMap, where the geographic information collected comes from multiple users and sources, causing the collected data to be accompanied by a high degree of heterogeneity. To minimize this high variability in terms of quality, OpenStreetMap has created a production model published in Wikipedia (Haklay and Weber, 2008), in order to indicate to users how to add and edit geographic data and thus standardize the coding of the information. However, due to the increasing use of VGI data, and the lack of reliability, many researchers have focused on studying its quality and usability (Haklay 2010; Fonte *et al.*, 2015; Yan *et al.*, 2017). All these studies have made it possible to delve into the classification of VGI data given the nature by which they were collected. Various authors such as Goodchild and Li (2012) have discussed alternatives for evaluating the quality of data, among them are the use of user groups to validate the edits made (*crowd-sourcing*) as well as the use of documentation to control its quality. Researchers such as Elwood *et al.* (2012), Foody *et al.* (2013), Jonietz and Zipf (2016), Wu *et al*. (2021) have created frameworks where VGI quality is evaluated and controlled using frameworks mentioned previously.

However, in response to the need to ensure the quality of geographic data, a series of quantitative and qualitative methods have been created, including measures and indicators for VGI (Antoniou and Skopeliti, 2015). The principles and guidelines of quality measures are given by the International Organization for Standardization (ISO) in its most recent version (ISO 19157: 2013) where the following quality elements have been defined: completeness, positional accuracy, logical consistency, temporal quality, thematic accuracy, among others.

Using the quality measures, several research projects focused on ensuring VGI quality have been developed. These quality measures have been evaluated by comparing objects referring to road nets with respect to official data (Dorn *et al.*, 2015; Mahabir *et al.*, 2017), and most authors have highlighted the presence of heterogeneity in quality. These methodologies are because official data are created under high quality standards (Antoniou and Skopeliti, 2015); and, therefore, it makes sense to try to use them as reference elements.

In the assessment of completeness, the most widely used method has been the one developed by (Goodchild and Hunter, 1997) where from two linear data sources, one called official and the other

as VGI source, buffers are created (Zhang *et al.*, 2019) around the reliable source and all those VGI elements that are within the range of influence of the buffer are selected, resulting in a count of elements and distances that allow determining the absence or excess of elements in the VGI source. A modification of this method has been used by (Da Costa, 2016a) where the completeness of the data was evaluated by quantifying the excesses and defects of polygonal objects. More advanced methods have proposed evaluating the completeness of the data by means of semi-automatic and automatic matching (*matching*), such is the case of research developed by Abdolmajidi *et al.* (2015) where the technique called *extended node-based* and consists of comparing the geometry by means of the coincidence of nodes and the topological evaluation of the elements. This study was based particularly on the matching of complex road structures. Others such as Bazeley and Jackson (2013) assessed completeness in point-like entities, using a more robust method than the one developed by (Haklay, 2010) and demonstrating that a count comparison is not sufficient to describe the differences between two data sources.

Regarding the evaluation of positional accuracy, the most common methods of evaluation consist of creating geometric matches and calculating the distance to the centroid of the line, where the positional accuracy is calculated by means of the Root of the Mean Square Error (RMSE) for the desired component. For example, Haklay (2010) found that OpenStreetMap data compared manually with an official UK source contained an error of 8.5 m. Ludwig *et al.* (2011) compared a German OpenStreetMap road grid to a private resource by matching from automatically formed linear objects to finally calculate positional accuracy. Graser *et al.* (2014), developed an algorithm to evaluate the quality of road networks addressing positional precision.

Regarding thematic accuracy, some researchers measure the percentage of correct classification of the type of road attributes (Antoniou and Skopeliti, 2015). Others measure accuracy using the univariate kappa index (Arsanjani *et al.*, 2015). There are also studies focused on determining the correct classification of attributes using a confusion matrix and a series of spatial crossings (Codescu *et al.*, 2011) where important differences in quality have been found between urbanized and rural areas. as well as on the type of entities studied.

Other relevant works regarding the quality of the VGI data are related to the quality measures of interest in the field of semantic similarity and the application of data mining techniques (Ipeirotis *et al.*, 2014). One of the advantages of using data mining is that it works on an approach independent of laws and knowledge of geography, and independent of social or multi-source approaches for assessing the quality of VGI (Mobasheri *et al.*, 2018).

Although the proposal to give credit to the information provided by users who are in direct contact with their environment has the support of experts in the area (Goodchild and Li, 2012), there are critical positions that affirm that by coming from these data from multiple users, their quality is compromised, directly affecting the reliability of the data (Ballatore and Zipf, 2015) and making the information often unusable because its quality is not determined and it has too much ambiguity (Esmaili *et al.*, 2013). The problem of using this type of information lies specifically in the lack of knowledge of its quality, therefore the authors mentioned in this article have investigated it in order to open the doors for its use (Antoniou, 2011). Measuring positional accuracy for road data from England contemplates the possibility that these will be used by state agencies, supported by the results found. Other authors argue that knowledge of VGI quality allows the use of data to complement the information gaps between data created by state agencies in developing countries (Mahabir *et al.*, 2017; Lin, 2018; Vannoni *et al.*, 2020).

Given the identified need for quality studies for the road network in Colombia, the objective of this article is to evaluate the quality of the VGI data on the Bogotá road network, considering the quality measures completeness, positional accuracy and thematic accuracy, using a semi-automated method.

2. Materials and methods

2.1. Study area

The study area is located on the eastern cordillera of the Andes at an altitude of 2,650 meters above sea level. The city of Bogotá, capital of the Colombian territory, has an area of 1,732 km^2 and it has a population of around 8.0 million inhabitants, making it the most populated city in the country. This city is located on 4.609°N, -74.082°W (Fig. 1). Its perimeter limits are inscribed within the coordinates: West: -74.450°; East: -73.986°; South: 3.731°; North: 4.837°.

Bogotá's road system is structured under the interconnection of four hierarchical grids according to their functional characteristics in terms of centrality. On the other hand, and according to the district secretary of finance, most of the marginal areas are in the south- west and south-east of Bogotá, where facilities provided by the government are scarce compared to the other localities. Bogotá's road network comprises 8,196 km of roads, of which 1,296 km correspond to arterial mesh-type sections, 1,691 km to an intermediate road mesh type, 5,091 km to a local road network. The rest of the km correspond to a rural road network, undefined and projected (Cristancho and Triana, 2018).

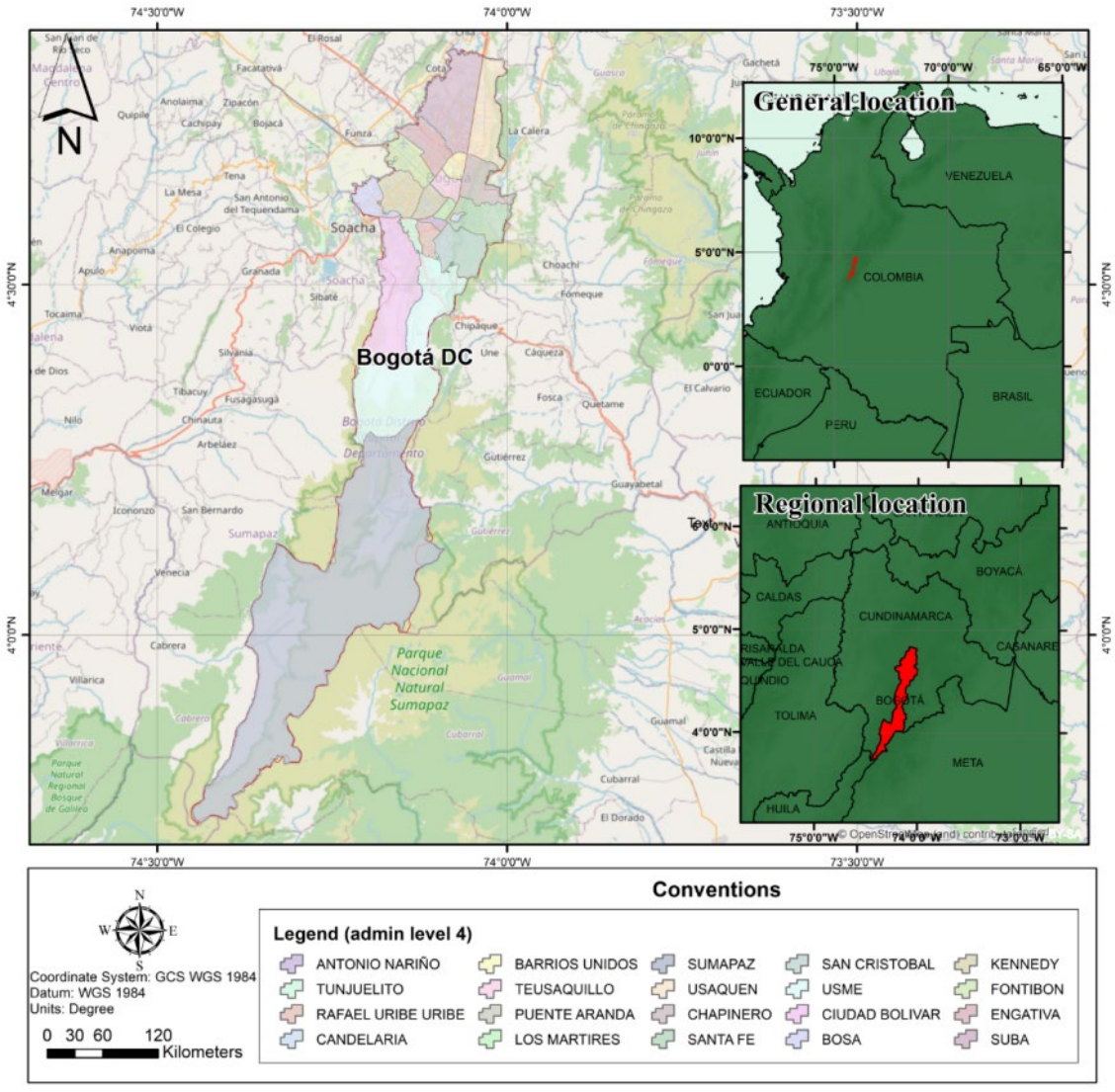

Figure 1. Location of the study area.

2.2. Voluntary Geographic Information quality measures

The elements for measuring the quality of geographic information are defined in ISO 19157: 2013 Standard that establishes a series of measures and sub-categories establishing the degree of quality of geographic information. From this standard three definitions have been extracted which refer to the measures used for the development of this work. Here is a brief description of them.

2.2.1. Completeness

The concept of completeness was defined in (Kresse, 2012) as the presence and/or absence of objects, attributes and relationships represented in the product with respect to its technical specification and a source of higher accuracy. On this item, there are two sub-elements of quality which are known as commission and omission. Commission refers to excess data, the number of elements within the dataset that should not be coded because they do not exist in reality (Number of excess elements) and a rate of excess elements, which is calculated as the ratio of excess elements to the number of elements that should have been present in the dataset. Omission counts the amount of missing data in a product according to the technical specification and a more accurate data source.

2.2.2. Spatial/positional accuracy

The accuracy of the position assesses how well the georeferenced value of an object is related to its respective reality in the field (Van Oort, 2006; Stein *et al.*, 2016). Its sub-elements of quality are absolute and relative accuracy. The absolute accuracy refers to the proximity between the observed values vs. their true values, while the relative accuracy refers to the position of an element with respect to the other elements contained in the data set.

2.2.3. Thematic accuracy

This evaluates the accuracy of the qualitative attributes, as well as the classification of the characteristics and their relationships. The thematic accuracy sub-item involved was the qualitative accuracy of an attribute. This consists of identifying the differences between the assigned qualitative values vs their real values. Illustrative examples are the labels on the road mesh. For the development of this project, thematic accuracy will be considered as a synergy between the completeness of the attributes and their accuracy, following the definition made by Koukoletsos *et al.* (2012).

2.3. Data

The data comes from two sources: (I) the IDECA (The Spatial Data Infrastructure of Bogotá) road grid in its version 09.17 (September 2017), considered as reference source, and (II) the VGI road grid coming from OpenStreetMap downloaded in September 2017. The IDECA reference data were produced under ISO 19157: 2013 quality standards and following the Colombian Technical Standard 5662 (NTC 5662 - Technical specifications for cartographic products). These data have zoom 18 (OSM, 2017a) and were coded under the MAGNA-SIRGAS spatial reference system (EPSG 4686). Positional errors do not exceed one meter (1m) distance at a reliability level of 95.0%, while thematic accuracy has a misclassification rate of 5.0% (IDECA, 2017). The IDECA road network consists of 136,958 sections and 8,196 km. The IDECA road grid contains the attributes: Road direction (From, To, Two-way); Hierarchy (Arterial, Intermediate, Local, Pedestrian, Rural, Planned and Undefined); Road nomenclature or Road type. The IDECA attributes for the geometric object of line sections are: type of road, road name and label, all of type string with abbreviation (*MVIType*), those of type long (Long) and with abbreviation (*MVITCla*) and finally the direction of the road of type string with label (*MVISVia*).

The OpenStreetMap data were downloaded from the *geofabrik* site (Geofabrik, 2018), a server that contains data extracts that are updated daily, within the data, the OpenStreetMap road geometry layer refers to a series of attributes very similar to those previously exposed, however, the classification of the attributes called (*Key*) in OpenStreetMap are governed by the parameters (*Values*), which are subcategories for each element within the OpenStreetMap classification. Information regarding the structure of the OpenStreetMap database is published on the OpenStreetMap Wiki page (OSM 2017b), where mapping guidelines for users can be found. The OpenStreetMap road grid contained at the time 73,454 road sections and 220,362 nodes and their spatial reference WGS84 (EPSG 4326). The attributes taken for the analysis of the road mesh for the online sections were: *Name* (string), *Fclass* (Long) and *Oneway* (string). In addition to this dataset, the Bogotá localities from IDECA were used. Table 1 summarizes the sources used for the development of the research.

Table 1. Data sources used.

Element	Format	Description	Note
OpenStreetMap data Road network	shp.zip (SHAPE)	Voluntary Data extracted from geofabric server	(Geofabrik, 2018)
IDECA Database	Geodatabase (GDB)	Official data of the city of Bogotá - Road network	(IDECA, 2017)
Digital Orthophoto of Bogotá 2014	Web Map Service (WMS)	WMS service from IDECA	Geographical extension of 49,000 ha
OpenStreetMap Wiki	HyperText Markup Language (HTML)	Website	OSM projects and guidelines

2.4. Extraction and Exploration of OSM Data

The extraction of the OpenStreetMap data was carried out using the *urllib2* Python library. This library allowed the creation of a connection to the URL page (Uniform Resource Locator) that stores this data making it possible to obtain them at will, allowing configuration of the Python code to the location data downloaded. Although semi-automatic data processing is not the main object of this research, it does play an important role in ensuring the credibility of the results (Abdolmajidi *et al.* 2015). As most of the spatial operations were performed using Python through the *arcpy* library, it was necessary to modify the original name of the shapes files coming from OpenStreetMap, since their original syntax did not allow the "*arcpy*" spatial tools to recognize the OpenStreetMap data. All these data came from the *geofabrik* server (Geofabrik, 2018), and were available at the country level. For the specific case in OpenStreetMap data from the Bogotá road network, an automatic spatial clipping had to be performed to select the data of interest framed in the Bogotá area. The functions *arcpy.MakeFeatureLayer_management*,*arcpy.SelectLayerByLocation_management*, and *arcpy.CopyFeatures_management* were used.

These functions, created by the group (Enviromental system Research Institute; ESRI) allow creating, according to the order shown, temporary layers to store results, to select data with respect to a spatial parameter, and finally to copy the results stored in the created temporary layer. Once the OpenStreetMap data had been downloaded and extracted according to the work area, the road mesh reference data was downloaded from the IDECA database.

The exploration of the data, important to give the first assessment of the state of the layers, was carried out using the attributes of the road mesh described above, for which a total count of records was made for each layer, subsequently, for each of the attributes, a count per domain was established, as well as a calculation of percentages. On the other hand, other basic descriptive statistics were calculated as well as the representation by means of histograms to study their distribution in each of the layers analyzed.

Because OpenStreetMap contains a set of rules for the coding of road attributes in each country, it was necessary to study the OpenStreetMap mapping guide created for Colombia, together with the definition of global parameters created in OpenStreetMap (*Map_Features*). These parameters assign labels to each coded element within OpenStreetMap, each label containing a *key* that represents the subcategories for each class. The guiding review criteria for OpenStreetMap were: Labelling of roads for classification, OpenStreetMap road classification proposed in the Map features page and Common coding errors detected by OpenStreetMap.

2.5. Entity-Relationship model

Once the coding guidelines for OpenStreetMap were understood, it was possible to clearly identify the fields that needed to be reclassified for the purpose of comparison by creating an Entity-Relationship (E-R) model, broadly following the work of (Da Costa 2016b) and using the variables described in Table 2. Some categories within the OpenStreetMap *Fclass* field could not be related to a category in IDECA, so they had to be excluded from the study.

Table 2. Attributes revised.

Fields	Description	Subcategories	Data
MVIType	Type of track (Cl, KR, TV, DG)	No	IDECA
MVIName	Name of the road	No	IDECA
MVISVia	Direction of the road	Yes	IDECA
MVIEtiquet	Main road name	No	IDECA
Type of classification	Road hierarchy	Yes	OpenStreetMap
Fclass	Classification of the road	Yes	OpenStreetMap
Name	Name of the road	No	OpenStreetMap
Oneway	Direction of the road	Yes	OpenStreetMap

The construction of the Entity Relationship model (Thalheim, 2013), the following guidelines were taken into account: I) *Relationship in the road sense fields*: where a road sense in IDECA could have more than one value in OSM, respecting the values To, From and Double sense. II) *Road hierarchy*: It was established that the OSM road network had too many categories, which could be reduced and standardized to those coded in IDECA, respecting that the many classes within the OSM road hierarchy could correspond to a single value in the IDECA hierarchical classification. III) *Road naming*: all OSM data had to be standardized according to the methodology followed in IDECA, separating the type of road and the corresponding name, respecting the one-to-one relationship. With the comparison structure defined, the standardization fields were created in the respective layers, after which automatic queries were created to select and group the elements according to the rules established in Figure 2.

Regarding the (E-R) model, it allowed to graphically observe the relationships between entities in a database (Silberschatz *et al.* 2002; Thalheim, 2013). On the other hand, the (E-R) model has cardinality correspondence restrictions, i.e. the degree of association that entities have. These restrictions are defined as follows: One to One (1:1), One to Many (1:*) and Many to Many (*:*). To create the model, the relationship in the road direction fields was taken into account, where a road direction in IDECA could have more than one value, in OpenStreetMap, respecting the values To, From and Two-way; the Road Hierarchy for which it was established that the OpenStreetMap road network had too many categories, which could be reduced and standardized to those coded in IDECA, respecting that the many classes within the OpenStreetMap road hierarchy could correspond to a single value in the IDECA hierarchical classification; the Road Naming, where all OpenStreetMap data had to be standardized according to the methodology followed in IDECA, separating the type of road and the corresponding name. Respecting the ratio (1:1).

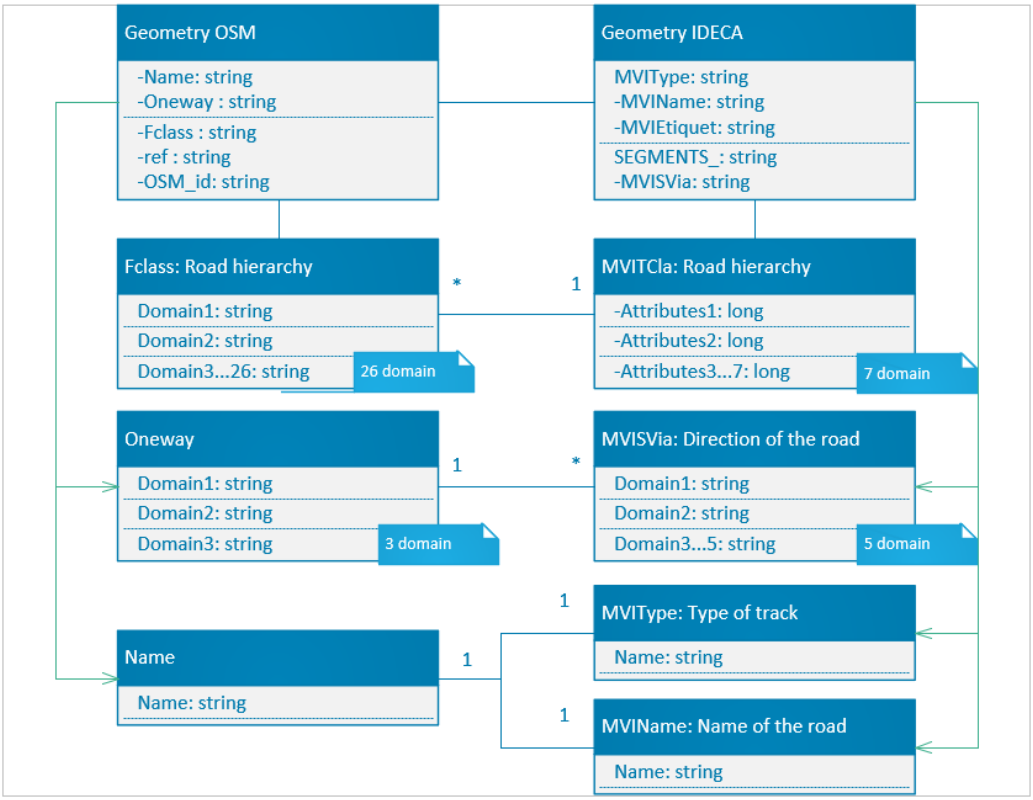

Figure 2. Entity-Relationship model.

With the comparison structure defined, we proceeded to create the standardization fields in the respective layers, after which automatic queries were created to select and group the elements according to the established rules. Some functions used to automate the process were: *arcpy.Select_analysis* (function to select data using Structured Query Language - SQL), *arcpy.AddField_management* (function to create new fields, which contained the new classification of the data according to the type of attribute described in the E-R model and the Creation of fields that would house the results of the attribute comparison).

Once the comparison structure was defined and standardized, the fields composed of character strings were compared. As expected, the OpenStreetMap data contained inconsistencies in terms of text typing, which forced us to standardize the names according to road type and street name. This study did not analyze the compound name because of the associated difficulties. Standardization of names was achieved using regular expressions known as *Regex* (Sharma and Nagpal, 2020). The regular expressions used are presented in Table 3.

Table 3. Standardization of road names.

Expression	Description
[Calle]+	Look for any patterns related to the letters C-a-l-e. Even if there are spaces between the letters.
[0-9]+[A-Za-z]*	Select expressions that begin with numbers and continue with letters.
^([0-9]+)$	Pattern for selecting expressions that start with numbers but do not have letters.
^([0+)$	Selects strings that start with 0 and have no other number associated with them.
^ [A-Z] +$	Select expressions that begin with letters and continue with letters.

With these expressions, general patterns were found to detect the type of track even if it was incorrectly typed. It was also possible to standardize the name of the main pathway, cleaning spaces at the beginning and end of each record. The following rules were established: (I) There cannot be two types of roads in the same label: 34th *Calle 34; calle 34*, (II) There cannot be spaces at the beginning, (III) There cannot be special characters = / # -$. (IV) There cannot be any combination of Street, Street and Avenue that does not end in a vowel. (V) Nomenclature and names must be separated for comparison. (VI) Bis must be in capital letters and without space between the number: 4 Bis: 4BIS. The standardized data was automatically copied into the column created to accommodate this standardization.

Once the data were standardized and columns were created to store the results of the comparison, homologous objects in both datasets were determined, a series of functions were employed to create spatial relationships through the use of indexing techniques, minimum distance calculation, node counting, use of buffers and calculation of distances (km). This could be done by broadly following the methodologies created by Goodchild and Hunter (1997) and Haklay (2010), used to measure positional accuracy and completeness. In our case, a moving buffer was applied depending on the distance of the nearest object to the central node of each line, and Bogotá localities were used instead of grids to measure distances (km) to detect missing elements.

In order to achieve this by means of an automatic process, the following steps were followed: I) the OpenStreetMap Road grid was transformed to the geographical coordinate system EPSG:4686, II) the OpenStreetMap roads were cut according to the intersection with other road segments, in order to get the roads with lengths as close as possible to the IDECA source, III) then the fields for measuring km were created in each of the layers. It is clarified that in this step all the fields of road hierarchy, direction of flow and name of the roads were already standardized, IV) the minimum distances to each road node center of each object were calculated, V) buffers with the minimum distance of the nearest segments were used, VI) the point to point spatial relations were indexed, VII) a spatial crossing with the localities was created in order to agglomerate results and finally VIII) the differences between the attributes were calculated.

The assessment of data completeness related to missing or excess elements was calculated using the automatic method, which created a relationship between nearest line nodes between the two data sources. The km and number of central nodes for each line were calculated. These results were grouped by location, where the number of elements and km in each of the locations were compared to determine the absence or excess of elements in the OpenStreetMap grid with respect to the reference source. Absolute positional accuracy, which refers to the accuracy of the position of an element with respect to another of higher precision, was calculated using Equation 1 (*Eq. 1*). Coordinates were calculated for each of the central nodes of the OpenStreetMap and IDECA road grid, then the differences were measured for records in the horizontal (x, y) component:

$$e_i = \sqrt{(X_{if} - X_{ir})^2 + (Y_{if} - Y_{ir})^2} \qquad \text{(Eq. 1)}$$

where e_i represented the i -th horizontal error (i =1, ..., n) at each point, (X_{if}, Y_{if}) are considered the coordinates of the source to be corroborated, in this case OpenStreetMap, while, $(X_{ir}, Y_{ir})$are the coordinates considered as reference or true (IDECA). For e_i, the RMSE for all records was then obtained. Finally, the positional accuracy was calculated according to the National Standard for Spatial Data Accuracy (NSSDA) and observing that the RMSEs in the x component is considered equal to those found in the y (Greenwalt and Schultz, 1962).

The assessment of thematic accuracy was carried out in a binary way. If the data standardized in OpenStreetMap did not exactly match the value encoded in IDECA then the box took the value of 1, otherwise it took the value of 0. At the end of this comparison a number of items of each object

incorrectly classified by locality was obtained for each of the attributes compared, which were direction of flow, road classification (road hierarchy) and name of the main road. To establish the measure of accuracy (% error) in the classification, the number of errors found was taken divided by the total number of items (multiplied by 100), and an acceptance criterion of up to 5% of misclassified items.

3. Results

This section will show the results obtained by measuring the quality of OpenStreetMap data using the three quality measures mentioned: Completeness, Positional Accuracy and Thematic Accuracy. The results are shown in the following order: I) Data Extraction and Exploration, II) Data Standardization, III) Automatic Data Comparison, IV) Completeness and Positional Accuracy Assessment and V) Thematic Accuracy Assessment.

3.1. Data Extraction and Exploration

Regarding the preliminary exploration of the data on the following elements: Number of records, Principal road name, Direction of flow and Road hierarchy it was found that OpenStreetMap had 73,454 records while IDECA 136,958 elements. This difference depended largely on the fact that OpenStreetMap did not have a division by road junctions. On the other hand, it was found that 28.3% of the OpenStreetMap records had road nomenclature assigned to them, while only 5.3% of the IDECA data were in the same condition. Regarding the direction of road, it was found that IDECA had five classes defined for this attribute: two-way, From the direction of digitization - no direction of road, Undefined, Towards the direction of digitization. Where 43.4% of the data referred to road direction "Both*"* and 27.0% was without any definition for the direction of flow. Regarding OpenStreetMap the direction of flow was defined under three categories where 79.4% of the data were recorded with direction "Both" (Fig. 3).

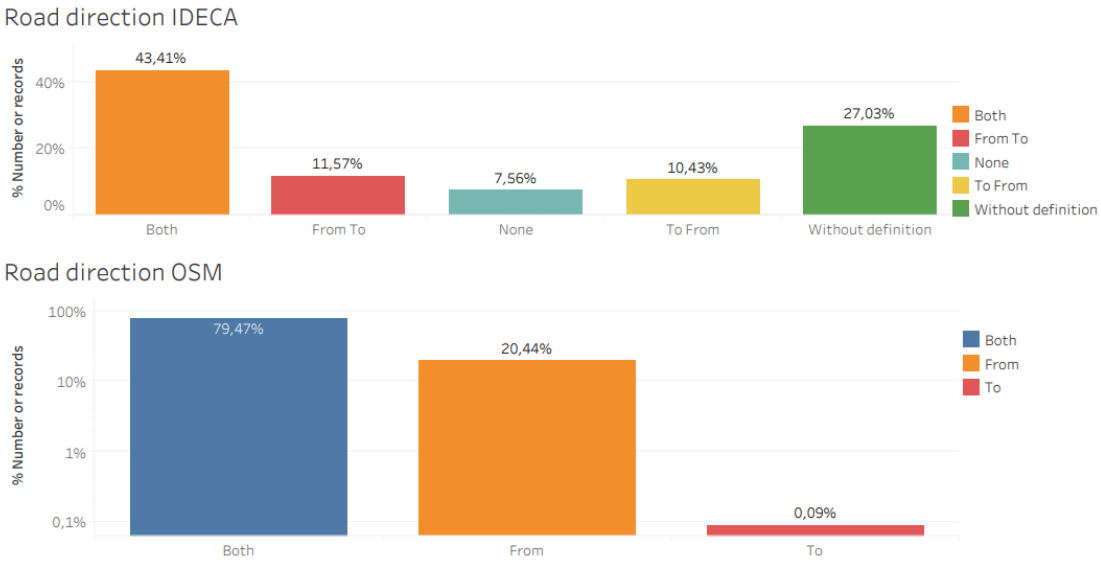

Figure 3. Percentage for each category in road direction

Regarding the Road Hierarchy (Element that categorises the road network according to its arterial functions) it was found that OpenStreetMap had a classification of 26 categories, where the residential type predominated with 36,980 records. This was followed by the *hidhway* category with 8,398 records. On the other hand, the IDECA Road Hierarchy had only seven categories. This eventually

resulted in the need to create standard categories for those attributes that could belong to more than one class. The last step to finalize the data exploration was to review and analyze the OpenStreetMap mapping guide. For the *highway* attribute with the subclassification *motorway*, it is not coded in Colombia, since OpenStreetMap established that these roads do not exist in the country. This is indeed correct as we do not have controlled access roads, so any coding found with these values was counted as an error. The following OpenStreetMap attributes were excluded from the study because they did not have a matching element in the IDECA data: ('*bridleway*', '*cycleway*', '*footway*', '*pedestrian*', '*path*', '*steps*')"), with 23,841 records removed and 116,987 data. Based on the (E-R) model and the data exploration carried out, the following model structure was obtained (Table 4).

Table 4. Result of applying the Entity-Relationship model (E-R).

OpenStreetMap Road Hierarchy	IDECA	Relation
Primary	Arterial road network	Many to one
Primary link		
Trunck		
Secundary	Intermediate road network	Many to one
Secundary Link		
Tertiary		
Tertiary link		
Trunck link		
Residential	Local road network	Many to one
Living street		
Track 1…5	Rural road network	
Unclasified	Not defined	Many to one
Unknow		
Services		
Sentido vial OpenStreetMap	IDECA	Relationship
Both	Both	One to Many
	None	
	without definition	
	Null	
From	From to	One by one
To	To from	
Road name OpenStreetMap	IDECA	Relationship
Name	name	One by one

3.2. Data Standardization

The fields were created to host the standardization results and the fields that would contain the comparison results in the fields created for the standardized data were: *New_Jerc*: New OpenStreetMap road hierarchy from 26 categories to 7. *Road_Dir_IDE*: Standardization of road direction in IDECA: 5 categories to 3 categories. *New_Jerc*: Road hierarchy: Domain change, now the road hierarchy is governed by numbers 1-7, and in the Fields created to store the results of the comparison: *NEAR_DIST*: Field to calculate the distances (meters) to the centroids of each object (Completeness). Exac_po: Field to hold the positional accuracy calculation. *Ch_New_Jer*: Boolean field to classify the comparison referring to Road Hierarchy. *Ch_Sent_F_*: Boolean field to classify the comparison referring to Road Direction *Fc_Name_*: Boolean field to classify the comparison referring to Main Road Name. Table 5 shows an example of name standardization on the main road, performed on OpenStreetMap with the help of regular expressions.

Just over 85,125 records were standardized. There were character strings that could not be standardized because no associated pattern was found. Some expressions that had to be corrected manually were those that started with a number, but had no associated nomenclature such as street or race or character strings longer than 20 records.

Table 5. Standardization of names.

Original Non-standardized road names	**Standard** Standardized name
Transversal 9 Bis Este	TV 9BISE
Transversal 9 Bis Este	TV 9BISE
Transversal 9B Este	TV 9B E
Variante Bosa San José - Travsnersal 80I	TV 80I
Avenida Agoberto Mejía - Transversal 80G	TV 80G
Calle 32 - Callw32	CL 32
Transversal 7 Bis Este	TV 7BISE
Avenida Tintal - Carrera 89	KR 89
Carrera 78H Bis Sur	KR 78HBIS S
Carrera 53F	KR 53F
Calera 15	KR 15
Calle 45/sur	CL 45 S
TransMilenio - Patio Calle 80	CL 80
TransMilenio - Intercambiador CL 6	CL 6
CL-57B	CL 57B

3.3. Automatic Data Comparison

Regarding the automatic comparison of the data to obtain the quality measures: positional accuracy, completeness and thematic accuracy, by means of nearest object *matching*, which consisted of a mobile buffer that depended on the distance of the nearest node, 114,183 records were matched, which meant a match of 85.42%. As can be seen in Figure 4, many nodes were not found close to their counterparts, however, the match was always successful depending on the simplicity of the geometry.

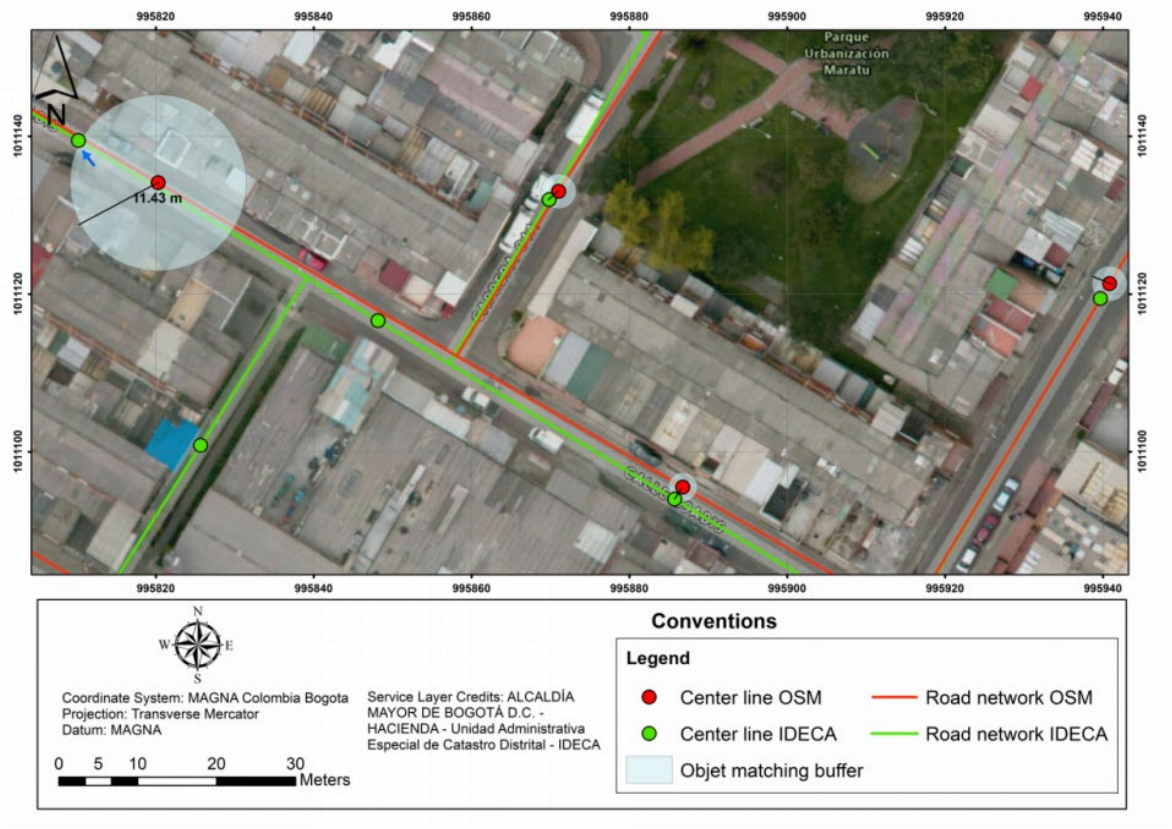

Figure 4. Mobile Buffer

14.58% of the data could not be joined due to problems with geometry, specifically on those complex elements such as depressions, *roundabouts* and in general terms on roads with complex geometry (Fig. 5).

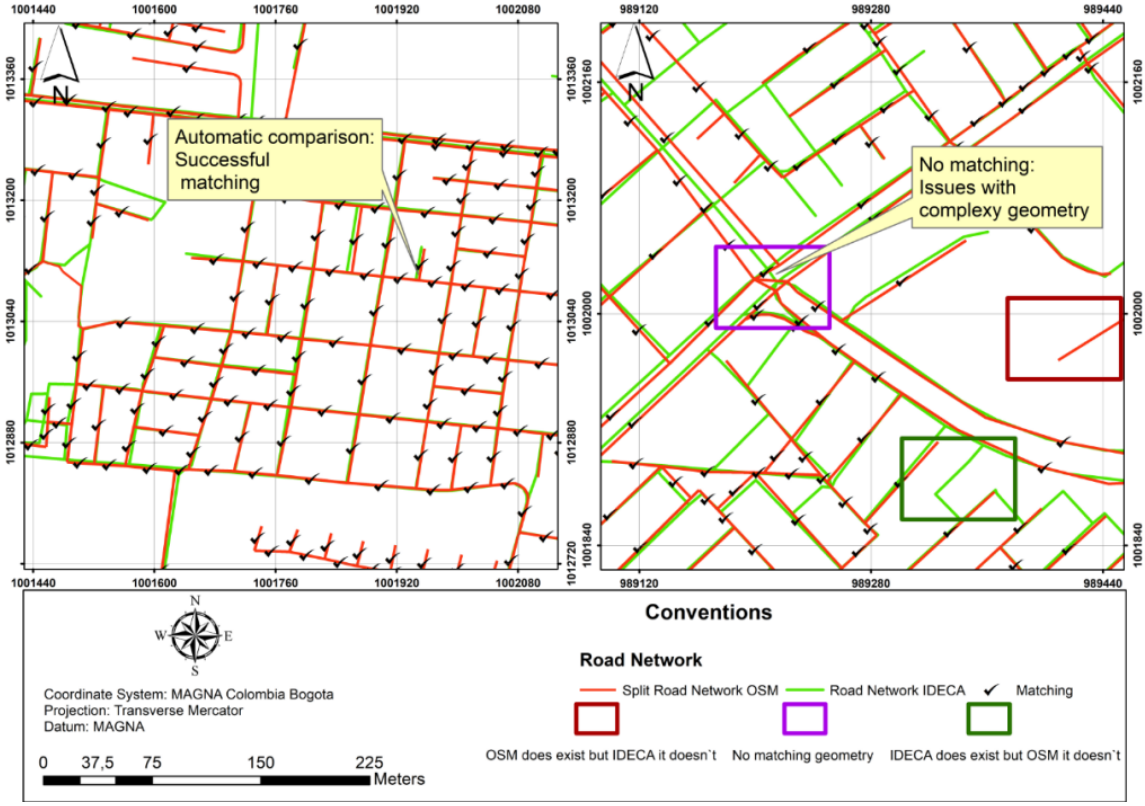

Figure 5. Mach between geometries

3.4. Completeness and Positional Accuracy Assessment

Completeness was assessed in general terms, over the entire city of Bogotá and particularly for each of the localities that make up the city. As can be seen in Table 6, it was found that OpenStreetMap has omitted 1,183.3 km, which is 12.6% less than the km reported by IDECA. On the other hand, the count derived from the automatic comparison showed that OpenStreetMap contains -14.6% fewer linear elements than those found in IDECA.

Table 6. Completeness results Bogotá level.

Variables	IDECA	OpenStreetMap	Delta	% Omission
km	9,379.21	8,195.91	1,183.29	12.60%
Line elements	1,369.58	1,169.64	199.94	14.60%

An excess in commission value of 36.42 km corresponding to 1,266 linear objects was found in OpenStreetMap. In percentage terms this corresponded to 0.44 and 1.11% respectively. On the other hand, evaluating the completeness results by locality, it was found that the localities in blue color contain the highest values in terms of missing segments in the OpenStreetMap road network (3,205-5,978). On the other hand, the localities (flattened) in the left image of Figure 6 refer to the presence of excess road segments (776). Finally, the image on the right refers to the number of km for each locality, where it can be seen that the rural area located to the south of the city has the highest number of km quantified as excess (-19.0 km). The localities in red contain the largest number of missing km (252-306). The

result of the summation of the mean squared errors found was: RMSr= 2.252. The extreme values found when evaluating the distance between related nodes were a minimum value found: 0.008162 m and a maximum value found: 5.000998 m. The factor used to compute the positional accuracy with a 95% confidence interval was 1.7308, considering that the RMSx, RMSy errors were equal.

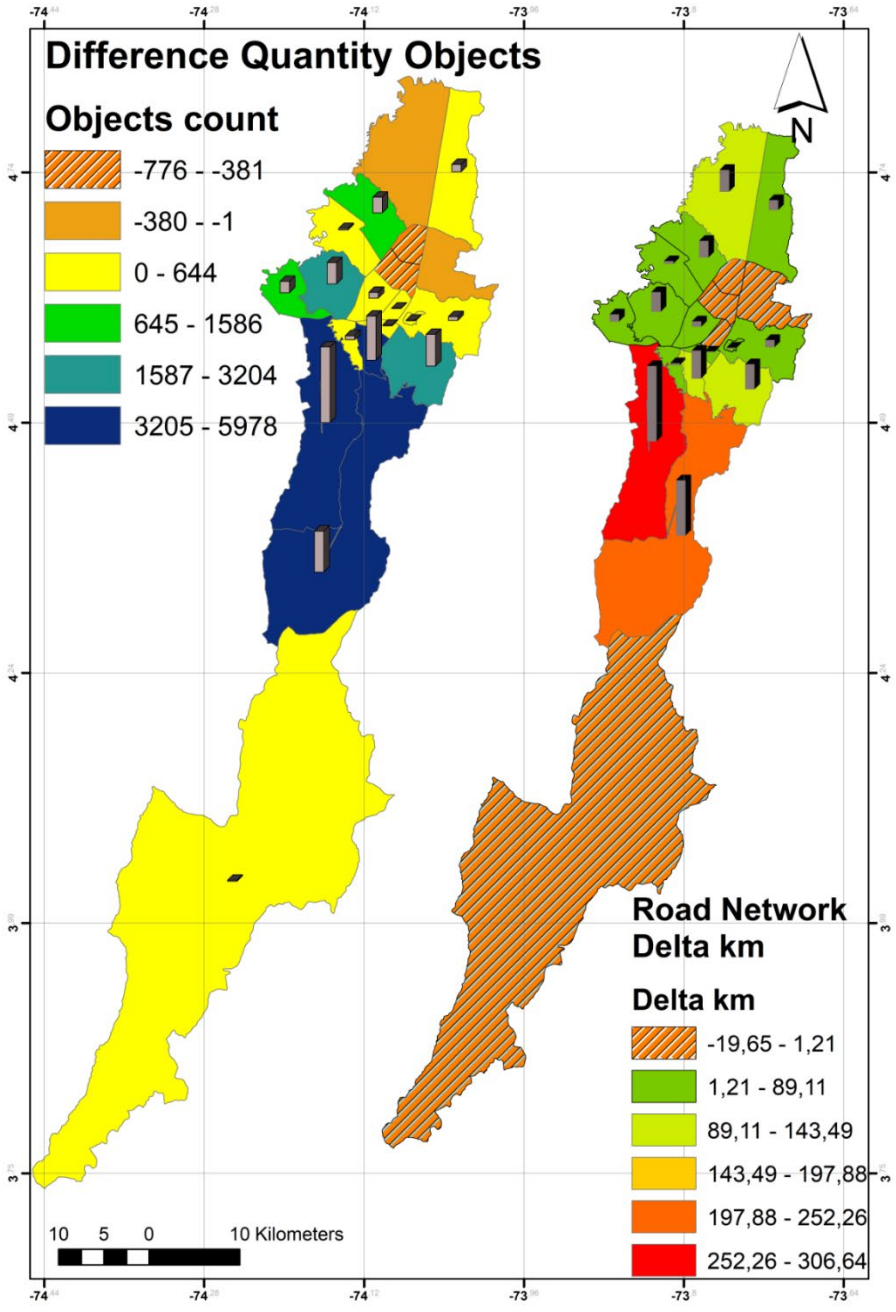

Figure 6. Presence of excess road segments

3.5. Thematic Accuracy Assessment

The comparison of attributes for Road Hierarchy found that all error values were found to be above 20%. The localities with the most errors in road classification were (Table 7): Sumapaz with 98.7% of incorrectly classified elements, followed by La Candelaria with 52.9%, and Ciudad Bolivar with 48.2%. In contrast and as can be observed, localities such as Bosa and Kennedy showed the lowest values in terms of error counts. This is due to the fact that these localities have high values for residential

roads, a category that is less affected in terms of the number of errors found. In general, it was observed that 35.8% of the evaluated data were misclassified with respect to Road Hierarchy.

Table 7. Thematic accuracy Road hierarchy.

LOCATION	Number of elements assessed	Number of errors found	% error
Antonio Nariño	1,937	662	34.2
Tunjuelito	2,263	700	30.9
Rafael Uribe Uribe	4,989	1,625	32.6
Candelaria	561	297	52.9
Barrios Unidos	4,273	1,650	38.6
Teusaquillo	4,729	1,978	41.8
Puente Aranda	5,473	1,586	29.0
Los Martires	2,358	962	40.8
Sumapaz	157	155	98.7
Usaquen	7,178	2,427	33.8
Chapinero	3,831	1,713	44.7
Santa Fe	2,685	1,059	394
San Cristobal	5,980	2,112	35.3
Usme	6,557	3,654	55.7
Ciudad Bolivar	9,983	4,811	48.2
Bosa	8,276	1,642	19.8
Kennedy	10,951	3,016	21.9
Fontibon	5,487	2,371	43.2
Engativa	10,783	3,164	29.3
Suba	15,709	5,262	33.5

The comparison of attributes for road sense found that there is a spatial distribution of error, with most of the errors located in the central part of the city, the first range found for the percentage of errors showed that between 0 and 10.0% of the errors are located in the south of the city. In orange and yellow we find 10.0-27.0% and finally in red, we find the errors above 27.0%. On this attribute, it was found that the range of error was between 22.9 and 100%. The localities with the most errors were located in the south of the city, among them Sumapaz with 100% of the erroneous naming and Usme with 71.8%. On the other hand, the lowest percentages in relation to the number of errors found were located in the localities of Bosa 32.1%, Kennedy 24.9% and Engativá with 25.5%. Overall, 34.6% of all roads were found to have a misclassification percentage.

It is important to emphasize that the results obtained associated with the quality of the VGI in the current investigation for the selected region are inherent to the locality studied and cannot be extended to other regions, much less question the important role played by the systems where this information is recorded, even if the same methodology is used, since the spatial bias may be intrinsic to each region (Zhang and Zhu, 2018).

4. Discussion

The analysis carried out for the evaluation of the VGI quality through automation of the process and using the measures of *completeness*, *positional accuracy*, and *thematic accuracy* evaluated the VGI quality for the Bogotá road network and allowed us to conclude that the VGI quality in terms of synergies is deficient with respect to the minimum quality levels established in the IDECA data model. However, the evaluation for each of the attributes shows varied quality. Well acceptable quality was noted in terms of completeness and thematic accuracy. In summary, the OpenStreetMap road mesh has the following quality:

Completeness: The results found regarding the omission of data were below 13.0% while the data for excess were only approximately 1.0%, showing that, although the quality in terms of completeness is not the best, if they are within an acceptable range. OSM data is an open source product that it is subjected to quality processes (Quality assurance) (OSM, 2017c) and although there are still gaps in terms of completeness data, we finding low omission and commission values, this implies that the VGI data is not so far from ideal quality metrics. This contrasts with what was concluded by Ziliestra and Zipf (2010) where at the time it was concluded that there was still strong heterogeneity in the OpenStreetMap data according to its completeness may mean that the OpenStreetMap has been improving considerably.

Positional accuracy: The obtained calculations showed that OpenStreetMap has a horizontal positional accuracy error of 3.98 m. Compared to the 2 m of accuracy reported by IDECA, it can be concluded that the positional accuracy of OpenStreetMap is good. However, it should be noted that the value found of 3.98 m refers to an average; therefore, the marginal values by locality, where errors were found in the horizontal position of 95% that oscillate between 4 and 6 m, showed that a sectional evaluation allowed a better observation of the distribution of errors. This variation may be due to the digitization process, data collection or even the coders' skills (Haklay, 2010).

Thematic accuracy: The results regarding the evaluation of thematic accuracy were deficient. This showed in general terms that the attribute Road hierarchy had a percentage of 35.8% of misclassified data. The evaluation of the road sense attribute showed that 15% of the evaluated data were found to be misclassified. This value was affected by the method used to perform the matching between the elements. The IDECA data contains an additional node for each coded flow direction. This greatly affected the calculations for counting the number of errors in the direction of the road. Finally, the thematic accuracy for the attribute name of road showed that, in general terms, 34.6% of all roads had a wrong percentage classification. For this reason, it was determined that taking into account the minimum quality level that established that only 5% of the data might be misclassified; thus, the quality of thematic accuracy for the Bogotá road network was considered of poor quality.

The standardization process was very important in order to arrive at these results. This work allowed the development of an entity-relationship model where the comparison of attributes was possible, which led, as in the work carried out by Da Costa (2016b), to guarantee unambiguity in the classification. However, after analyzing the results, it was concluded that the services category should have been analyzed separately in order to obtain better results in terms of road hierarchy comparison.

As far as the standardization of names is concerned, the knowledge of the data structure was essential, because the comparison of characters is very sensitive. The advantage found when applying standardization by means of regular expressions allowed a standardization of road names with more certainty at the cost of higher processing times (Zandbergen, 2009; Nelli, 2015; Gao *et al.*, 2017).

Regarding the creation of the automatic process to compare sources and calculate quality measures, this work allowed the creation of a series of scripts to download, cut, compare and calculate attributes in the road segments, which undoubtedly speeded up the process with respect to a semi-automatic development, as implemented in the proposals made by Goodchild and Li (2012). However, it should be noted that the process implemented in Python is based on the node matching technique (Koukoletsos *et al.*, 2012) with the difference that here no attributes were used to refine the *match*, only the nodes and the minimum distance for the creation of a mobile buffer. Although in terms of processing this method consumed about 3 hours to calculate the *match* in geometry, compare each of the attributes and generate the calculations of the 3 quality averages, there are methods that spend about 5 hours only comparing geometry to determine the completeness as is the case of *Segment-based algorithm* (Abdolmajidi *et al.*, 2015).

The study by Ali and Schmid (2014) who used a classifier using machine learning techniques and where cities in Germany, United Kingdom and Austria were involved showed how for the case of

the first two countries a classification accuracy of 70% to 90% was found for parks and gardens. The classifiers generated showed how between 10% and 30% of all entities analyzed in each city could be incorrectly classified. Poorer results were obtained in Austria, which could be attributed to the relatively low number of entities in the available dataset or to already existing classification problems.

As the results published in many countries reflect values inherent to each region, the statistical or data science methodology used to measure these errors may also be important at the time of generating the metrics, so it is recommended to continue exploring methods to standardize criteria for measuring and judging the quality of VGI (Elias *et al.*, 2018).

On the other hand, some limitations of this development are: I) problems with geometries that do not have a considerable range of separation, as is the case of mass transit road segments II) complex geometries such as at-grade crossings with *roundabouts*, in which *matching* becomes unpredictable. III) Scalability, as the code is not the subject of this work, it is not scalable, IV) the code is highly dependent on the attributes compared, in this case, Road hierarchy, Road direction and road naming, so it only works for them. All these limitations will be the subject of a later study. This is one of the first works carried out on the evaluation of VGI quality for Colombian data, therefore, it is a small effort to begin to know the current state of VGI and its possible applications, given that the lack of knowledge of the quality of a product impedes its usability.

The IDECA and OSM data are collected with different resources and at different scales, so the results presented may be conditioned by this. Another aspect to be considered is that some aspects of how the data are stored within the OSM database may have been overlooked, due to the use of a modified version of the raw data.

5. Conclusions

The VGI quality assessed through completeness, positional accuracy, thematic accuracy is jointly considered as poor, with respect to the minimum quality levels established in the IDECA data model. However, assessed separately through the indicated averages, the VGI data was found to have acceptable completeness, optimal positional accuracy and poor thematic accuracy.

The assessment of the completeness of the data resulted in omission values below 13.0%, as well as the commission was found to be approximately 1.0%. The calculation of positional accuracy resulted in a 95% horizontal error of 3.98 meters. Therefore, a good quality was concluded for the position on the OpenStreetMap road grid.

Thematic accuracy was the worst performing attribute, showing for the road hierarchy a misclassification error rate of 35.8%. For the road sense attribute, misclassification errors of 15.0% were found and finally naming had a misclassification error rate of 34.6%.

The automatic process used here, allowed to automatically compare the sources and generate the calculations to measure the VGI quality, however, as it is not the full object of this research, algorithmic improvements, such as match on complex geometries and scalability of the tool, are still to be developed. The evaluation by positional accuracy allows us to affirm that the VGI data corresponding to the road network of Bogotá can be used as reference data for official entities.

Data Availability Statement (DAS)

The data that support the findings of this study are available from the corresponding author, Aquiles Enrique Darghan Contreras (aqedarghanco@unal.edu.co), upon reasonable request.

References

Abdolmajidi, E., Mansourian, A., Will, J., Harrie, L. 2015. Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-Spatial Information Science,* 18(2-3), 65-80. https://doi.org/10.1080/10095020.2015.1071065

Ali, A.L., Schmid, F. 2014. Data quality assurance for Volunteered Geographic Information. In: *Proceedings of the 8hth International Conference on Geographic Information Science,* Springer International Publishing Switzerland, pp. 126-141.

Antoniou, V. 2011. *User generated spatial content: an analysis of the phenomenon and its challenges for mapping agencies.* Doctoral thesis. University College London. Retrieved from: http://discovery.ucl.ac.uk/1318053/

Antoniou, V., Skopeliti, A. 2015. Measures and indicators of VGI Quality: An Overview. I*SPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5, 345-351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Arsanjani, J., Mooney, P., Zipf, A., Schauss, A. 2015. Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In *OpenStreetMap in GIScience,* pp. 37-58, Springer, Cham. https://doi.org/10.1007/978-3-319-14280-7_3

Ballatore, A., Zipf, A. 2015. A conceptual quality framework for Volunteered Geographic Information. In *International Conference on Spatial Information Theory*, pp. 89-107. Springer, Cham. https://doi.org/10.1007/978-3-319-23374-1_5

Bazeley, P., Jackson, K. 2013. *Qualitative data analysis with NVivo.* Sage Publications Limited, London, Thousand Oaks and New Delhi.

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., Rau, R. 2011. *Osmonto-an ontology of openstreetmap tags*. State of the map Europe (SOTM-EU), 23-24.

Cristancho, C., Triana, E. 2018. *Análisis demográfico y proyecciones poblacionales de Bogotá.* Alcaldía Mayor de Bogotá, DC, Secretaría Distrital de Planeación, Alcaldía Mayor de Bogotá, Bogotá.

Da Costa, J.N. 2016a. Novel tool for examination of data completeness based on a comparative study of VGI data and official building datasets. *Geodetski vestnik* 60(3), 495-508. https://doi.org/10.15292/geodetski-vestnik.2016.03.495-508

Da Costa, J.N. 2016b. Towards building data semantic similarity analysis: OpenStreetMap and the Polish Database of Topographic Objects. In *2016 Baltic geodetic congress* (BGC Geomatics), pp. 269-275. IEEE. https://doi.org/10.1109/BGC.Geomatics.2016.55

Darwish, A., Lakhtaria, K.I. 2011. The impact of the new Web 2.0 technologies in communication, development, and revolutions of societies. *Journal of Advances in Information Technology* 2(4), 204-216. https://doi.org/10.4304/jait.2.4.204-216

Dorn, H., Törnros, T., Zipf, A. 2015. Quality evaluation of VGI using authoritative data-A comparison with land use data in Southern Germany. *ISPRS International Journal of Geo-Information* 4(3), 1657-1671. https://doi.org/10.3390/ijgi4031657

Elias, E., Fernandes, V., Junior, M. 2018. Positional Accuracy Assessment of the VGI Data from OpenStreetMap - Case Study: Federal University of Bahia Campus in Brazil. In *Proceedings of the 4th International Conference on Geographical Information Systems Theory, Applications and Management - GISTAM,* pp. 231-238. https://doi.org/10.5220/0006707702310238

Elwood, S., Goodchild, M. F., Sui, D.Z. 2012. Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers* 102(3), 571-590. https://doi.org/10.1080/00045608.2011.595657

Esmaili, R., Naseri, F., Esmaili, A. 2013. Quality Assessment of Volunteered Geographic Information. *American Journal of Geographic Information System* 2(2), 19-26. https://doi.org/10.5923/j.ajgis.20130202.01

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. https://doi.org/10.1080/13658816.2015.1018266

Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D.S. 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. *Transactions in GIS* 17(6), 847-860. https://doi.org/10.1111/tgis.12033

Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y. 2017. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems* 61, 172-186. https://doi.org/10.1016/j.compenvurbsys.2014.02.004

Geofabrik, 2018. *OpenStreetMap data downloads.* Retrieved from http://download.geofabrik.de/ (accessed 10.05.2018).

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

Goodchild, M.F., Li, L. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics* 1, 110-120. https://doi.org/10.1016/j.spasta.2012.03.002

Graser, A., Straub, M., Dragaschnig, M. 2014. Towards an open source analysis toolbox for street network comparison: Indicators, tools and results of a comparison of OSM and the official Austrian reference graph. *Transactions in GIS* 18(4), 510-526. https://doi.org/10.1111/tgis.12061

Greenwalt, C.R., Schultz, M. 1962. Principles of Error Theory and Cartographic Applications, Aeronautical Chart and Information Center. *Technical Report* No. 96.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: *Planning and Design* 37(4), 682-703. https://doi.org/10.1068/b35097

Haklay, M., Weber, P. 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing* 7(4), 12-18. https://doi.org/10.1109/MPRV.2008.80

Hudson-Smith, A., Batty, M., Crooks, A., Milton, R. 2009. Mapping for the Masses Accessing Web 2.0 Through Crowdsourcing. *Social Science Computer Review* 27(4), 524-538. https://doi.org/10.1177/0894439309332299

IDECA (Infraestructura de Datos Espaciales para el Distrito Capital), 2017. *Malla Vial Integral. Bogotá D. C.* UAECD (Unidad Administrativa Especial de Catastro Distrital). Online consultations: http://www.ideca.gov.co/recursos/mapas/malla-vial-integral-bogota-dc

Ipeirotis, P.G., Provost, F., Sheng, V.S., Wang, J. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 28(2), 402-441. https://doi.org/10.1007/s10618-013-0306-1

ISO 19157, 2013 *Geographic information-data quality.* ISO (International Organization of Standardization).

Janelle, D.G., Goodchild, M.F. 2011. Concepts, principles, tools, and challenges in spatially integrated social science. T*he SAGE Handbook of GIS and Society* 27-45. https://doi.org/10.4135/9781446201046.n2

Jonietz, D., Zipf, A. 2016. Defining fitness-for-use for crowdsourced points of interest (POI). *ISPRS International Journal of Geo-Information* 5(9), 149. https://doi.org/10.3390/ijgi5090149

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

Kresse, W. 2012. Springer handbook of geographic information (pp. 118-120). D. M. Danko (Ed.). Springer, Science & Business Media, Berlin. https://doi.org/10.1007/978-3-540-72680-7

Lin, W. 2018. Volunteered Geographic Information constructions in a contested terrain: A case of OpenStreetMap in China. *Geoforum* 89, 73-82. https://doi.org/10.1016/j.geoforum.2018.01.005

Ludwig, I., Voss, A., Krause-Traudes, M. 2011. A Comparison of the Street Networks of Navteq and OSM in Germany. In *Advancing geoinformation science for a changing world.* Springer, Berlin, pp. 65-84. https://doi.org/10.1007/978-3-642-19789-5_4

Mahabir, R., Stefanidis, A., Croitoru, A., Crooks, A.T., Agouris, P. 2017. Authoritative and volunteered geographical information in a developing country: A comparative case study of road datasets in Nairobi, Kenya. *ISPRS International Journal of Geo-Information* 6(1), 24. https://doi.org/10.3390/ijgi6010024

Mobasheri, A., Zipf, A., Francis, L. 2018. OpenStreetMap data quality enrichment through awareness raising and collective action tools-experiences from a European project. *Geo-spatial Information Science* 21(3), 234-246. https://doi.org/10.1080/10095020.2018.1493817

Moreri, K.K., Fairbairn, D., James, P. 2018. Volunteered geographic information quality assessment using trust and reputation modelling in land administration systems in developing countries. *International Journal of Geographical Information Science* 32(5), 931-959. https://doi.org/10.1080/13658816.2017.1409353

Nelli, F. 2015. An introduction to data analysis. In *Python Data Analytics*. Apress, Berkeley, pp. 1-12.

OpenStreetMap Contributors, 2017a. *Zoom Levels.* Retrieved September 12, 2017, from https://wiki.openstreetmap.org/wiki/Zoom_levels

OpenStreetMap Contributors, 2017b. *Completeness.* Retrieved September 25, 2017, from https://wiki.openstreetmap.org/wiki/Completeness

OpenStreetMap Contributors, 2017c. *Quality Assurance.* Retrieved September 25, 2017, from https://wiki.openstreetmap.org/wiki/Quality_assurance

Sharma, P., Nagpal, B. 2020. Regex: an experimental approach for searching in cyber forensic. *International Journal of Information Technology* 12(2), 339-343. https://doi.org/10.1007/s41870-019-00401-y

Silberschatz, A., Korth, H.F., Sudarshan, S., Pérez, F. S., Santiago, A. I., Sánchez, A. V. 2002. *Fundamentos de bases de datos (Vol. 11).* McGraw-Hill. Ciudad de México, México.

Stein, A., Shi, W., Bijker, W. 2016. *Quality aspects in spatial data mining.* CRC Press.

Thalheim, B. 2013. *Entity-relationship modeling: foundations of database technology.* Springer Science & Business Media.

Van Oort, P. 2006. *Spatial data quality: from description to application.* Wageningen University and Research Netherlands.

Vannoni, M., McKee, M., Semenza, J.C., Bonell, C., Stuckler, D. 2020. Using volunteered geographic information to assess mobility in the early phases of the COVID-19 pandemic: a cross-city time series analysis of 41 cities in 22 countries from March 2nd to 26th 2020. *Globalization and Health* 16(1), 1-9. https://doi.org/10.1186/s12992-020-00598-9

Wu, H., Lin, A., Clarke, K.C., Shi, W., Cardenas-Tristan, A., Tu, Z. 2021. A comprehensive quality assessment framework for linear features from Volunteered Geographic Information. *International Journal of Geographical Information Science* 35(9), 1826-1847. https://doi.org/10.1080/13658816.2020.1832228

Yan, Y., Feng, C.C., Wang, Y.C. 2017. Utilizing fuzzy set theory to assure the quality of volunteered geographic information. *GeoJournal* 82(3), 517-532. https://doi.org/10.1007/s10708-016-9699-x

Zandbergen, P.A. 2009. Geocoding quality and implications for spatial analysis. *Geography Compass* 3(2), 647-680. https://doi.org/10.1111/j.1749-8198.2008.00205.x

Zhang, G., Zhu, A.X. 2018. The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS* 24(3), 151-162. https://doi.org/10.1080/19475683.2018.1501607

Zhang, W.B., Leung, Y., Ma, J.H. 2019. Analysis of positional uncertainty of road networks in volunteered geographic information with a statistically defined buffer-zone method. International *Journal of Geographical Information Science* 33(9), 1807-1828. https://doi.org/10.1080/13658816.2019.1606430

Zielstra, D., Zipf, A. 2010. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE International Conference on Geographic Information Science*, pp. 1-15.