DOI 10.1590/s1982-21702024000100007

ORIGINAL ARTICLE

Proposal of a method for evaluating the spatial distribution pattern of linear features

Marconi Martins Cunha[1] - ORCID: 0000-0003-0797-8530
Afonso de Paula dos Santos[2] - ORCID: 0000-0001-7248-4524
Marcelo Antonio Nero[1] - ORCID: 0000-0003-2124-5018
Nilcilene das Graças Medeiros[1] - ORCID: 0000-0003-0839-3729

[1]Universidade Federal de Viçosa, Departamento de Engenharia Civil, Viçosa - MG, Brasil.
E-mail: marconi.cunha@ufv.br, afonso.santos@ufv.br, nilcilene.medeiros@ufv.br

[2]Universidade Federal de Minas Gerais, Departamento de Cartografia, Belo Horizonte - MG, Brasil.
E-mail: marcelo.nero@gmail.com

Received in 11th November 2023.
Accepted in 19th March 2024.

Abstract:

Positional accuracy of cartographic products is typically evaluated using positional discrepancies and point-based techniques. However, using linear features has some advantages over the point-based method, such as a greater amount of geometric and positional information and the fact that approximately 80% of the features on a cartographic basis are lines. Despite these advantages, important parameters for evaluating accuracy using lines have not yet been established or determined, such as the spatial distribution pattern, although it is a relevant factor that can affect the results and determine the validity of an evaluation process. This study proposes a method based on the modification of the Nearest Neighbor Method for points, which can be used to evaluate the spatial distribution pattern of linear features. Instead of the traditional Euclidean distance used by the method for points, the method proposes using the Hausdorff Distance as a measure of the spacing between lines. The proposed method, called Nearest Neighbor Method for Linear Features (NNMLF), was applied to simulated and real data. All experiments with simulated data showed that the NNMLF was effective in estimating spatial distribution pattern up to the third order. Its use on real data showed NNMLF is simple to apply.

Keywords: Cartographic Quality Control; Linear Features; Hausdorff Distance; Nearest Neighbor Method for Linear Features.

How to cite this article: CUNHA MM, SANTOS AP, NERO MA, MEDEIROS NG. Proposal of a method for evaluating the spatial distribution pattern of linear features. *Bulletin of Geodetic Sciences*. 30: e2024007, 2024.

1. Introduction

The increasing availability of spatial information to users also increases the need to assess the quality of data and products generated, to identify whether they meet user requirements (Xavier et al. 2019). Cartographic Quality Control (CQC) evaluates the quality of a cartographic product, in order to increase its reliability and enable its use in an adequate way to meet the real needs of the user. This assessment is carried out on the quality elements of the product. ISO 19157 standard (ISO 2013) defines that the basic quality elements to be observed and evaluated in cartographic products are: completeness, logical consistency, positional accuracy, temporal accuracy, thematic accuracy and usability. It should be noted that positional accuracy is the most used element to verify spatial data quality (Jakobsson and Vauglin 2002; Drobnjak et al. 2017).

Traditionally, positional accuracy assessment is performed using point-based techniques (FGDC 1998; Van Niel and Mcvicar 2001; Ariza-López et al. 2012; Nero et al. 2017; Ruiz-Lendínez et al. 2017; Ariza-López et al. 2018a; Mozas-Calvache 2021; Santo Filho et al. 2022). In these techniques, there is a statistical comparison of the positional discrepancies observed among well-identified homologous points of test and reference products.

However, the technological advances of GNSS (Global Navigation Satellite System) receivers and the development of kinematic survey methods boosted the possibility of acquiring and using lines in positional control (Mozas-Calvache 2007, 2021). The use of lines in CQC has some advantages over points. Cuenin (1972), Thapa (1988) and Li (2006) state that approximately 80% of the entities in a cartographic base are made up of linear features. Mozas-Calvache and Ariza-López (2011) point out that linear elements have varied geometric and positional information, such as vertices, angles, length of line segments, orientation, sinuosity, among others. This additional information can help and provide new possibilities in the positional accuracy evaluation process. In view of this, Mozas-Calvache and Ariza-López (2010) stated that it is possible that positional quality control in cartography using linear features represents an evolution of methods based on point features. This idea is also seen in studies by Seo and O'Hara (2009) and Wu et al. (2021).

Despite these advantages, important parameters in the evaluation of positional accuracy using linear features have not yet been established or determined, such as the sample size and its distribution pattern (Ariza-López et al. 2011; Ariza-López et al. 2018b).

Spatial distribution of a sample, subject of this work, can determine the validity and quality of a sampling evaluation process, since a bad spatial distribution affects the representativeness of the sample, reducing the adequate demonstration of the population, which can result in a wrong estimate (Ariza-López and Atkinson-Gordo 2008).

When it comes to spatial distribution of the point-type control elements, there are metrics, statistics or methods that evaluate their distribution, such as the Nearest Neighbor Method, presented by Clark and Evans (1954), and Ripley's K Function, introduced by Ripley (1977). In Cartographic Quality Control, Nearest Neighbor Method was applied at the studies of Santos et al. (2016) and Silva (2020); while Ripley's K Function was used in research by Zanetti (2017) and Oliveira et al. (2018). Standards from some countries also provide guidelines for the spatial distribution of checkpoints, such as the National Standard for Spatial Data Accuracy (NSSDA) (FGDC 1998), from the United States of America (USA), and UNE 148002:2016 – Control of la Calidad Positional en Conjuntos de Datos Espaciales (AENOR 2016), from Spain.

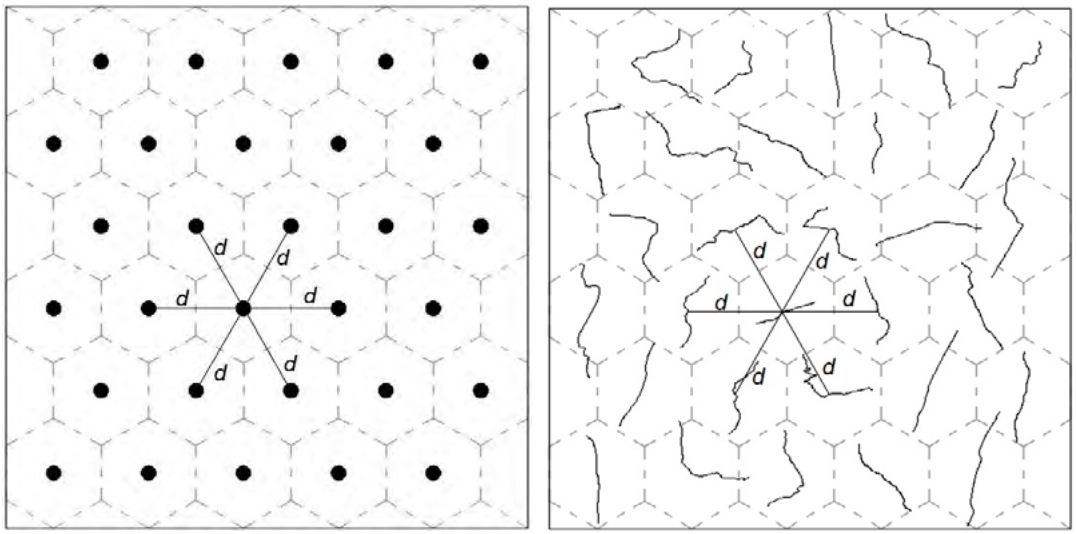
On the other hand, the assessment of positional accuracy through distribution pattern of linear features seems to have been little or not explored at all in some aspects. An analysis in the literature shows us, for example, that there is no method to evaluate the spatial distribution for linear features. Therefore, the aim of this paper is to propose a method for evaluating the spatial distribution pattern of linear features used in the evaluation of positional accuracy.

2. The proposed method

The proposed method was named Nearest Neighbor Method for Linear Features (NNMLF). As the name suggests, this method is based on a modification of the Nearest Neighbor Method for points.

The Nearest Neighbor Method compares the average between an observed distance and an expected distance in a theoretical set of point features that have a random distribution pattern. This theoretical set is obtained by partitioning the study area into hexagons of the same size, with a point in the center of each hexagon, which means that each element is equidistant from the other six and the space between the elements is maximized (Clark and Evans 1954; Lee and Wong 2001). Figure 1a illustrates the theoretical distribution of 30 points, showing the distances d between six of them.

This method has distances as input parameter, and this is the basic premise for its adaptation to the method based on linear features. The main idea of the proposed method is to obtain a set of distances that represent the average spacing between the linear features and compare it with a theoretical spatial distribution pattern, as we do in the traditional method. The theoretical pattern can be the same used at the Nearest Neighbor Method, assuming that the area is divided into hexagons of the same size, with a linear feature in the center of each hexagon (Figure 1b). Although the lines have different geometric parameters, such as direction, length and sinuosity, the same theoretical distribution pattern can be used because the distances that maximize the spacing are obtained between the hexagon's centers, just like in the point-based method. The use of the same theoretical pattern makes it possible to apply the framework of the point-based method to the NNMLF, whose input parameters will also be distances but between lines.



Source: (a) Adapted from De Vos (1973).

Figure 1: Theoretical spatial distribution pattern for points (a) and lines (b).

As a measure of spacing between lines, we chose to use the Hausdorff Distance metric. Hausdorff Distance is widely used for information retrieval and analysis of geometric similarity between vector objects (for points, lines or polygons) or images (Huttenlocher et al. 1992; Ariza-López and Mozas-Calvache 2012; Chehreghan and Ali Abbaspour 2017; Marošević 2018; Wang et al. 2019). According to Atkinson-Gordo and Ariza-López (2002), this distance is recurrently applied to the evaluation of positional quality (Mozas-Calvache and Ariza-López 2015; Santos et al. 2015; Mozas-Calvache et al. 2017a, 2017b; Saito et al. 2019) and in processes to control the effectiveness of cartographic generalization (Zhai et al. 2017; Guo et al. 2019; Liu et al. 2020). In these cases, the Hausdorff Distance

portrays the distance between pairs of homologous lines, whereas in this research the Hausdorff Distance will represent the spacing of different linear features. This application is also provided by Hangouet (1995).

The NNMLF algorithm is given by the ten-step procedural sequence listed below.

1) Obtain a set of n lines, from which you want to evaluate the spatial distribution pattern.

2) Determine the size A of the area of the region where the linear features to be evaluated are located.

3) Extract the coordinates of all lines' vertices from the set of n lines.

As discussed by Hangouet (1995), the Hausdorff Distance does not need to be calculated only from vertices. However, using any point on the segment of a line for the calculation would make proposed method's implementation computationally difficult. Therefore, NNMLF is based only on the coordinates of lines' vertices.

4) Considering a line i as belonging to the set of n lines, that is, $i \in \{1, 2, 3, ..., n\}$, calculate the smallest Euclidean distance of each vertex of line i in relation to any segment from line 1. In this step, the Euclidean distance is obtained between a vertex of line i and a vertex or any point on the line segments that make up line 1, always choosing the one that provides the smallest distance. Repeat the process, calculating the smallest Euclidean distance from the vertices of line i for any segment of lines 2, 3, ..., n, except for line i itself.

5) Repeat step 4 for all lines belonging to the set of n lines. That is, $\forall i \in \{1, 2, 3, ..., n\}$.

6) For each line belonging to the set of n lines, take the average of all the smallest distances calculated in relation to all other lines.

7) Obtain the Hausdorff Distance between all pairs of lines in the set of n lines. Consider j a line also belonging to the set of n lines, that is, $j \in \{1, 2, 3, ..., n\}$. Since d_{ij} is the average of the smallest distances from line i to line j and d_{ji} is the average of the smallest distances from line j to line i, the Hausdorff Distance (dh) is defined as the largest value between d_{ij} and d_{ji}, calculated according to Equation 1.

$$dh = max\{d_{ij}, d_{ji}\} \tag{1}$$

As the average of the smallest distances from line i to line j will probably be different from the average of the smallest distances from line j to line i, this step is important to obtain a unique distance between each pair of lines, as in the case of distance between points. According to Mozas-Calvache (2007), this difference is provided by the asymmetry of this type of measurement based on lines. Hangouet (1995) cites asymmetry as one of the properties of the Hausdorff Distance.

In its classical approach, the values of d_{ij} and d_{ji} are given by the maximum distances of the smallest distances from line i to line j and from line j to line i, respectively. However, Maiseli (2021) alerts to the fact that, by using maximum values in the calculation of d_{ij} and dji, this way of obtaining the Hausdorff Distance becomes very sensitive to gross errors or noise. To get around this problem, Mozas-Calvache (2007) proposed using the average of the smallest distances, instead of the maximum values, to obtain d_{ij} and d_{ji}. That is why this way of calculating the Hausdorff Distance was used in the proposed NNMLF method.

8) For each line, select the smallest observed Hausdorff Distance, which represents the distance to the nearest neighbor (dv_i).

This step must be modified when you want the nearest neighbor of orders greater than 1. For example, if you want to obtain the nearest neighbor of second order, you must select the second smallest observed Hausdorff Distance; and so on.

The description of the distribution pattern of a set of features based only on the first-order distance to the nearest neighbor is not complete, as it disregards other spatial relationships (Clark and Evans 1954). Therefore, it is interesting to use higher orders in the analysis of the spatial distribution pattern, regardless of the feature's type.

With the steps presented here, it is possible to evaluate the NNMLF up to the sixth order.

9) With the set of smaller distances obtained in the previous step, as shown by Clark and Evans (1954) and Lee and Wong (2001), the R index must be calculated, given by Equation 2.

$$R(k) = \frac{R_{OBS}(k)}{R_{ESP}(k)} \tag{2}$$

$R_{OBS}(k)$ and $R_{ESP}(k)$ are given by Equations 3 and 4, respectively.

$$R_{OBS}(k) = \frac{\sum_{i=1}^{n} dv_1\,(k)}{n} \tag{3}$$

$$R_{ESP}(k) = \gamma_{1_k}\sqrt{\frac{A}{n}} \tag{4}$$

Where,

- k is the nearest neighbor order;
- $R_{OBS}(k)$ is the observed average of the distances from each line to its k nearest neighbor;
- $R_{ESP}(k)$ is the expected average of the distances between the k nearest neighbors for a random distribution;
- $dv_i(k)$ is the Hausdorff Distance of a line i to its k nearest neighbor;
- γ_{1k} is a constant, given in terms of order k, as can be seen in Table 1;
- n is the number of lines whose spatial distribution pattern is being evaluated;
- A is the area of the region under analysis.

Table 1: Constants for calculating the nearest neighbor, depending on the order.

Order (k)	γ_1	γ_2
1	0.5000	0.2613
2	0.7500	0.2722
3	0.9375	0.2757
4	1.0937	0.2775
5	1.2305	0.2784
6	1.3535	0.2789

Source: Wong and Lee (2005).

10) Apply the Z Test.

Clark and Evans (1954) claim that applying a significance test to assess whether the observed average distance to the nearest neighbor is statistically equal to the expected average distance from the random distribution increases the reliability of the method. The Z statistic is given by Equation 5 (Lee and Wong 2001).

$$Z(k) = \frac{R_{OBS}(k) - R_{ESP}(k)}{SE_r(k)} \tag{5}$$

Where, SEr(k) is given by Equation 6.

$$SE_r(k) = \gamma_{2_k}\sqrt{\frac{A}{n^2}} \tag{6}$$

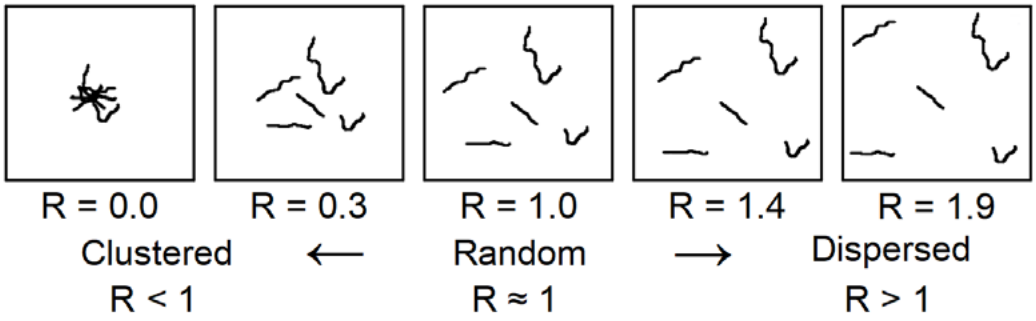
Where,

- γ_{2k} is a constant, given in terms of order k (Table 1);
- SEr(k) is the standard error of the difference between the mean values of the expected distances and the observed distances between the nearest neighbors for order k. SEr(k) describes the probability that any differences occur purely by chance;
- A is the area of the region under analysis;
- n is the number of lines whose spatial distribution pattern is being evaluated.

In this test, the null hypothesis is that the data have a random spatial distribution pattern. If the value of Z calculated for order k (Equation 5) is greater than the value of Z tabulated, the null hypothesis is rejected.

If the null hypothesis is not rejected, it means that $R_{OBS}(k)$ is statistically equal to $R_{ESP}(k)$, given that $R_{ESP}(k)$ represents the expected random pattern. In other words, it means saying that R(k) is statistically equal to 1, since it is given by the ratio between $R_{OBS}(k)$ and $R_{ESP}(k)$. If the null hypothesis is rejected and the value of R(k) is less than 1, we can say the value of $R_{OBS}(k)$ is statistically less than the expected value for the random pattern. In this case, the spatial distribution pattern is identified as clustered. On the other hand, if the null hypothesis is rejected and the value of R(k) is greater than 1, we can say that the value of $R_{OBS}(k)$ is statistically greater than the expected value for the random pattern, being the distribution pattern space considered dispersed.

A graphical representation of the R index values and their relationship with the observed spatial distribution pattern is shown in Figure 2.



Source: Adapted from Santos et al. (2016).

Figure 2: Relationship between the R index and the spatial distribution pattern of linear features.

A disadvantage of NNMLF is its computational cost. By extracting all vertices and calculating the Hausdorff Distance of these points for all lines, the method demands a considerable computational effort. Although it does not prevent the use of the method, the processing time must be considered for datasets that have lines with a high density of points, as is common in GNSS surveys using the kinematic positioning method. This limitation can be circumvented by applying some linear element generalization method before executing the NNMLF, such as the Douglas-Peucker algorithm (Douglas and Peucker 1973).

We performed two experiments to illustrate the application of the proposed method and evaluate its performance. Furthermore, to facilitate the application of the NNMLF, a computational application based on its methodology was implemented.

To facilitate the use of NNMLF in a Geographic Information System (GIS) and aiming at the dissemination of the proposed method, a plugin for QGIS version 3.0.0 or higher (QGIS 2023) was developed. This plugin called Nearest Neighbor Method for Linear Features (NNMLF) is available in the official QGIS Python plugin repository and

can be installed directly on this GIS. In addition, users interested in using this method have its version developed with parallel programming in the R language (R Core Team 2023). This code can be obtained from: https://figshare.com/s/c80a5de8afeb34bcd249.

3. Experiments and results

To analyze the behavior and validate the proposed method, two experiments were performed using real and simulated data.

3.1 Experiment 1: Simulated data and NNMLF orders

In the first experiment, two simulated datasets with known spatial distribution patterns were created to assess whether NNMLF is able to estimate these patterns correctly, up to the third order.

Of all possible spatial distribution patterns (clustered, random, and dispersed), for any order, random is probably the most difficult to simulate. This is because clustered and dispersed patterns can be emulated as extreme cases, where the distance between features is too small or too large, respectively. On the other hand, trying to portray a condition in which the spatial distribution pattern is random can be especially complicated when dealing with linear features, as these, unlike points, have length, sinuosity and direction that influence distances and make it difficult to predict a set of distances between lines that can provide the random pattern. In addition, as previously mentioned, there is no knowledge of a previous method to compare and validate the performance of NNMLF.

However, a strategy can be used for this analysis based on the description of compression by De Vos (1973). This author calls compression the process in which the distances between features decrease while the relative position between them remains unchanged. With compression, the distance to the nearest neighbor decreases, while the expected distance for a random spatial distribution pattern remains the same, given that the number of features and the area are the same (De Vos 1973).

Therefore, to assess whether the NNMLF correctly returns the expected spatial distribution pattern, the first set of simulated lines was used, divided into three subsets of ten lines each, each with a compression level in the same area, as can be seen in the Figure 3.

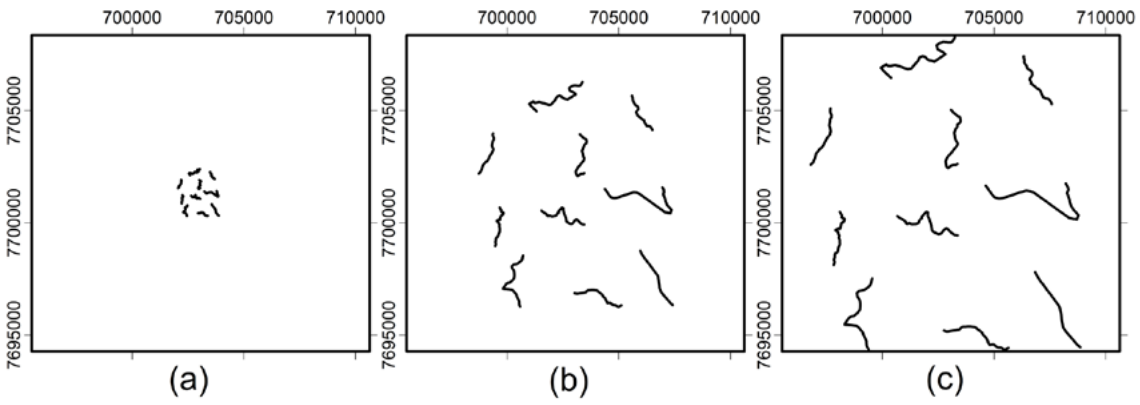


Figure 3: First set of simulated lines.

To simulate a group of dispersed lines, the linear features were arranged in such a way as to occupy the entire available area, maximizing the distance between the features, as can be seen in Figure 3c. In order to portray clustered linear features, compression was performed on the lines so that they occupied a region with a dimension of 2.25% of the area, as shown in Figure 3a. Finally, to simulate the random pattern, a compression was performed on the lines so that they occupy a region equivalent to the average of the regions of the clustered and dispersed patterns, for the same area. In this way, the middle ground between clustered and dispersed patterns is expected to provide the random spatial distribution pattern. Figure 3b presents this case.

On this set of lines, the method we propose, the NNMLF, was applied. This stage of the experiment aimed to evaluate only the first order NNMLF. The results are shown in Table 2.

Table 2: First order NNMLF result for simulated data.

Representation	**Figure 3a**	**Figure 3b**	**Figure 3c**
R_{OBS} (m)	526.720	2514.775	3495.837
R_{ESP} (m)	2281.149	2281.149	2281.149
R	0.231	1.102	1.532
Z	-4.654	0.620	3.222
Z-Score Table		1.960	
A (km^2)		208	
Confidence Level		95%	
Expected Pattern	Clustered	Random	Dispersed
NNMLF Result	Clustered	Random	Dispersed

As can be seen in Table 2, the R values are in line with what was expected for the strategy used. This shows agreement between the strategy of using compression and the value returned by NNMLF. We highlight the R index for the random pattern, which was close to 1.

For a 95% Confidence Level, the tabulated Z value is 1.96. Therefore, the null hypothesis of the Z Test, that the linear features present a random spatial distribution pattern, was rejected for the first order NNMLF for the subsets represented in Figures 3a and 3c, considering that in both cases the Z value was greater than 1.96 ($|Z(1)| > 1.96$). As the value of R of the first subset (Figure 3a) is less than one ($R(1) < 1$), it presents a clustered distribution pattern. The third subset has an R value greater than one ($R(1) > 1$), which means that its distribution pattern is dispersed. As for the second subset, the null hypothesis is not rejected for the first order NNMLF, since $|Z(1)| < 1.96$, showing that Figure 3b represents a random spatial distribution pattern.

In general, the results showed that the first order NNMLF method was effective in estimating the expected spatial distribution patterns for the simulated lines, succeeding in the three proposed cases.

To evaluate the performance of this method in higher orders (second and third orders), a second set of simulated lines was used, as represented in Figure 4. The distribution of these lines was inspired by the work of Lee and Wong (2001) and Santos et al. (2016).

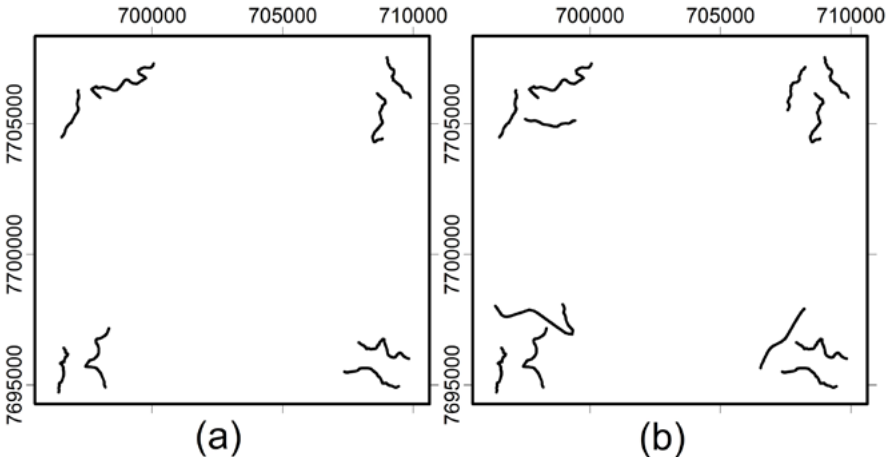


Figure 4: Second set of simulated lines.

The first subset (Figure 4a) is composed of eight linear features, arranged in such a way that the distance to the first order nearest neighbor is small and the distance to the second and third order nearest neighbors are large. Thus, for this subset of simulated data, the first order NNMLF is expected to be clustered, and the second and third order NNMLF are expected to be dispersed.

The second subset (Figure 4b) is formed by 12 lines. These features were simulated in such a way that the distances to the first and second order nearest neighbors are small. On the other hand, in this subset the distance to the third-order nearest neighbor is large. Therefore, it is expected that the first and second order NNMLF will be clustered, and the third order NNMLF will be dispersed.

NNMLF method was applied to this set of lines and the results are shown in Table 3.

Table 3: Second and third order NNMLF result for simulated data.

Representation		**Figure 4a**			**Figure 4b**	
Order	First	Second	Third	First	Second	Third
R_{OBS} (m)	1392.300	9030.152	9914.311	1301.640	1720.466	8277.309
R_{ESP} (m)	2550.402	3825.603	4782.004	2082.395	3123.592	3904.490
R	0.546	2.360	2.073	0.625	0.551	2.120
Z	-2.458	10.602	10.322	-2.485	-4.288	13.192
Z-Score Table		1.960			1.960	
A (km^2)		208			208	
Confidence Level		95%			95%	
Expected Pattern	Clustered	Dispersed	Dispersed	Clustered	Clustered	Dispersed
NNMLF Result	Clustered	Dispersed	Dispersed	Clustered	Clustered	Dispersed

The results showed that the proposed method returned small distances between features for the first order NNMLF of the first data subset (Figure 4a) and for the first and second order NNMLF of the second subset (Figure 4b), as expected. Likewise, applying the method returned large values for the distances between lines for the second and third order NNMLF of the first subset (Figure 4a) and for the third order NNMLF of the second data subset (Figure 4b).

Also considering a 95% Confidence Level, for the first subset (Figure 4a), the null hypothesis (random spatial distribution pattern) was rejected for the three NNMLF orders. This is because the calculated values of Z were

greater, in module, than 1.96: |Z(1)| > 1.96, |Z(2)| > 1.96 and |Z(3)| > 1.96. Since R(1) < 1, the first order NNMLF is considered clustered. On the other hand, as R(2) > 1 and R(3) > 1, the second and third order NNMLF were considered dispersed. For the second subset (Figure 4b), the null hypothesis was also rejected for the three NNMLF orders (|Z(1)| > 1.96, |Z(2)| > 1.96 and |Z(3)| > 1.96). As R(1) < 1 and R(2) < 1, the first and second order NNMLF were considered clustered. The third order NNMLF resulted in a dispersed pattern (R(3) > 1).

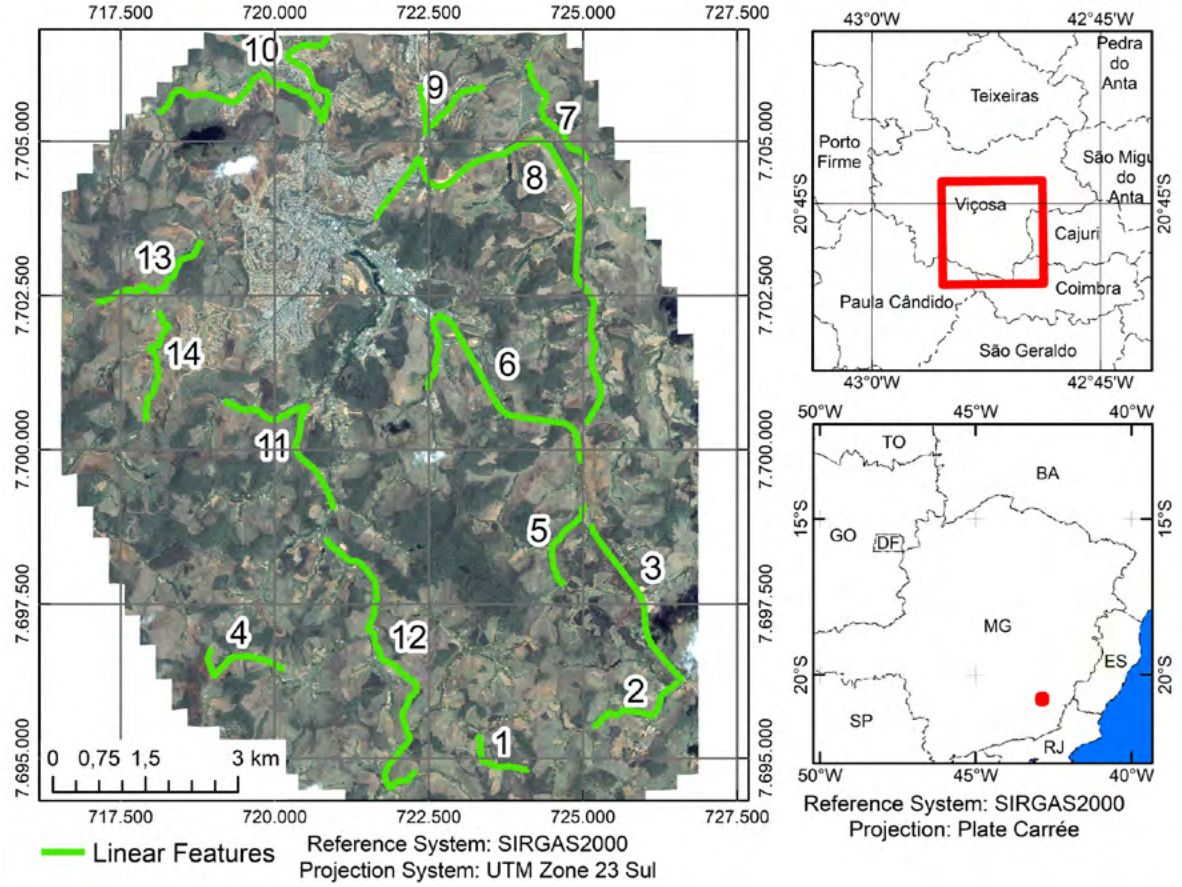
As in the case of the first data set, the results demonstrated that the NNMLF method was effective in predicting the expected spatial distribution patterns for the simulated lines, succeeding in the six proposed cases.

Overall, this first experiment demonstrated that the Nearest Neighbor Method for Linear Features is effective in estimating all possible spatial distribution patterns (clustered, random and dispersed) and in all three orders (first, second and third order NNMLF). Therefore, we can say that the proposed method performs well and is effective for estimating the spatial distribution pattern for linear features.

3.2 Experiment 2: Real data in a case study

The second proposed experiment is a case study that aimed to demonstrate the step-by-step use of the Nearest Neighbor Method for Linear Features in a real situation.

The study area of this example (Figure 5) comprises part of the municipality of Viçosa, State of Minas Gerais, Brazil, and has approximately 110 km².



Source: Adapted from Cunha et al. (2019).

Figure 5: Area for the case study.

For this experiment, 14 roads were used, collected through GNSS receivers, which were obtained from the work of Santos et al. (2015).

Following the sequence of procedures described in the previous section, the average distances of each line in relation to all the others were obtained. The result, given in meters, can be seen in Table 4.

Table 4: Average distances between lines, in meters.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	2339.22	4100.55	2568.92	2927.21	4230.00	4636.52	5227.44	8508.49	9260.03	8803.84	11555.08	11081.70	7460.88
2	1894.69	0	1502.96	896.72	4484.50	5700.09	3407.35	5366.95	6893.42	9442.90	7207.66	10073.62	10215.71	7267.44
3	3573.92	1130.69	0	652.48	6015.26	7116.67	3827.18	6257.20	6517.15	10064.48	6829.64	9562.64	10254.40	7875.32
4	3679.84	2288.36	1975.42	0	5217.21	6008.42	2183.53	4728.71	4730.33	8164.91	5038.60	7819.23	8302.70	6030.35
5	2679.44	4875.23	6705.11	3620.22	0	738.01	4488.11	2333.83	7719.49	6115.31	8064.51	10139.80	9065.93	5056.19
6	4056.35	5767.25	7449.63	4309.39	1170.92	0	4454.03	1359.50	7103.65	4715.32	7547.21	9244.59	7998.36	3990.73
7	5143.78	4302.24	4072.45	1479.24	5299.18	5407.74	0	3070.24	2479.08	5800.72	2775.02	5594.81	5728.14	3690.86
8	5117.57	5373.45	6516.92	3788.26	3127.63	2301.68	2412.46	0	4652.55	3681.19	5121.85	6860.32	5780.24	1829.95
9	7939.12	7073.03	6268.62	2041.73	7640.18	7075.41	1444.11	3780.72	0	4929.78	623.29	2773.51	3332.10	3090.60
10	8898.22	9132.99	9938.33	6898.97	6165.97	4192.93	5424.64	2974.26	5407.04	0	6197.68	5926.54	4143.03	2235.05
11	9334.73	8352.75	7215.35	3030.01	9145.94	8519.72	2485.23	5180.12	1319.11	5870.79	0	1433.43	3268.31	3894.71
12	10828.12	9951.05	8910.95	4705.50	10296.51	9277.63	4141.33	5923.62	2385.80	5804.76	816.36	0	2075.11	3638.04
13	10769.58	10246.82	9823.27	5648.88	9405.36	7961.44	5053.35	4881.59	3055.76	3799.33	3111.00	1645.20	0	1395.22
14	9036.70	8810.45	8972.21	5320.03	7228.94	5694.66	4546.70	2805.35	3404.44	2065.02	4135.88	3673.20	2044.09	0

The next step was to obtain the Hausdorff Distance, which is defined as the highest value of the average distances between pairs of lines. As mentioned in the method description, the distances between two lines are not equal. For example, the average distance from all vertices of the line with ID 1 to any segment of the line with ID 2 is 2339.22 m. The average distance of all vertices of the line with ID 2 for any segment of the line with ID 1 was 1894.69 m. Therefore, for this case, the Hausdorff Distance between lines ID 1 and 2 would be 2339.22 m.

Next, you must obtain the smallest value of the Hausdorff Distance for each line, in relation to all the others, that represents the nearest neighbor. Doing this up to the third order, we have the values presented in Table 5, in meters.

Table 5: Nearest Neighbor Method for Linear Features up to third order.

ID	First order (m)	Second order (m)	Third order (m)
1	2339.216	2927.211	3679.836
2	1502.956	2288.359	2339.216
3	1502.956	1975.417	4072.448
4	1975.417	2183.534	2288.359
5	1170.923	2927.211	3127.631
6	1170.923	2301.682	4229.999
7	2183.534	2479.078	2775.016
8	2301.682	2805.346	3070.235
9	1319.111	2479.078	2773.512
10	2235.050	3681.185	4143.029
11	1319.111	1433.431	2775.016
12	1433.431	2075.112	2773.512
13	2044.093	2075.112	3268.314
14	2044.093	2235.050	2805.346

With this set of smaller distances, we calculated the average observed distance values of each line to its k nearest neighbor (R_{OBS}), and the average expected distance for the random spatial distribution pattern (R_{ESP}) for the order k and the R index, with the help of Equations 3, 4 e 2, respectively. Table 6 presents these values for the first three NNMLF orders.

Table 6: R_{OBS} and R_{ESP} values and R Index.

	First order	Second order	Third order
R_{OBS} (m)	1753.036	2419.058	3151.534
R_{ESP} (m)	1401.530	2102.295	2627.868
R	1.251	1.151	1.199

The last step of the proposed method consisted of applying the Z Test to verify whether the calculated R index was statistically equal to the R value of the random distribution. The Z values for the three orders are presented below, in Table 7.

Table 7: Z values.

	First order	Second order	Third order
Z	1.796	1.553	2.535

Considering a Confidence Level of 95%, the tabled Z value is 1.96. Therefore, the null hypothesis of the Z Test, that the linear features present a random spatial distribution pattern, is rejected for the third order NNMLF ($Z(3) > 1.96$) and not rejected for the first and second orders NNMLF ($Z(1) < 1.96$ and $Z(2) < 1.96$). Considering that $R(3) > 1$, we have the following results:

- First order Nearest Neighbor Method for Linear Features: Random;
- Second order Nearest Neighbor Method for Linear Features: Random; and
- Third order Nearest Neighbor Method for Linear Features: Dispersed.

This sequence of procedures demonstrates that the method proposed in this research is simple to apply, an important feature for the end user and essential for computational implementation.

4. Conclusion

The development of a method to infer about the spatial distribution pattern of lines is of great importance for the evaluation of positional accuracy using linear features. In addition to making it possible to move forward in a little-explored issue, the spatial distribution of a sample can determine the validity of an evaluation process. As a practical effect, the establishment of a method for inferring the spatial distribution pattern of lines allows the development of a methodology for assessing positional accuracy that takes into account the spatial distribution of sampling elements. It is also worth mentioning that most features of a mapping are composed of lines.

The results of this study showed that the Nearest Neighbor Method for Linear Features (NNMLF) was successful in estimating the expected spatial distribution patterns in all proposed experiments, with a simulated dataset, for the first three orders of the NNMLF. Application of the proposed method on real data proved it is simple to use. On the other hand, a disadvantage of the method can be the processing time, when applied to linear features with a large number of vertices.

An important outcome is the computational tool created for the application of the NNMLF. The plugin developed will be very useful for users who wish to use this method in QGIS, enabling the popularization and greater dissemination of this research. It is worth noting that the NNMLF plugin was registered with the Brazilian National Institute of Industrial Property (Instituto Nacional da Propriedade Industrial - INPI), with registration number BR512023000701-3.

As a recommendation for future studies, based on the literature analysis, we suggest the development of a methodology for standardizing the sample size that considers the risks of the user and the producer, which is still a pending issue. Such research will allow the continuation of advances in questions little explored in the evaluation of positional accuracy using linear features. As a general conclusion, given the results presented, we can say that the NNMLF is a robust method and can be used in the evaluation of the spatial distribution pattern of linear features.

ACKNOWLEDGMENT

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) for granting Author 1 a scholarship [Funding Code 001].

AUTHOR´S CONTRIBUTION

Author 1: conceptualization, methodology, drafting, code writing, carrying out the tests, writing; Author 2: carrying out the tests (plugin), writing - revision, review, editing, final approval; Author 3: writing – revision, review, editing, final approval; Author 4: writing – revision, review, editing, final approval.

REFERENCES

Ariza-López, F. J. Atkinson-Gordo, A. D. 2008. Analysis of some positional accuracy assessment methodologies. *Journal of surveying Engineering*, 134 (2), pp.45-54.

Ariza-López, F. J. et al. 2011. Influence of sample size on line-based positional assessment methods for road data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (5), pp.708-719.

Ariza-López, F. J. García-Balboa, J. L. Ureña-Cámara, M. A. Reinoso-Gordo, J. F. 2012. Propuesta de metodología para la evaluación de la calidad de elementos lineales 3D. In: *X Congreso TOPCART - I Congreso Iberoamericano de Geomática Y. de la Tierra*. Madrid, Spain, 16 – 19 October 2012.

Ariza-López, F. J. Mozas-Calvache, A. T. 2012. Comparison of four line-based positional assessment methods by means of synthetic data. *GeoInformatica*, 16 (2), pp.221-243.

Ariza-López, F. J Rodríguez-Avi, J. Alba-Fernández, V. 2018a. A Positional Quality Control Test Based on Proportions. In: *The Annual International Conference on Geographic Information Science*, Springer: Cham, Melbourne, Australia, 28-31 August 2018.

Ariza-López, F. J Ruiz-Lendínez, J. Ureña-Cámara, M. 2018b. Influence of Sample Size on Automatic Positional Accuracy Assessment Methods for Urban Areas. *ISPRS International Journal of Geo-Information*, 7 (6), 200, pp.1-16.

Asociación Española De Normalización Y Certificación [AENOR] 2016. UNE 148002:2016: Metodología de evaluación de la exactitud posicional de la información geográfica. Madrid, España.

Atkinson-Gordo, A. D. Ariza-López, F. J. 2002. Nuevo Enfoque Para El Análisis de La Calidad Posicional En Cartografía Mediante Estudios Basados En La Geometría Lineal. In: *XIV Congreso Internacional de Ingeniería Gráfica*. Santander, España, 5-7 Junio 2002.

Chehreghan, A. Ali Abbaspour, R. 2017. An assessment of the efficiency of spatial distances in linear object matching on multi-scale, multi-source maps. *International Journal of Image and Data Fusion*, 9 (2), pp.95-114.

Clark, P. J. Evans, F. C. 1954. Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. *Ecology*, 35 (4), pp;445–453.

Cuenin, R. 1972. *Cartographie générale: Méthodes et techniques de production*. (Vol. 1), Eyrolles.

Cunha, M. M. Secatto, G. Z. Galindo, J. R. F. Santos, A. P. 2019. Proposta de um método de avaliação da acurácia posicional baseado na modificação do Buffer Simples. *Revista Brasileira de Cartografia*, 71 (4), pp.1193-1218.

De Vos, S. 1973. The use of nearest neighbor methods. *Tijdschrift voor economische en sociale geografie*, 64 (5), pp.307-319.

Douglas, D. H. Peucker, T. K. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10 (2), pp.112-122.

Drobnjak, S. Banković, R. Bakrač, S. Kostic, M. 2017. Visualization of Horizontal Positional Accuracy Assessment Results for Digital Topographic Maps at Scale 1: 25000. In: *Sinteza 2017-International Scientific Conference on Information Technology and Data Related Research*. Singidunum University, Belgrade, Republic of Serbia, 21 April 2017.

Federal Geographic Data Committee [FGDC] 1998. Geospatial Positioning Accuracy Standards Part 3: National Standard for Spatial Data Accuracy. United States.

Guo, Q. Xu, X. Wang, Y. Liu, J. 2019. Combined matching approach of road networks under different scales considering constraints of cartographic generalization. *IEEE Access*, 8, pp.944-956.

Hangouet, J. F. 1995. Computation of the Hausdorff distance between plane vector polylines. In: *Proceedings of Autocarto 12*. AUTOCARTO-CONFERENCE. Charlotte, USA. Feburary 27–March 1, 1995.

Huttenlocher, D. P. Rucklidge, W. J. Klanderman, G. A. 1992. Comparing images using the Hausdorff distance under translation. In: *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Champaign, IL, USA, 15-18 June 1992.

International Organization for Standardization [ISO] 2013. ISO 19157: Geographic Information—Data Quality. Geneva, Switzerland.

Jakobsson, A. Vauglin, F. 2002. Report of a questionnaire on data quality in National Mapping Agencies. *CERCO Working Group on Quality. Comité Européen des Responsables de Cartographie Officielle*, Marne-la-Vallée.

Lee, J. Wong, D. W. S. 2001. *Statistical analysis with ArcView GIS*. John Wiley & Sons.

Li, Z. 2006. *Algorithmic foundation of multi-scale spatial representation*. CRC Press

Liu, B. et al. 2020. A vector line simplification algorithm based on the Douglas–Peucker algorithm, monotonic chains and dichotomy. *ISPRS International Journal of Geo-Information*, 9 (4), pp.1-14.

Maiseli, B. J. 2021. Hausdorff Distance with Outliers and Noise Resilience Capabilities. *SN Computer Science*, 2 (5), pp.1-12.

Marošević, T. 2018. The Hausdorff distance between some sets of points. *Mathematical Communications*, 23 (2), pp.247-257.

Mozas-Calvache, A. T. 2007. *Control de calidad posicional en cartografía por elementos lineales*. PhD. Universidad de Jaén.

Mozas-Calvache, A. T. Ariza-López, F. J. 2010. Methodology for positional quality control in cartography using linear features. *The Cartographic Journal*, 47 (4), pp.371-378.

Mozas-Calvache, A. T. Ariza-López, F. J. 2011. New method for positional quality control in cartography based on lines. A comparative study of methodologies. *International Journal of Geographical Information Science*, 25 (10), pp.1681-1695.

Mozas-Calvache, A. T. Ariza-López, F. J. 2015. Adapting 2D positional control methodologies based on linear elements to 3D. *Survey Review*, 47 (342), pp.195-201.

Mozas-Calvache, A. T. Pérez-García, J. L. Fernández-Del Castillo, T. 2017a. Monitoring of landslide displacements using UAS and control methods based on lines. *Landslides*, 14 (6), pp.2115-2128.

Mozas-Calvache, A. T. Ureña-Cámara, M. A. Ariza-López, F. J. 2017b. Determination of 3D displacements of drainage networks extracted from digital elevation models (DEMs) using linear-based methods. *ISPRS International Journal of Geo-Information*, 6 (8), pp.1-17.

Mozas-Calvache, A. T. 2021. Positional quality assessment based on linear elements. *Revista cartográfica*, 103, pp.11-31.

Nero, M. A. et al. 2017. A computational tool to evaluate the sample size in map positional accuracy. *Boletim de Ciências Geodésicas [online]*, 23 (3), pp.445-460.

Oliveira, G. D. et al. 2018. Correção geométrica de imagens orbitais a partir das coordenadas de vértices de imóveis certificados pelo INCRA. *Revista Brasileira de Cartografia*, 70 (1), pp.290-324.

QGIS 2023. QGIS Geographic Information System [software]. QGIS Association. Available from: http://www.qgis.org [Accessed 08 February 2023].

R Core Team 2023. R: A language and environment for statistical computing [software]. R Foundation for Statistical Computing. Available from: https://www.R-project.org/ [Accessed 08 February 2023].

Ripley, B. D. 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 9 (2), pp.172-192.

Ruiz-Lendínez, J. J. Ureña-Cámara, M. A. Ariza-López, F. J. 2017. A polygon and point-based approach to matching geospatial features. *ISPRS International Journal of Geo-Information*, 6 (12), pp.1-24.

Santos, A. P. Medeiros, N. G. Santos, G. R. Rodrigues, D. D. 2015. Controle de qualidade posicional em dados espaciais utilizando feições lineares. *Boletim de Ciências Geodésicas*, 21 (2), pp.233-250.

Santos, A. P. Rodrigues, D. D. Santos, N. T. Gripp Junior, J. 2016. Avaliação Da Acurácia Posicional Em Dados Espaciais Utilizando Técnicas De Estatística Espacial: Proposta De Método E Exemplo Utilizando A Norma Brasileira. *Boletim de Ciências Geodésicas*, 22 (4), pp.630-650.

Santos Filho, H. D. Cornero, C. Pereira, A. Nero, M. A. 2022. Cartographic Accuracy Standard (CAS) of the digital terrain model of the digital and continuous cartographic base of the state of Amapá: case study in the city of Macapá. *Boletim de Ciências Geodésicas*, 28 (03), pp.1-20.

Saito, Y. K. et al. 2019. Influência da densidade de vértices nos métodos Distância de Hausdorff e Influência do Vértice. *Revista Brasileira de Cartografia*, 71 (2), pp.598-618.

Seo, S. O'Hara, C. G. 2009. Quality assessment of linear data. *International Journal of Geographical Information Science*, 23(12), pp.1503-1525.

Silva, D. C. 2020. *Estudo da qualidade posicional e análise de padrões espaciais na distribuição de erros altimétricos em modelos digitais de elevação*. Master. Universidade Federal de Pernambuco.

Thapa, K. 1988. Automatic line generalization using zero-crossings. *Photogrammetric Engineering and Remote Sensing*, 54 (4), pp.511-517.

Van Niel, T. G. Mcvicar, T. R. 2001 Assessing positional accuracy and its effects on rice crop area measurement: an application at Coleambally Irrigation Area. *Australian Journal of Experimental Agriculture*, 41 (4), pp.557-566.

Wang, S. Guo, Q. Xu, X. Xie, Y. 2019. A study on a matching algorithm for urban underground pipelines. *ISPRS International Journal of Geo-Information*, 8 (8), pp.1-23.

Wong, D. W. S. Lee, J. 2005. *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. John Wiley & Sons.

Wu, H. et al. 2021. A comprehensive quality assessment framework for linear features from Volunteered Geographic Information. *International Journal of Geographical Information Science*, 35(9), pp.1826-1847.

Xavier, E. Ariza-López, F. J. Ureña-Cámara, M. A. 2019. Automatic evaluation of geospatial data quality using web services. *Revista Cartográfica*, 98, pp.59-73.

Zanetti, J. 2017. *Influência do número e distribuição de pontos de controle em ortofotos geradas a partir de um levantamento por VANT*. Master. Universidade Federal de Viçosa.

Zhai, J. et al. 2017. Quality Assessment Method for Linear Feature Simplification Based on Multi-Scale Spatial Uncertainty. *ISPRS International Journal of Geo-Information*, 6 (6), pp.1-25.