Artigo

USE OF GEOSTATISTICS ON ABSOLUTE POSITIONAL ACCURACY ASSESMENT OF GEOSPATIAL DATA

Uso da Geoestatística na Avaliação da Acurácia Posicional Absoluta de dados Geoespaciais

Alex da Silva Santos ¹
Nilcilene das Graças Medeiros ²
Gérson Rodrigues dos Santos ³
Jugurta Lisboa Filho ⁴

- ¹ Diretoria de Geociências, Coordenação de Cartografia, Instituto Brasileiro de Geografia e Estatística-IBGE, Rio de Janeiro, RJ, Brasil. Fundação Presidente Antônio Carlos FUPAC, Curso de Engenharia Ambiental e Sanitária, Leopoldina, MG,. Brasil. E-mail: alex.santos@ibge.gov.br.
- ² Departamento de Engenharia Civil, Universidade Federal de Viçosa- UFV, Viçosa, MG, Brasil. E-mail: nilcilene.medeiros@ufv.br
- ³ Departamento de Estatística, Universidade Federal de Viçosa UFV, Viçosa, MG, Brasil. E-mail: gerson.santos@ufv.br
- ⁴ Departamento de Informática, Universidade Federal de Viçosa UFV, Viçosa MG, Brasil. E-mail: jugurta@ufv.br

Abstract:

In the area of Geosciences it is intuitive to think of spatial correlation as a phenomenon under study, and Geostatistics has tools to identify and represent the behavior of such dependency. The spatial analysis of the results of an inspection of the quality of a cartographic product is generally not addressed in the standards, which are restricted to descriptive and tabular findings, based on the assumption of the Classical Statistics of independence of observed data. At the Brazilian National Infrastructure of Spatial Data (INDE), various cartographic products should be made available to society, along with their metadata. This paper proposes a methodology for quality inspection based on international standards and on the Cartographic Accuracy Standard (PEC), using geostatistical methods and spatial representation of this benchmarking, through positional quality maps. The method of evaluating the quality of data was applied to Brazil's Continual Cartographic Base, at the scale of 1:250000 - BC250, with a focus on absolute positional accuracy. The quality map generated presented regionalizations of the planimetric error confirmed by the producer team of the referred cartographic base of IBGE. Such information can help users and producers to understand the spatial behavior of cartographic product quality in study.

Keywords: Geostatistics, Spatial data quality, Accuracy positional, Visualisation Uncertainty

Resumo: Na área de Geociências é intuitivo pensar na correlação espacial de um fenômeno em estudo, e a Geoestatística dispõe de ferramentas para identificar e representar o comportamento desta dependência. A análise espacial dos resultados de uma inspeção de qualidade de um produto cartográfico, geralmente, não é abordada nas normas, restrita apenas à conclusões descritivas e tabulares, baseadas no pressuposto da Estatística Clássica de independência dos dados observados. Na Infraestrutura Nacional de Dados Espaciais do Brasil (INDE), diversos produtos cartográficos devem ser disponibilizados à sociedade, junto com seus metadados. A presente pesquisa propõe uma metodologia de inspeção da qualidade, baseada em normas internacionais e no Padrão de Exatidão Cartográfica (PEC), usando métodos geoestatísticos e a representação espacial desta aferição, por meio de mapas de qualidade posicional. O método de avaliação da qualidade de dados foi aplicado sobre a Base Cartográfica Contínua do Brasil, na escala 1:250000 - BC250, com foco na acurácia posicional absoluta. O mapa de qualidade gerado apresentou regionalizações do erro planimétrico, confirmadas pela equipe produtora da referida base cartográfica do IBGE. Tais informações podem subsidiar usuários e os próprios produtores no entendimento do comportamento espacial da qualidade do produto cartográfico em estudo.

Palavras-chave: Geoestatística, Qualidade de dados espaciais, Acurácia posicional, Visualização de incertezas

1. Introduction

The cartographic process, incorporated into computational means, has allowed the increasing number of producers and principally users of spatial data. To meet the demands of mapping with satisfactory quality products and to keep track of technological developments, standards and minimum parameters should be adopted (Dalmolin, 2001).

ISO 19100 series of standards are intended for geographic information, specifically the ISO 19157:2013 and ISO 19115:2003 standards, which refer to quality of spatial dataset. Geospatial metadata facilitates the location, identification, and even provides a qualitative and quantitative description of the quality of a set of geospatial data. They are all of great value for the Geosciences field, but the results of quality inspection, complete or by sampling, proposed by these standards, are purely descriptive and tabular. The spatial representation of the results of quality inspection, although trivial and intuitive, is not addressed in this series of standards.

The main purpose of this paper is to present the use of Geostatistics in the identification of probable regionalizations of quality in relation to the planimetric error of a cartographic product. The paper also shows the assessment and spatial representation of positional quality, identifying the group of possible systematic error pointed out by the Classical Statistics in graphic and descriptive form.

2. Spatial data quality

Concepts of quality, the kinds of uncertainty to depict and the application of its principles are essential in evaluating the spatial dataset. In order to evaluate the quality of spatial data, the quality elements must be considered, from the perspectives of the producer and the user of the spatial information.

ISO 19157:2013 standard addresses 6 (six) elements of spatial data quality: positional accuracy, thematic accuracy, temporal accuracy, completeness, logic consistency, usability and their respective sub-elements, as shown in Table 1

It should be noted that cartography is dedicated to represent real-world models, and the accuracy of a cartographic product is determined by its compliance with an abstract model of reality, which is described in the technical specifications of the spatial dataset or nominal ground (Longley et al, 2013).

Table 1: *Elements and sub-elements of quality*

Element	Sub-element		
Positional accuracy	Absolute accuracy		
	Relative accuracy		
	Positional accuracy of the data in the grid		
Thematic accuracy	Correction of classification		
	Degree of correction of quantitative attributes		
	Accuracy of quantitative attributes		
Temporal accuracy	Accuracy of a temporal measure		
	Temporal consistency		
	Temporal validity		
Completeness	Commission		
	Omission		
Logical consistency	Conceptual consistency		
	Domain consistency		
	Format consistency		
	Topological consistency		
Usability			

Source: based on ISO 19157:2013

2.1 Absolute positional accuracy

According to López (2002), positional accuracy is a traditional emblematic aspect of cartographic productions. This refers to the planimetric and altimetric accuracy of the spatial dataset, and according to the ISO 19157:2013 standard, there are 3 (three) types of positional accuracy:

Absolute or external accuracy: proximity of coordinate values observed to values accepted as true.

Relative or internal accuracy: proximity of coordinate values observed to the relative position of the dataset features and their respective positions accepted as true.

Accuracy of a data grid: proximity to the position of a data grid to the acceptable or real value.

2.2 Cartographic Accuracy Standard (PEC)

In Brazil, the assessment of positional quality is based on the Cartographic Accuracy Standard - PEC, established in Decree Law 89,817 of June 20, 1984. This is designed for measurement of analog cartographic products (Brasil 1984).

The planimetric and altimetric reference systems for Brazilian Cartography are those that define the Brazilian Geodetic System - SGB, as established by the Brazilian Institute of Geography and Statistics - IBGE, in its specifications and standards (Brasil 2005).

According to Oliveira (2011), PEC considers to be equivalent the terms "standard error" terms (EP), "standard deviation" (σ) and "root mean square error" (RMSE), without presenting mathematical formulations for calculation, causing dubious interpretations. The standard deviation and standard error portray the variability of the data around the arithmetic mean of the sample space. The RMSE is associated with the difference between an observed value, in the spatial dataset, and the value taken as a reference, for measuring the planimetric error (e_p) and/or altimetric error (e_p) .

The classification of the positional accuracy of cartographic products must meet the two criteria of the PEC simultaneously, which are RMSE associated with the precision and PEC with the trend of observed data (Brasil 1984).

2.3 Sampling techniques applied to spatial data

Sampling is the process of selecting a sample, which enables the study of the characteristics (parameters) of a population. In order to estimate the behavior of the population from samples it is necessary that the subset is collected in such a way that each observation has the same chance of being selected (Landim, 1998).

Most of the time, when dealing with a set of spatial data, one can note its complexity, due to territorial dimensions and the quantity of information represented. Thus, in various situations, it is not feasible to perform a thorough inspection of the dataset, except for processes that can be automated, such as some evaluations of logical consistency. The rest demand samples representing the cartographic product as a whole.

The specification of the tolerable sampling error, must be made under a probabilistic approach. In addition, financial constraints may be considered in determining the sample size (n) and, consequently, in the calculation of the desired tolerable sampling error (Barbetta, 2012).

According to Barbetta (2012), a first calculation of sample size (n_0) , based only on the tolerable sampling error (E_0) , independent of the size of the population (N) may be determined using Equation 1,

$$n_0 = \frac{1}{E_{0^2}} \tag{1}$$

In relation to the coverage area spatial distribution of samples, there exists a good relationship between systematic random sampling and methods of spatial analysis of the results of sampling inspection, on the basis of comprehensiveness in the spatial distribution of samples (Yamamoto, 2013)

3. Geostatistics applied to visualizing spatial data quality

Geostatistics is a science based on the theory of regionalized variables, which have a spatial behavior and show intermediate characteristics between truly random variables and fully deterministic variables. In the study of regionalized variables two tools are fundamental: the variogram and kriging (Guerra, 1998; Andriotti, 2003).

The determination of the variogram is described as the first and most important step in the procedure of a geostatistical estimate, because it will influence the whole process of kriging, the results and conclusions. In practice it is the mathematical tool that allows for one to analyze the natural dispersion of regionalized variables, representing the degree of continuity of the studied phenomenon, as show in Equation 2 (Guerra, 1988),

$$2\hat{\lambda}(h) = \frac{1}{n(h)} \sum_{i=1}^{n(h)} [Z(x_i) - Z(x_i + h)]^2$$
 (2),

where: $Z(x_i)$ is the value of the variable at the point x_i ; $Z(x_i + h)$ is the value of the variable at the point $x_i + h$; n(h) is the number of pairs at a distance h.

Similar to the statistical method, there are 3 (three) types of variograms: the observed, or experimental, the real, and the theoretical variogram. In practice, only the experimental variogram is known and the theoretical variogram serves as reference to estimate the real variogram. Figure 1 illustrates the three types of variograms (Guerra, 1988).

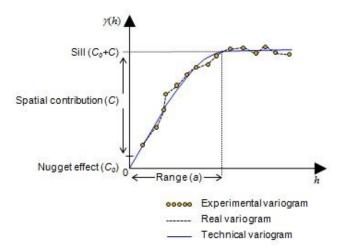


Figure 1: The three variograms of Geostatistics

The main parameters of a variogram are: the range (a), which is the distance from which the samples become independent, that is, the spread between two values observed are aleatory and without spatial correlation, this is the principal of Classical Statistics; the nugget effect (C_0) , which is the discontinuity at the origin of the variogram, being a particular correlation between samples, where $\gamma(0)=0$ and $\gamma(e)=Cov(0)$ (nugget effect), e being the shortest distance between samples. This can be attributed to measurement errors, or to the fact that the data has not been collected at intervals small enough to show the underlying spatial behavior of the phenomenon under study (Landim, 1998; Vieira, 2000).

The sill $(C_0 + C)$ of a variogram represents the value of the total variance of the variable under study and equivalent to the point of its stabilization, the spatial dispersion or variance is represented by (C). Theory shows that the variance of the observed data of the studied population is equivalent to this point of stabilization and limits the growth of the semivariogram (Guerra, 1988; Andriotti, 2003).

The semivariogram expresses the spatial behavior of the regionalized variable and its residues, such as: the size of the zone of influence around a sample, anisotropy, and the continuity by means of form of the variogram (Landim, 1998).

Kriging allows the prediction of values in non-observed locations and is show to be superior to other interpolation methods, since it determines the spatial distribution and the accuracy of the phenomenon under study. Kriging is known as BLUP (Best Linear Unbiased Predictor). This geoestatistic estimator is based on an adjustment of a theoretical variogram to an experimental variogram, in order to make inferences about the real variogram. In this paper the kriging result permits the generation of quality maps of the evaluated dataset, which allow the visualization of uncertainty. Equation 3, shows the mathematical formulation of one of the predictors to predict values of the variable under study in non-observed points,

$$\hat{Z}_k(x_0) = \sum_{i=1}^n \lambda_i . Z(x_i)$$
(3),

where: $\sum_{n=1}^{n} = 1$ (condition of universality, without trend); $\sigma_k^2 = E[Z - \hat{Z}_k]^p$ is minimal (optimality condition); λ_i are the weights associated with the experimental data $Z(x_i)$; n is the total number of such data; \hat{Z}_k is the kriging estimator and provides the variance of the estimate; and Z is the observed value (Guerra, 1988; Landim, 1998; Santos, 2010b).

4. Materials and methods

Following the objective of this research, which focuses on the study of use of Geostatistics coupled with the adoption of international standards for evaluating the quality of cartographic products and the spatial representation of the result of this measurement, data and inputs which are available to society were used.

For measuring the absolute positional accuracy of dataset, Brazil' Continual Cartographic Base, at scale of 1:250000 - BC250, easily identifiable features on the terrain such as: roads, junction of roads, level crossings and bridges were considered. The level of compliance of the absolute

positional accuracy planimetric of the dataset, in the scale of 1:250000, was based in PEC, as show Table 2. For a cartographic product to be considered Class A, it must have 90% of the planimetric error less than 125 m (0.5 mm of scale). The control points for the points collected in the field are from other projects of IBGE's Coordination of Cartography (CCAR) and were not used for the production of assessment spatial dataset.

Table 2: Conformance levels to the absolute positional accuracy
--

Class PEC(meters)		EP (meters)	
A	125 m	75 m	
В	200 m	125 m	
С	250 m	150 m	

EP - Standard error

The flowchart shown in Figure 2 presents the overview of the methodology of evaluation of the quality of cartographic products based on ISO 19157:2013 and ISO 19115:2003 standards, proposed in Santos (2013), with emphasis on spatial analysis of the results of sampling inspection using Geostatistics and subsequent reporting of this benchmarking in metadata.

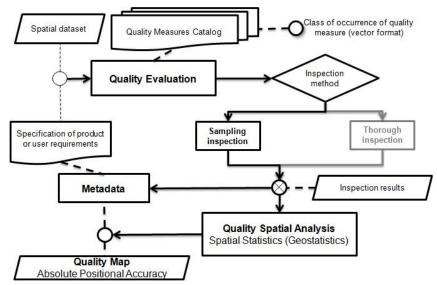


Figure 2: Flowchart for geospatial data quality assessment

4.1 Inspection method for positional accuracy

According to ISO 19157:2013, in order to supplement the assessment of quality required in PEC, a quality measure was drafted before performing the positional quality assessment. In addition, a class of occurrence to aid in counting and measuring of the planimetric error was created, as shown in Figure 2.

The goal was to have one point noted per 1:250000 sheet. The inspection method chosen was systematic random sampling. The grid utilized was articulations of systematic mapping, at scale

of 1:100000. Each cell delineated an inspection area, similar to search windows, where field points (GPS) could be identified and compared with their counterparts, on the referred cartographic base, as shown in Figure 3.

However, despite the significant number of field points, in some inspection areas it was not possible to perform the correlation between the field points and their correspondent on BC250, because some geographic features are not representable in scale of 1:250000, making it impossible to identify their counterparts. It should be noted that there are no field points in the Legal Amazon region.

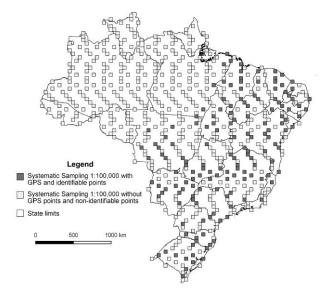


Figure 3: Spatial distribution of the sampling areas for collection of points relating to the evaluation of absolute positional accuracy

5. Discussion of the results

The result of evaluating the quality of absolute positional accuracy in relation to the planimetric error of the dataset is expressed in quantitative and tabular values, present in each sampling unit. Table 3 presents the exploratory analysis of observed data, and Figure 3 presents its respective histogram, relating to the evaluation of absolute positional accuracy of BC250.

Table 3: Descriptive statistics on the absolute positional accuracy (planimetric error) of BC250

Number of observed points	204
Average	82.849 m
Standard deviation	91.577 m
Variance of observed data	8386.495 m ²
Maximum value	789.600 m
Upper quartile	98.900 m
Median	57.840 m
Lower quartile	33.320 m
Minimum value	3.916 m

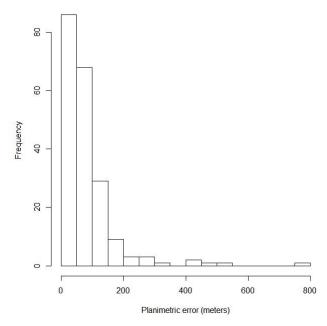


Figure 4: Histogram referring to the planimetric error of BC250

Figure 5 shows the spatial distribution of points and the density of values under study, according to quartile division. As mentioned, the lack of data observed in the northern region of Brazil occurs due to the absence of field points, which made the analysis of results in this region unfeasible.

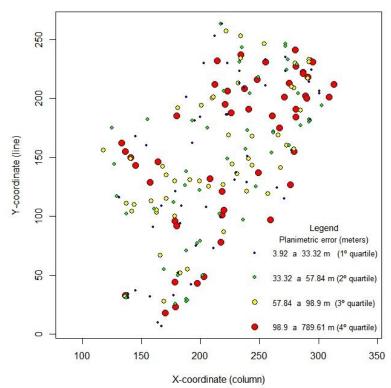


Figure 5: Spatial distribution of the observed data concerning absolute positional accuracy

In Figure 5 it can be observed that most of the observed data is below the fourth quartile, whose minimum value is less than the Class A of PEC. Thus, considering the size of the sample of 204

sampling units, the calculated sampling tolerable error (E_0) , based on Barbetta (2012), was of 7.01% as the Equation 4

$$E_0 = \sqrt{\frac{1}{204}} = 0,0701 \tag{4}$$

The observed values were used to calculate the percentage of planimetric errors that were above Class A of PEC for the 1:250000 scale. Table 4 summarizes the quantities and percentages observed.

	Table 4. Symmesis of Be230 classification according to 1 Be					
Class	Interval	Percentage of points in the interval	Amount of points in the interval			
A	0 to 125 meters	84,31%	172			
В	0 to 200 meters	94,12%	192			
С	0 to 250 meters	95,59%	195			

Table 4: Synthesis of BC250 classification according to PEC

The RMSE (Root Mean Square Error) calculated was of 123.63 meters for the 204 points observed in the BC250. Thus, it is concluded that their absolute positional accuracy can be classified as Class B, in accordance with PEC for the scale of 1:250000 and the Standard Error considered. However, the geostatistical analysis of this quality inspection identified regionalizations of these positional uncertainties, in addition to spatial dependence of the observed data, breaking one of the principal of the Classical Statistics of independence of the observed data.

On the independence (Chi-square) and normality test (Shapiro-Wilk), as recommended by Santos (2010a), it was found that the data does not follow the normal distribution (p-value=2.2e-16, for the confidence level of 90%). This conclusion is confirmed by Geostatistics, where spatial dependence was identified from the data observed through the variographic analysis.

5.1 Variographic analysis

To find the best adjustment from the theoretical variogram to the experimental variogram, three theoretical models were considered: spherical, exponential, and Gaussian. The best adjustment to the experimental variogram was obtained by the spherical model, as shown in Table 5 with the respective variogram parameters adjusted, illustrated in Figure 6.

Table 5: Adjustment parameters of the variogram for absolute positional accuracy

Spherical	
6100 m ²	
9000 m ²	
8386.495 m ²	
2900 m ²	
130 km	

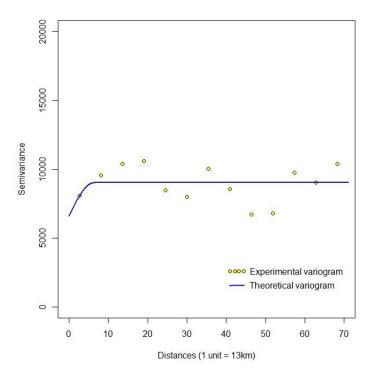


Figure 6: Adjustment of the theoretical variogram to the experimental variogram of commission data

The behavior of the variogram, as show in Figure 6, indicates the existence of regionalizations in different areas, but with similar spatial dependence observed by the isotropy of the phenomenon.

5.2 Kriging of absolute positional accuracy data

Kriging was performed in order to estimate the values of planimetric error for non-inspected areas and to provide a visualization of spatial data quality of cartographic product. A cartographic editing of the quality map for positional accuracy was then conducted. In this case, in addition to the benchmarks of quality maps, the delimitation of the Legal Amazon was added to highlight the area where there are no field control points for measuring positional quality field of the dataset under study. This step concluded the process of quality evaluation and generated the spatial representation, depicted in the data quality map, as shown in Figure 7.

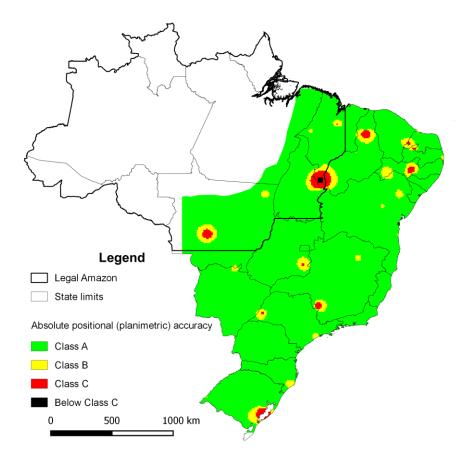


Figure 7: Absolute positional accuracy quality map of BC250

The quality map shows the concentration of regions with high planimetric error, close to the demarcation of Legal Amazon, indicating possible systematic errors, according to the BC250 producer team. These are places of difficult access, such as the Pantanal Mato-grossense, or areas with excess of clouds, which caused difficulties in collecting and identifying points to orthorectificate the orbital images of the project.

6. Conclusion

The process of assessing the positional accuracy of the BC250 proved to be efficient. The systematic errors were depicted on the regionalizations on absolute positional accuracy quality map. The visualization of data quality of cartographic products may serve as an aid to a future update and improvement for producers and show an overview of this data quality for users.

For the positional accuracy of cartographic products, the PEC assumes that the planimetric error values in the spatial dataset are independent. However, as noted, there is a spatial dependence and even though it's minimal, it does break an assumption of the Classical Statistics. Therefore, Geostatistic is recommended to assess the spatial dataset.

Systematic random sampling technique proved to be efficient, showing a spatial distribution suitable for the coverage area of the evaluated dataset. In practice, factors such as difficult accessibility of certain regions, or the presence of clouds in the images analyzed, among others, hampered the collecting and the identifying of control points of the projects. Thus, a larger

search window would facilitate the collection, selection and identification of field points for measuring the planimetric error and performing the geostatistical analysis.

It is recommended that, a study is performed in relation to the size of the sample, its spatial distribution and the size of sampling units, considering the nominal scale cartographic product (technical specifications), the element of quality or spatial phenomenon to be assessed, its tolerable sampling error and financial factors.

In function of the identification of spatial dependence by Geostatistics, it is recommended to find a common denominator between the amount of field control points, their spatial distribution and as far as is necessary to optimize the quantity of points collected in the coverage area. It should be noted that if the field points are collected at intervals over a greater distance than the range of the spatial dependence, Geostatistics will not identify the spatial dependence and Classical Statistics can be used.

The use of Geostatistics combined with Classical Statistics allows the possibility of collecting the field points in an efficient form in relation to their spatial distribution, the quantity collected and the reduction of costs.

Acknowledgements

The authors would like to thank the Brazilian Institute of Geography and Statistics - IBGE, the Federal University of Viçosa - UFV, the Brazilian National Council for Scientific and Technological Development - CNPq and the Minas Gerais State Foundation for Research Development - FAPEMIG for providing the realization of this survey.

REFERENCES

Andriotti, J. L. S. 2003. *Fundamentos de Estatística e Geoestatística*. São Leopoldo, RS. Editora UNISINOS. 165p.

Barbetta, P. A. 2012. *Estatística aplicada às Ciências Sociais*. 8rd ed. Florianópolis, SC. Editora da UFSC. 318p

Brasil, Decreto n.º 89.817, de 20 de junho de 1984. Dispõe sobre as instruções reguladoras das normas técnicas da cartografia nacional. Diário Oficial (da República Federativa do Brasil), Poder Executivo, Brasília, 22 de julho de 1984.

Brasil, Decreto n.º 5.334, de 6 de janeiro de 2005. Dá nova redação ao art. 21 e revoga o art. 22 do Decreto nº 89.817, de 20 de junho de 1984, que estabelece as Instruções Reguladoras das Normas Técnicas da Cartografia Nacional. Diário Oficial (da República Federativa do Brasil), Poder Executivo, Brasília, DF, 6 de janeiro de 2005.

Dalmolin, Q.; Leal, E. M. 2001. Análise da qualidade posicional em bases cartográficas geradas em CAD. Curitiba. *Boletim de Ciências Geodésicas*, 7(1), pp. 21-40.

Guerra, P. A. G. 1988. *Geoestatística Operacional*. Brasília. Departamento Nacional da Produção Mineral.

ISO 19115. 2003. *Geographic Information – Metadata*. Genebra, Suíça. International Standardization for Organization.

ISO 19157. 2013. *Geographic Information – Data quality*. Genebra, Suíça. International Standardization for Organization. ISO TC/211 Secretariat.

Landim, P. M. B. 1998. *Análise Estatística e dados geológicos*. São Paulo: Fundação Editora da UNESP.

Longley, P. A., Goodchild, D. J M. and Rhind, D. W. 2013. *Sistema de Informações Geográficas*. 3rd ed. Bookman Companhia Editora Ltda.

López, F. J. A. 2002. Calidad en La Producción Cartográfica. Madrid. Editora RA-MA.

Oliveira, L. F. 2011. Análise de modelos digitais de superfície gerados a partir de imagens do sensor PRISM/ALOS. M.Sc.Viçosa, MG. UFV.

Santos, A.P. 2010a. Avaliação da acurácia posicional em dados espaciais com o uso da estatística espacial. M.Sc. Viçosa, MG. UFV.

Santos, G. R. 2010b. Hierarquização geométrica dos preditores geoestatísticos. PhD. Lavras, MG. UFLA.

Santos, A. S. 2013. Geoestatística aplicada na avaliação da qualidade de dados geoespaciais. M.Sc. Viçosa, MG. UFV.

Vieira, S.R. 2000. Geoestatística em estudos de variabilidade espacial do solo. In: Novais, R. F.; Alvarez, V. H.; Schaefer, G. R. *Tópicos em ciência do solo*. Viçosa, MG. Sociedade Brasileira de Ciência do Solo. Ch.1, pp.1-54.

Yamamoto, J. K.; Landim, P. M. B. 2013. *Geoestatística: conceitos e aplicações*. São Paulo: Oficina de Textos.

Recebido em 29 de março de 2016. Aceito em 7 de fevereiro de 2017.