TRANSPORTES
revistatransportes.org.br
anpet
associação nacional de
pesquisa e ensino em transportes
anpet.org.br

Should we account for network distances or anisotropy in the spatial estimation of missing traffic data?

Devemos considerar distâncias em rede ou anisotropia na estimativa espacial de dados faltantes de tráfego?

Samuel de França Marques[1], **Renan Favero**[2], **Cira Souza Pitombo**[1]

[1]University of São Paulo, São Carlos, São Paulo – Brazil

[2]University of Florida, Gainesville, Florida – United States

contato: samuelmarques@usp.br, (SFM); renanfavero@ufl.edu, (RF); cirapitombo@gmail.com, (CSP)

Submitted:
28 July, 2022
Accepted for publication:
6 March, 2023
Published:
3 May, 2023

Editor in charge:
José Reynaldo Setti

Keywords:
Annual average daily traffic.
Ordinary Kriging.
Network distances.
Anisotropy.

Palavras-chave:
Volume médio diário anual.
Krigagem Ordinária.
Distâncias em rede.
Anisotropia.

DOI: 10.58922/transportes.v31i1.2822

ABSTRACT

In light of the unavailability of traffic volume data for all road segments, the scientific literature proposes estimating this variable using spatial interpolators. However, most of the methods found use the Euclidean distance between the database points as a proximity measure, in addition to ignoring the anisotropy of the phenomenon. Thus, the objective of the present study was to apply Ordinary Kriging (OK) with network distances and considering the anisotropy in traffic volume data on highways in the state of São Paulo (Brazil). Additionally, the two previous results were compared to the traditional isotropic approach with Euclidean distances. Goodness-of-fit measures confirmed the good performance and better suitability of OK with network distances over the analyses that use Euclidean distances. Addressing the anisotropy of the traffic volume data also helped to improve the results. The proposed method can effectively support estimating traffic volume in segments without flow data.

RESUMO

Tendo em vista a indisponibilidade de dados de volume de tráfego para todos os trechos viários, a literatura científica propõe a estimativa dessa variável a partir de interpoladores espaciais. Contudo, a maioria das abordagens encontradas utiliza a distância euclidiana entre os pontos do banco de dados e ignora a anisotropia do fenômeno. Dessa forma, o objetivo do presente trabalho foi aplicar a Krigagem Ordinária (KO) com distâncias em rede e anisotropia ao volume de tráfego em rodovias do estado de São Paulo, comparando seus resultados aos da abordagem isotrópica com distâncias euclidianas. Métricas de aderência confirmaram o bom desempenho e melhor adequabilidade da KO com distâncias em rede, em detrimento das análises com distâncias euclidianas. Tratar a anisotropia do volume de tráfego também contribuiu para a melhoria dos resultados. O método proposto pode servir efetivamente como suporte à estimativa do volume de tráfego em trechos sem dados de fluxo.

1. INTRODUCTION AND BACKGROUND

The Annual Average Daily Traffic (AADT) is an important variable of interest for the Traffic Engineering area, as it is the basis for pavement design, accident modeling, identification of critical segments, level of service analysis (DNIT, 2006) etc. However, this data is only attained directly in segments provided with counting devices, survey stations, tolls, and others. Bearing in mind the need to know the traffic volume in segments without counting stations, the scientific literature has methods, from the simplest to the most sophisticated, to estimate the AADT along an entire road network.

In this context, there are deterministic interpolators, such as inverse distance weighting, trend analysis, historical average and nearest neighbor (Klatko et al., 2017; Yang et al., 2018), which depend only on data referring to the variable of interest itself. Solutions based on the sequential transport planning method (Ortúzar and Willumsen, 2011) can also be found, which, in addition to the information provided by the counting stations, also require obtaining an origin and destination matrix (UFRJ, 2018; Wang et al., 2013). Multivariate approaches, which also use explanatory variables, have been consistently used to estimate AADT and can be divided into two main groups: 1) statistical models, such as generalized linear and local spatial models (Apronti et al., 2016; Duddu and Pulugurtha, 2013; Pulugurtha and Kusam, 2012; Pulugurtha and Mathew, 2021); and 2) machine learning algorithms, such as neural networks and support vector regression (Duddu and Pulugurtha, 2013; Khan et al., 2018; Sharma et al., 2001).

Based on the observation that the AADT is usually spatially dependent, that is, traffic volumes in segments close to each other are more related than in distant segments (Tobler, 1970), from the 2000s onwards, Geostatistics started to be proposed as a solution to the lack of traffic volume data (Eom et al., 2006). It consists of a set of interpolators that treat spatially dependent variables as random, allowing to incorporate statistical inference in their estimates (Matheron, 1971). Conversely, traditional spatial interpolators, such as inverse distance weighting and nearest neighbor, are deterministic, which means that they are not able to provide uncertainty measures for the calculated estimates (for example, variance and confidence intervals).

Another advantage of Geostatistics is the fact that some of its interpolators do not require additional data to carry out the prediction, and their computational routine is freely available on software such as R (R Core Team, 2021; Pebesma, 2004; Ribeiro Jr. and Diggle, 2016; Ver Hoef, 2018), SGeMS (Remy et al., 2009) and GSLIB (Deutsch and Journel, 1998). Some of these interfaces also allow the user to incorporate modifications into the calculation code. On the other hand, approaches such as the four-step modeling, statistical models and machine learning algorithms strictly depend on explanatory variables, which may not be easily available and require time for collection. In the case of the four-step modeling, the analysis often relies on a costly software, such as the TransCAD. In turn, machine learning algorithms may fail in accounting for the spatial dependence of traffic data (Song and Kim, 2022).

Geostatistics was formerly created to model spatially continuous variables. In spite of that, geostatistical interpolators were expanded, due to their convenience, to areas such as epidemiology, aquaculture, agriculture, forest sciences (Carvalho et al., 2015; Goovaerts, 2009; Kerry et al., 2016; Stelzenmüller et al., 2005), whose variables of interest are spatially discrete.

The lack of data on travel demand variables, which are usually spatially discrete, has led to an increasing number of geostatistical applications to travel demand modeling, with results that represent an important contribution to the planning and operation of transport systems (Gomes et al., 2018; Lindner and Pitombo, 2019; Marques and Pitombo, 2021a; Yang et al., 2018; Zhang and Wang, 2014). Several studies using Geostatistics for spatially estimate travel demand variables can be found in the bibliographic review by Marques and Pitombo (2020). Along these studies, the spatial dependence of travel demand variables is confirmed by the well-structured variograms calculated in the geostatistical modeling step. Even though Geostatistics allows calculating the variable of interest in all geographic coordinates of the database, case studies using travel demand variables seek to obtain an estimate only in the points where the phenomenon occurs, which may be along a road network or bus route, for example.

Geostatistical applications to Annual Average Daily Traffic modeling cover various types of interpolation techniques, commonly called Kriging after Krige (1951), whose pioneering study in mining engineering inspired the first steps into the development of the Geostatistics framework. The interpolators are: Simple Kriging (SK), Ordinary Kriging (OK), Universal Kriging (UK), Regression Kriging (RK), Empirical Bayesian Kriging (EBK), and Spatio-temporal Kriging (STK). Of these models, only UK and RK use explanatory variables, meaning that the remaining ones are univariate interpolators. Table 1 summarizes the studies found in the literature addressing the lack of traffic volume data using Geostatistics.

Table 1: Geostatistical applications to AADT modeling

Source	Number of points	Variographic models	Variable(s)	Methods used	Comments
Eom et al. (2006)	200	Exp*, Gau, and Sph*	AADT	UK and LR	Best results from UK; variation of errors according to the density of counting stations.
Wang and Kockelman (2009)	27738	Exp*, Gau, and Sph	AADT	UK	Best results for the case of intermediate traffic volumes, compared to the low and high-volume cases.
Chi and Zheng (2013)	91	Exp, Gau, Sph, and Cauchy*	AADT / Transport carbon footprint	OK with network distances	Percentage of error variation according to the magnitude of real values (error increases as the real value increases); reduced number of points for calculating the semivariogram.
Selby and Kockelman (2013)	3145, 667, 3017, 1053, 6256, 3532	Exp*, Gau and Sph	AADT	UK, with network and Euclidean distances, GWR and non-spatial model	UK performed best; variation of errors according to the density of counting stations.
Shamo et al. (2015)	4992, 7485, 7734	Exp, Gau and Sph	AADT	SK, OK and UK	No pattern of best technique was found.
Sarlas and Axhausen (2015)	314	Exp*, Gau and Sph	AADT	LR; NBR; SEM (Euclidean, network distance, and network time); SLM (Euclidean, network distance, and network time); GWR; Kriging	GWR and Kriging had the best results in calibration and validation samples, respectively; little difference was seen between kriging and SAR models; network results similar to Euclidean ones; network distances were not used in kriging interpolation.
Kim et al. (2016)	127	Exp, Gau and Sph*	AADT / VMT	HPMS (Highway Performance Monitoring System method), LR, and RK	Best results from RK.

Table 1: Continued…

Source	Number of points	Variographic models	Variable(s)	Methods used	Comments
Klatko et al. (2017)	93, 223, 71, 203, 148, 80, 147, 116, 129, 51	not reported	AADT / VMT	OK, IDW, natural neighbor and trend	Considering only the validation results, no pattern of best technique was found.
Yang et al. (2018)	For training, data collected every 30s from 1 sensor covering a whole day was used	not reported	Traffic volume	STK, historical average and k-nearest neighborhood	Best results from STK.
Song et al. (2019)	627	not reported	Vehicles/(km.day)	IDW; OK; Segment-based OK; LR; UK; RK; Segment-based RK	Best results from SRK for heavy vehicles, and RK for light vehicles; Kriging yields better results compared to non-spatial models (LR) and non-stochastic interpolators (IDW).
Mathew and Pulugurtha (2021)	12899	Exp*, Sph, Power, and Linear	AADT in local roads	LR; GWR; SK, OK, UK, and EBK; IDW; and natural neighbor interpolation	Considering the validation results, GWR and EBK yielded the best results, but GWR performed slightly better. However, the kriging technique does not use predictor data information.

Note: Exp, Gau, and Sph are Exponential, Gaussian, and Spherical semivariogram models. VMT expresses Vehicle Miles Traveled. IDW, LR, NBR, SAR, SEM, SLM, GWR, SK, OK, UK, RK, EBK, and STK stand out, respectively, for Inverse Distance Weighting, Linear Regression, Negative Binomial Regression, Spatial Autoregressive models, Spatial Error Model, Spatial Lags Model, Geographically Weighted Regression, Simple Kriging, Ordinary Kriging, Universal Kriging, Regression Kriging, Empirical Bayesian Kriging, and Spatio-temporal Kriging.

* Model with the best results among the models being compared.

Table 1 shows that geostatistical interpolation of AADT has yielded better results than traditional techniques, such as Linear Regression and Negative Binomial Regression, which do not account for the spatial dependence of traffic data (Eom et al., 2006; Kim et al., 2016; Mathew and Pulugurtha, 2021; Sarlas and Axhausen, 2015; Selby and Kockelman, 2013; Song et al., 2019). Kriging also performed better than other interpolation techniques, such as Inverse Distance Weighting (IDW), natural neighbor, historical average, and k-nearest neighborhood (Mathew and Pulugurtha, 2021; Song et al., 2019; Yang et al., 2018). The potential of kriging to predict AADT in uncounted locations has proven, in some case studies, to be superior to some other spatial models, such as Geographically Weighted Regression (Selby and Kockelman, 2013) and spatial autoregressive models (Sarlas and Axhausen, 2015).

However, most of the studies found rely on the Euclidean distance as the distance measure that helps explain spatial dependence between points in the database. As the traffic flow occurs along a road network, using network distances, rather than straight-line ones, could provide better estimates of AADT in the kriging interpolation (Eom et al., 2006; Wang and Kockelman, 2009). However, Table 1 shows that only two studies applied network distances in the spatial interpolation of AADT (Chi and Zheng, 2013; Selby and Kockelman, 2013), but only Selby and Kockelman (2013) provided a comparison between results from Euclidean and network distances. In this case, the authors concluded that using distances along the road network did not contribute significantly to improving the kriging estimates, and may not compensate the high computational demand process to perform kriging with network distances.

Little difference was seen between results from the two types of distance in the study of Sarlas and Axhausen (2015). Although the authors did not use network distances in kriging, they applied them in the spatial weights matrix of the spatial autoregressive models used.

Network distances have already been tested in the geostatistical modeling of other variables related to travel demand, however there are no consensual results. To assure positive definiteness of covariance matrices, Zou et al. (2012) used an approximate road network distance for modeling urban travel speeds using Universal Kriging. The outcomes were better than those obtained from Euclidean distances. Conversely, little or no improvement was seen in the geostatistical modeling of transit ridership variables, such as boardings per metro station (Zhang and Wang, 2014) and boarding, alighting and loading at the bus stop level or route segment level (Marques and Pitombo, 2021b, 2021c).

Nevertheless, a recent study (Wong and Kwon, 2021) proved that network distances can, in fact, yield better estimates of winter weather collisions using Regression Kriging when compared to the Euclidean results. However, accidents can occur at any point along the road network, while traffic data assumes only one value in the length of a segment. Although collisions and traffic volume are positively related, the former consists of a variable with high occurrence of zeros, and it is much more random in nature than vehicle flow. Moreover, to better capture the weather variability along the road network, the authors divided the road segments into segments of up to 5km. Although better results were achieved in the network distance approach, the AADT modeling, which usually keeps the original length of road segments, could yield even better outcomes, as in this case the difference between network and straight-line distances is higher.

Although using network distances is a promising approach for the spatial estimation of AADT, geostatistical modeling that uses non-Euclidean distances can generate estimates with negative variances (Ver Hoef, 2018). Depending on the theoretical semivariogram model, applying distances along the road network could make it harder to attain positive covariance matrices, which are necessary to obtain positive uncertainty measures when estimating AADT at an unsampled point. However, the studies that addressed the AADT using Geostatistics with network distances (Chi and Zheng, 2013; Selby and Kockelman, 2013) did not carry out any type of inspection to verify the occurrence of this problem and thus circumvent it, if necessary.

The absence of anisotropic analysis is also observed in all cited studies. However, spatial variables may present spatial dependence primarily along a given spatial direction, that is, the spatial behavior varies according to the analyzed direction. Thus, the accuracy of estimates resulting from interpolators that ignore this feature, when it exists, can be negatively affected (Oliver and Webster, 2015). Nevertheless, since, in the network distance approach, the direction under analysis is the network direction, there would not only be a single possible direction between pairs of segments, but numerous ones. Consequently, an anisotropic spatial modeling with network distances may not be feasible.

Based on the problems exposed above, the following research gaps are enumerated:

- **Spatial modeling of AADT using distances along the road network**: the use of network distances to estimate a travel demand variable is still little explored.
- **Anisotropic spatial modeling of AADT**: although using network distances can better represent the spatial behavior of AADT and, consequently, generate better estimates, this approach does not allow the inclusion of anisotropic analysis. Using network distances can be a way to capture the directional variation of AADT, but the usual treatment of anisotropy in Euclidean space may eventually compete against the isotropic method with network distances, given the large computational efforts

normally required to calculate network distances. The anisotropic approach, however, has not yet been addressed in previous studies.

- **Verification of gains from the application of network distances or anisotropy, in comparison with the traditional isotropic approach with Euclidean distances, in the spatial modeling of AADT**.

Figure 1 illustrates the main research gaps and associated justification. In addition, it represents the spatial approaches used in this paper.

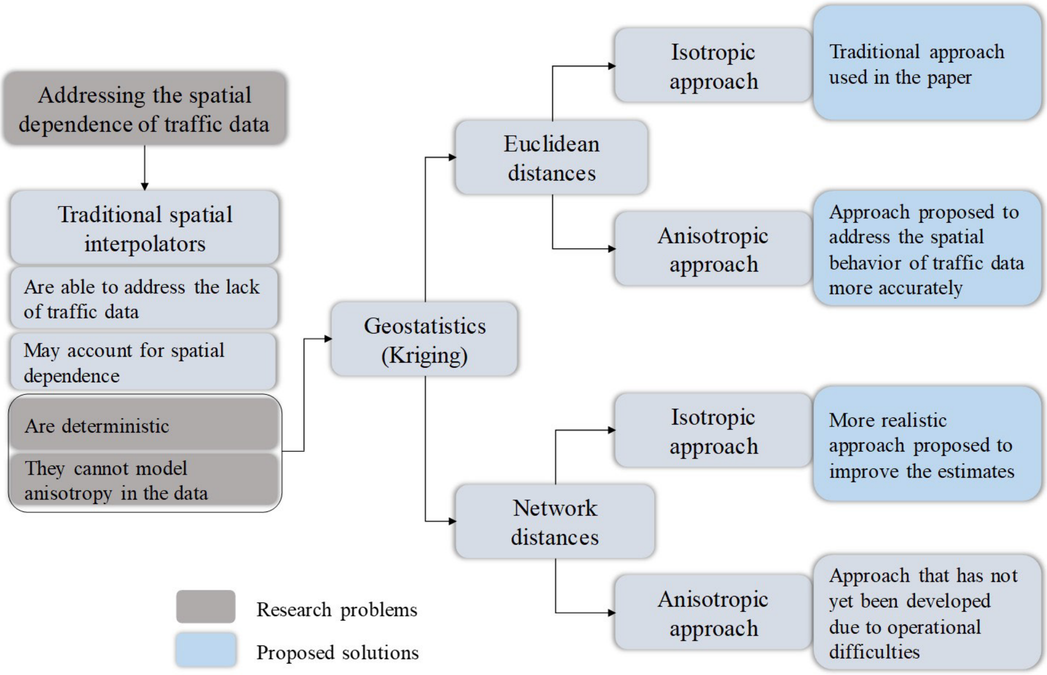

Addressing the spatial dependence of traffic data
Traditional spatial interpolators
Are able to address the lack of traffic data
May account for spatial dependence
Are deterministic
They cannot model anisotropy in the data
Geostatistics (Kriging)
Euclidean distances
Isotropic approach
Traditional approach used in the paper
Anisotropic approach
Approach proposed to address the spatial behavior of traffic data more accurately
Network distances
Isotropic approach
More realistic approach proposed to improve the estimates
Anisotropic approach
Approach that has not yet been developed due to operational difficulties
Research problems
Proposed solutions

Figure 1. Flowchart of proposed methods based on research gaps

Therefore, the main objective of this article is to model the AADT from Ordinary Kriging (OK) with network distances and OK with anisotropy treatment. As a specific objective, the results of these two approaches will be compared with each other and with the results of the isotropic OK with Euclidean distances in a case study focused on the road network in the state of São Paulo - SP (Brazil).

This article is divided into 5 sections. Section 2 highlights the impact of network distances and anisotropy on the spatial estimation of AADT from illustrative examples. The third section details the database used as a case study and the method stages applied. Section 4 shows the results obtained and discusses answers to the title question of the article. The last section summarizes the contributions achieved by the present study and lists suggestions for future research.

2. INFLUENCE OF NETWORK DISTANCES AND ANISOTROPY

Figure 2 illustrates the impact of using network distances on the spatial estimation of AADT.

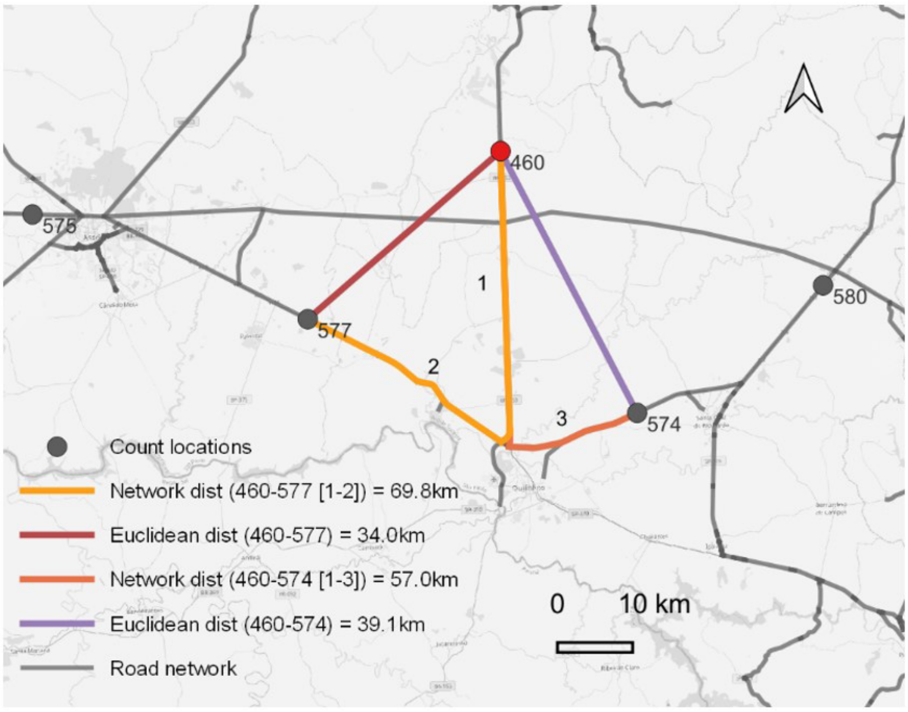

460
575
577
580
574
1
2
3
Count locations
Network dist (460-577 [1-2]) = 69.8km
Euclidean dist (460-577) = 34.0km
Network dist (460-574 [1-3]) = 57.0km
Euclidean dist (460-574) = 39.1km
Road network
0
10 km

Figure 2. Comparison between network and Euclidean distances

Assume that neighbors 574 and 577 will be used to estimate the AADT at point 460. Figure 2 shows that, in addition to the network distance being clearly greater than the straight-line distance, in the exposed case the magnitude pattern between network and Euclidean distances do not hold for the distance pairs between point 460 and its neighbors 574 and 577. Point 577 is 34 km away from point 460, considering the straight-line distance. For point 574, this distance increases to 39 km. In contrast, this pattern is inverted when network distances are used, as the distance between points 460 and 574 is smaller (57 km) than between points 460 and 577 (approximately 70 km).

Thus, in traditional modeling with Euclidean distances, point 577 tends to receive greater weight than point 574 in the AADT estimate at point 460. However, the actual distance between points 460 and 574 is smaller than between 460 and 577. Based on the assumption of spatial continuity of the regionalized phenomenon along the road network and not in a straight line, point 574 should have more weight in the AADT estimate at 460 than point 577.

Although using network distances makes it difficult to incorporate anisotropy in the geostatistical modeling, this approach is also a way to deal with the directional variation of AADT. Regarding anisotropy, there is a main direction in which the spatial continuity of AADT is greater, that is, the spatial dependence, expressed by the semivariogram function, prevails over a greater distance than in the other directions. This distance refers to the range parameter of the theoretical semivariogram. Thus, points along this main direction tend to receive greater weight in the AADT estimate at an unsampled point than other neighboring points.

Figure 3 provides a hypothetical example of the difference between the three approaches that will be compared in the present study. The AADT at the unsampled point in red will be estimated based on four neighbors, equidistant from it in a straight line. The weight assigned to neighbors in each case is expressed by the value λ. The main direction is the vertical axis. Applying these approaches to a real case study is detailed in Section 3.

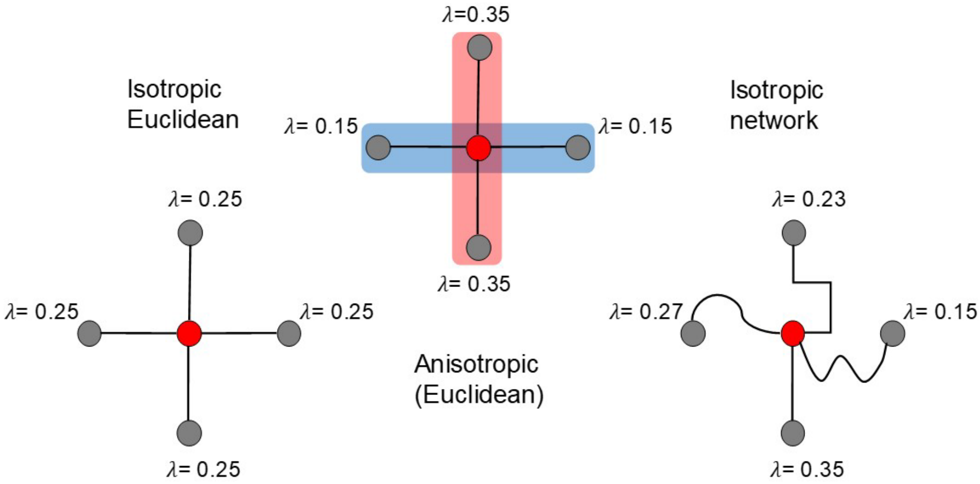


Figure 3. Influence of network distances and anisotropy on the kriging weights

3. MATERIALS AND METHOD

Figure 4 highlights the spatial variability of the Brazilian road network and count locations in 2017. The data necessary for geostatistical modeling were made available by the National Department of Transport Infrastructure (DNIT, *Departamento Nacional de Infraestrutura de Transportes*). There is a greater concentration of highways in the eastern half of the country, mainly in the state of São Paulo, which also has the largest number of count locations. The density of count locations increases significantly in the vicinity of the homonymous capital, which is the most populous city in the country (IBGE, 2021).

In view of the sizeable Brazilian road network (Figure 4) and to limit the study region and enhance the results of the geostatistical modeling, only the state of São Paulo was chosen to participate in the case study. This state was chosen because of its higher density of counting stations and relative ease in acquiring data regarding its road network. In addition, the parameters of the semivariogram models, necessary for calculating the spatial estimates, may be heterogeneous across different geographic units (Wong and Kwon, 2021). Reducing the spatial coverage of the database, for instance AADT, ensures that the models capture local characteristics and, consequently, generate better estimates.

To test the geostatistical potential in estimating the AADT with different network distances and sample quantities, the present case study covered two scenarios: a more restrictive one, that is, with a smaller number of samples; and a more complete one. The database used comprises the spatially distributed Annual Average Daily Traffic (variable of interest), with respective geographic coordinates, and the road layout that connects these points. In this context, the National Department of Transport Infrastructure made available two data sources: 1) shapefile containing the AADT in 742 count locations, for 2017; and 2) spreadsheet with the results of the AADT modeling carried out by DNIT in 2017 covering a total of 5,722 road segments of the National Road Traffic System (SNV, *Sistema Nacional de Viação*).

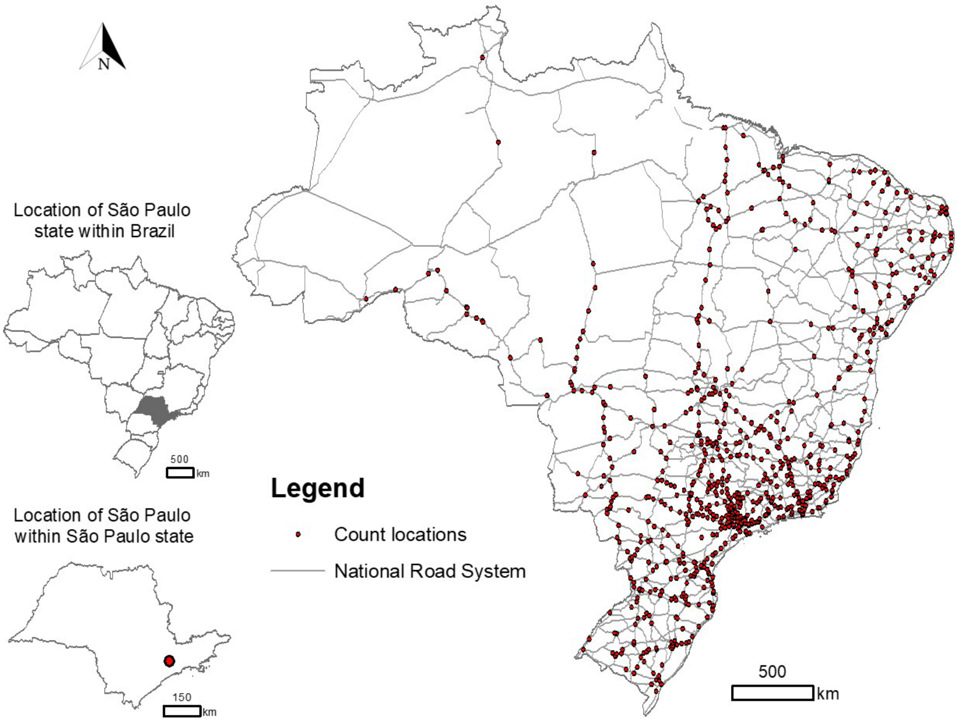

N
Location of São Paulo state within Brazil
500 km
Location of São Paulo within São Paulo state
150 km
Legend
● Count locations
National Road System
500 km

Figure 4. Brazilian road network and count locations in 2017

Of the 742 counting stations, the state of São Paulo had 143 in 2017, while that year, of the 5,722 road segments of federal highways, 339 were from São Paulo. Since Geostatistics deals only with data in the form of points, the AADT information in these 339 segments was assigned to the midpoint of the segment, using, as a basis, the SNV shapefile corresponding to the 2017 DNIT modeling (version 201801B). Since both the DNIT spreadsheet and the SNV shapefile had a common identifier for the segments, it was possible to proceed with the spatialization of the AADT modeled by the DNIT along the reference segments. Thus, the two scenarios analyzed were: 1st) AADT in 143 count locations; 2nd) AADT in 339 road segments, represented by the respective midpoint.

In terms of the road network, three data sources were used to calculate network distances: 1) SNV shapefile corresponding to the 2017 DNIT modeling (version 201801B); 2) kmz file of segments of state highways under concession in São Paulo; and 3) OpenStreetMap road network corresponding to the state of São Paulo. The data from the OpenStreetMap were added to those from the DNIT and the highways under concession until the final layout was equivalent to the map of federal and state highways in the state of São Paulo, available as a Web Map Service (WMS) layer for 2014 on the DataGEO website (https://datageo.ambiente.sp.gov.br/ [Accessed in jun. 2022]). Network distances were calculated in GRASS GIS. The method steps are described in the next subsections.

3.1. Exploratory analysis

The exploratory analysis stage consists of two sub-steps: 1) Spatial dependence verification; and 2) Asymmetry correction. In this context, the spatial files of the 143 counting stations and 339 segment midpoints, with associated AADT, were submitted to an exploratory analysis of spatial dependence based on the calculation of the Moran index

(Moran, 1948). The elements of the spatial weight matrix adopted in the present study were equivalent to the inverse of the network distance between points i and j.

The Moran index helps to anticipate the suitability of the variable of interest to the geostatistical treatment: the closer to 1, the greater the spatial dependence, and the better the results of the spatial interpolators will be. Results close to zero reflect the absence of spatial structure, and the closer to -1, the greater the spatial dispersion/dissociation of the variable of interest.

Subsequently, measures of central tendency and dispersion were calculated for the AADT in both scenarios. Due to their remarkable positive asymmetry, these data were converted to normal distribution by the Box-Cox transformation (Box and Cox, 1964), since Ordinary Kriging assumes normality for the dependent variable. The Moran index and Box-Cox transformation were calculated in the open and free programming tool R (R Core Team, 2021; Millard, 2013; Paradis et al., 2004).

3.2. Minimum eigenvalue analysis

Ver Hoef (2018) showed that the occurrence of negative variances in kriging estimates can be analyzed based on the minimum eigenvalues of the covariance matrices, which are associated with the autocorrelation models selected for spatial interpolation. In the present article, three models were initially adopted: exponential, spherical and Gaussian (Chiles and Delfiner, 2012).

The procedure consists of calculating the eigenvalues for different range values of the theoretical models and applying the network distance matrix of the scenarios considered. Minimum eigenvalue analysis allows identifying range parameter values whose covariance matrices are positive definite. If negative minimum eigenvalues are found for small range values, there is a high chance the autocorrelation model generates negative variances. Consequently, this model is disregarded in the geostatistical modeling stages. If the theoretical semivariogram results in a range greater than the maximum range for which a minimum eigenvalue greater than zero is obtained, this semivariogram model should also be discarded. The geostatistical modeling stages are described in the next subsections.

3.3. Empirical semivariogram

The semivariogram $\gamma(h)$, also known as variogram $2\gamma(h)$, is an important tool to identify the spatial autocorrelation of the sampled values. The empirical semivariogram function is expressed by Equation 1 (Cressie, 1993):

$$\gamma(h) = 1/2N(h) \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (1)$$

where, $Z(x_i)$ is the variable of interest at geographic position x_i, h is the distance between pairs of points, and $N(h)$ represents the number of pairs situated at a distance h. The search for pairs to calculate the empirical semivariogram is performed based on five main parameters: direction (φ, angle measured from the horizontal axis in a counterclockwise direction), angular tolerance ($\Delta\varphi$), lag (h), lag tolerance (Δh) and maximum width (Oliver and Webster, 2015). When considering the phenomenon as isotropic, that is, that the spatial variation occurs in a similar way in all directions, the omnidirectional semivariogram is calculated. In this case, all pairs of points located at a distance $h \pm \Delta h$ are selected, regardless of their direction. However, if the variable presents anisotropy, the pairs of points are limited to the $\varphi \pm \Delta\varphi$ directions.

The empirical semivariogram was calculated for the two scenarios described above and considering three different situations: 1) anisotropic phenomenon with Euclidean distances; 2) isotropic phenomenon with Euclidean distances; and 3) isotropic phenomenon with network distances. The presence of anisotropy was identified based on an exploratory analysis, in which the empirical semivariogram is calculated for several different directions.

3.4. Theoretical semivariogram

To perform the kriging interpolation, theoretical models of semivariograms are fitted to the empirical semivariograms calculated. The theoretical models are characterized by three parameters: nugget effect (C_0), which corresponds to the semivariance for very small distances and reflects the spatial randomness of the phenomenon. It may also represent lack of information or sampling error; partial sill (C), which is termed as the spatial variance between the points; and range (A/a), which represents the distance from which there is no more autocorrelation between the points (Matheron, 1971). For this article, we tested the semivariogram models corresponding to the autocorrelation models presented in subsection 3.2. They are: exponential (Equation 2), spherical (Equation 3) and Gaussian (Equation 4) (Olea, 2006):

$$\gamma(h) = C_0 + C[1 - exp(-h/a)] \tag{2}$$

$$\gamma(h) = \begin{cases} C_0 + C[1.5(h/a) - 0.5(h/a)^3] \; if \; h < a \\ C_0 + C \; if \; h \geq a \end{cases} \tag{3}$$

$$\gamma(h) = C_0 + C[1 - exp(-(h/a)^2)] \tag{4}$$

If the semivariograms present different values of range and/or partial sill for different spatial directions, it is concluded that the variable of interest exhibits anisotropy. After verifying that the range and sill vary as a function of the direction, obtaining an isotropic semivariogram that accounts for this characteristic is performed as described in the following subsection.

3.5. Anisotropy

Geometric (range) anisotropy is accounted for by modeling the range as the axes of an ellipse. The coordinate system, whose axes are equivalent to the main and secondary directions, is initially used. The main direction is the one with the greatest range and the secondary direction is assumed to be perpendicular to the main direction (Eriksson and Siska, 2000). The anisotropic semivariogram is converted to isotropic through a process of rotation and stretching/shrinking of the initial axes, so they coincide with the axes of the coordinate system in which the variable is measured (Deutsch and Journel, 1998; Eriksson and Siska, 2000; Isaaks and Srivastava, 1989).

When both range and partial sill vary with spatial direction, an isotropic semivariogram is initially calculated in the direction of lowest partial sill (C_1), represented by the angle θ, which is also the direction with the greatest range. A second structure is added to this structure, with a partial sill equal to the difference between the highest and lowest partial sills (C_2), and with anisotropic range. Thus, the exponential model takes the form shown in Equation 5:

$$\gamma(h,\varphi) = C_0 + C_1[1 - exp(-h/a_1)] + C_2\big[1 - exp\big(-h/a_\varphi\big)\big] \tag{5}$$

where a_1 is the isotropic range of the first structure. For the second structure, the range along the lowest sill direction (θ) is considered to be extremely large so that the term $(h/a_\varphi) \rightarrow 0$ when the semivariogram function approaches the θ direction. Thus, the influence of the second structure on the semivariogram function is negligible in the θ direction, but exhibits its maximum contribution in the highest sill direction (Eriksson and Siska, 2000).

Since the second structure has geometric anisotropy, the range a_φ is calculated based on Equation 6 (Eriksson and Siska, 2000):

$$a_\varphi = h/\sqrt{b_1 \Delta x^2 - b_2 \Delta x \Delta y + b_3 \Delta y^2} \tag{6}$$

where $b_1 = (cos\ \theta/a_{max})^2 + (sin\ \theta/a_{min})^2, b_2 = 2sin\ \theta cos\ \theta\big(1/a_{min}^2 - 1/a_{max}^2\big)$, and $b_3 = (sin\ \theta/a_{max})^2 + (cos\ \theta/a_{min})^2$, Δx and Δy are, respectively, the horizontal and vertical components of the distance vector h in the original data coordinate system. In this case, a_{min} is the range in the θ direction and a_{max} was assumed as 10^{30}.

3.6. Ordinary Kriging

Ordinary Kriging estimates are given by Equation 7 (Matheron, 1971):

$$Z_{x_0}^* = \sum_{i=1}^{n} \lambda_i\, Z(x_i) \tag{7}$$

where $Z_{x_0}^*$ is the attribute estimated at point x_0; λ_i are kriging weights; and $Z(x_i)$ is the observed value of the variable Z at the i-th point. Thus, it is a linear combination of sampled neighbor values associated with optimal weights.

The OK weights are calculated by a system of equations, represented in matrix form according to Equation 8 (Cressie, 1993):

$$\begin{bmatrix} \gamma(x_1 - x_1) & \gamma(x_1 - x_2) & \dots & \gamma(x_1 - x_n) & 1 \\ \gamma(x_2 - x_1) & \gamma(x_2 - x_2) & \cdots & \gamma(x_2 - x_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(x_n - x_1) & \gamma(x_n - x_2) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(x_0 - x_1) \\ \gamma(x_0 - x_2) \\ \vdots \\ \gamma(x_0 - x_n) \\ 1 \end{bmatrix} \tag{8}$$

where the values of $\gamma(x_n - x_n)$ represent the modeled semivariance of the n-th sample with respect to itself; the values of λ_n are the optimal weights of kriging; μ is the Lagrange multiplier; and $\gamma(x_0 - x_n)$ is the modeled semivariance between the n-th sample and the point to be estimated x_0. Therefore, the weight values are obtained from the product of the inverse of the semivariance matrix between the sampled points $[K]$ and the semivariance matrix between the point to be estimated and its sampled neighbors $[M]$ (Equation 9):

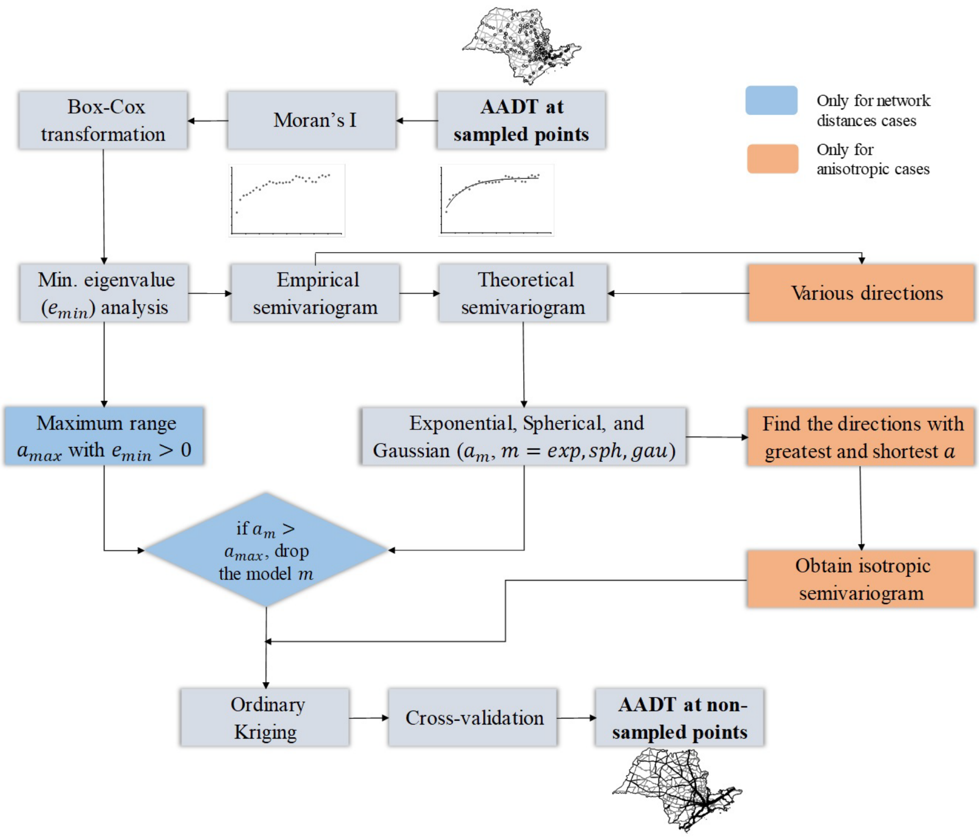
$$[\lambda] = [K]^{-1} \cdot [M] \tag{9}$$

3.7. Cross-validation

Finally, in order to evaluate the performance of Ordinary Kriging, a cross-validation step is carried out, known as leave-one-out (Cressie, 1993). The procedure is performed by removing each sampled point from the database and this point is estimated from the neighboring values according to the method described in the previous subsections. It is then possible to compare the observed values to those estimated by OK based on several goodness-of-fit measures, such as: Absolute Percentage Error (APE), Root Mean Squared Error (RMSE) and Pearson's linear correlation coefficient (R) (Hollander and Liu, 2008).

In addition, the nonparametric Mann-Whitney test was applied to the comparison between real and predicted values. The null hypothesis states that the two samples are from populations with the same distribution function. Therefore, for a 95% confidence level, we can accept the null hypothesis if the resulting p-value is greater than 0.05.

To illustrate the functionality of Ordinary Kriging, 27 unsampled segments of the National Road Traffic System had their AADT estimated from sampled neighbors located within the range region. This example was carried out using data from the second scenario since it has a greater number of samples. Figure 5 summarizes the method used in the present article.



Box-Cox transformation
Moran's I
AADT at sampled points
Only for network distances cases
Only for anisotropic cases
Min. eigenvalue (e_{min}) analysis
Empirical semivariogram
Theoretical semivariogram
Various directions
Maximum range a_{max} with $e_{min} > 0$
Exponential, Spherical, and Gaussian (a_m, $m = exp, sph, gau$)
Find the directions with greatest and shortest a
if $a_m >$ a_{max}, drop the model m
Obtain isotropic semivariogram
Ordinary Kriging
Cross-validation
AADT at non-sampled points

Figure 5. Flowchart of methodological steps (a is the range parameter)

3.8. Validation analysis

To attest to the robustness of the proposed method in estimating missing traffic data, the datasets of the first and second scenarios were split into two samples: 1) calibration, with 70% of the original database; and 2) validation, with the remaining 30%. In this paper, calibration samples refer to the points used to calculate the empirical semivariogram and fit a model to it. The validation samples are the remaining points whose values are estimated based on the theoretical semivariogram adjusted previously. Estimates were calculated in both sampled (calibration) and missing (validation) points. Only network distances, coupled with a permissive semivariogram model, were used at this stage.

The calibration samples were randomly selected based on the density of points in the original database using the R package “spatialEco” (Evans, 2021). Isotropic modeling with Euclidean and network distances was performed using the open-source programming interface R (R Core Team, 2021; Ver Hoef, 2018), while anisotropic modeling was performed using the Isatis software. The results of these three approaches for the two considered scenarios are presented in Section 4.

4. RESULTS AND DISCUSSION

Figure 6 shows the spatial and frequency distribution of the AADT along the 143 counting stations (on the left) and 339 midpoints of federal road segments in the state of São Paulo (on the right). The road network laid out corresponds to state and federal highways in São Paulo.

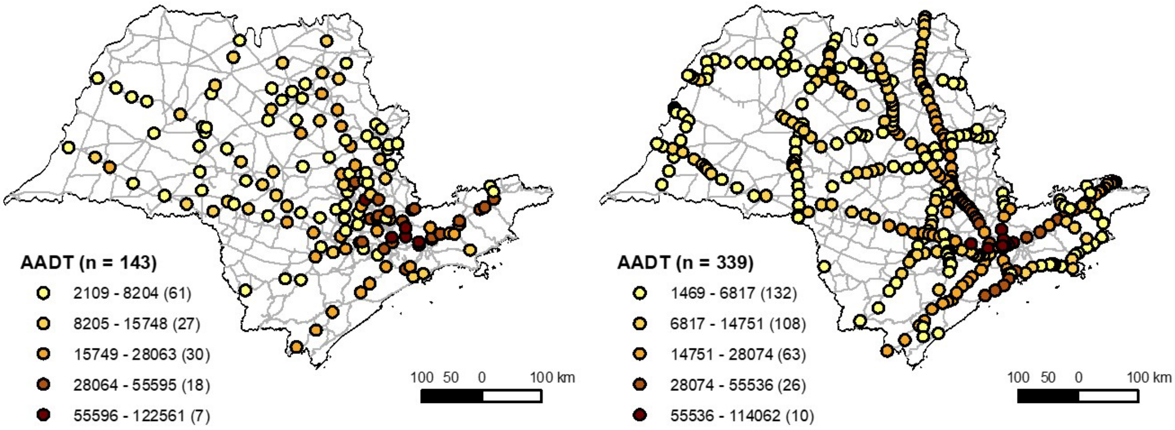


Figure 6. Spatial distribution of AADT in both scenarios

Since the AADT in the 339 reference segments of the National Traffic System is based on the information from the 143 traffic counting stations, both maps have a similar pattern, showing the highest traffic volumes in the segments closest to the state capital. From the city of São Paulo, the AADT gradually decreases along the northwest direction. The main flows of goods and people are in the northwest direction due to a historical development policy in the interior of the state (Souza and Silveira, 2009).

The high amplitude of traffic volume variation represents the inclusion of different values of the dependent variable in the modeling and, therefore, a more comprehensive analysis. The spatial pattern of AADT also reveals that the variable of interest demonstrates spatial autocorrelation, that is, points close to each other have more similar traffic volume than distant points. In fact, the Moran index obtained with the AADT values of the first and second scenarios was 0.157 and 0.201, respectively, both with statistical significance ($pseudo\ p\ value = 0$).

The frequency of points in each AADT category confirms the positive asymmetry discussed in Section 3. Both scenarios have a mean greater than the median: in the first case, the mean and median are equivalent to 17,682 and 10,532 vehicles, respectively; in the second case, these values are 13,703 and 8,651 vehicles. To correct the asymmetry of the variables of interest, the power of the Box-Cox transformation was 0.2160749, in the first scenario; and 0.2149832 in the second.

Then, the geostatistical modeling of AADT was carried out in the two scenarios considering three different approaches: anisotropic with Euclidean distances, isotropic with Euclidean distances and isotropic with network distances. However, the possibility that negative variances occur in the kriging estimates with network distances was verified before this step, based on the analysis of minimum eigenvalues. The graphs in Figure 7 show the eigenvalues' variation due to the spatial arrangement of the two scenarios and the range parameter, considering the three theoretical models of autocorrelation typically used: exponential, Gaussian and spherical.

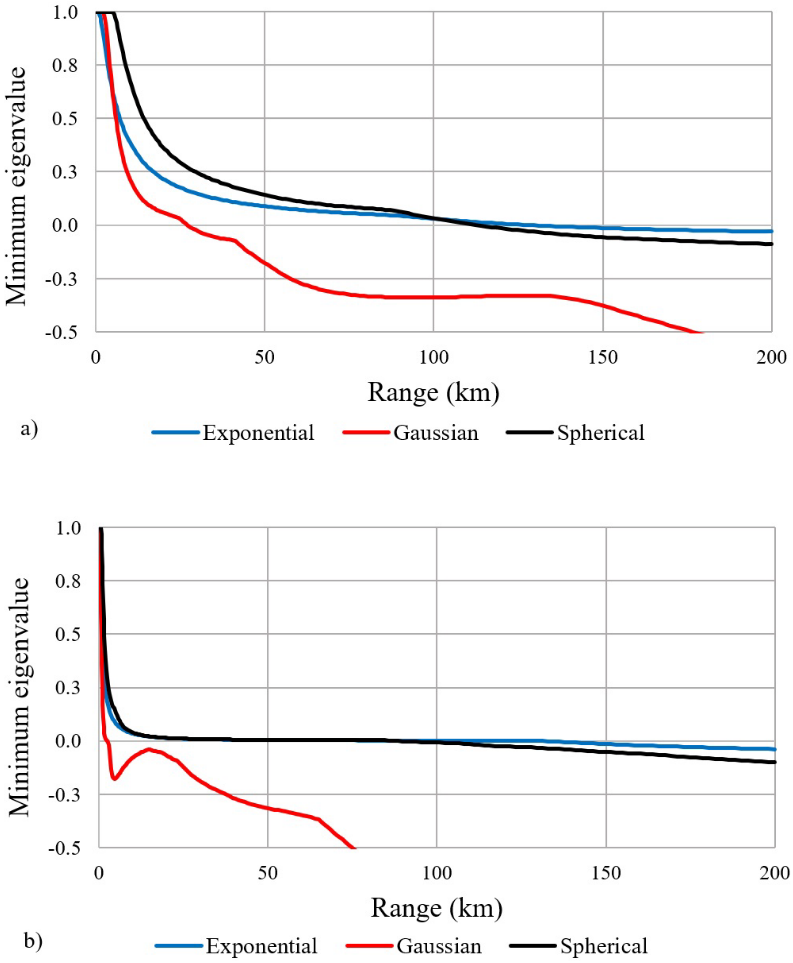


Figure 7. Minimum eigenvalues for the first (a) and second (b) scenarios

In both scenarios the Gaussian semivariogram reaches negative minimum eigenvalues quickly at small range values. Thus, only exponential and spherical semivariograms were used in the geostatistical modeling stage. For the exponential model, negative eigenvalues arise from an approximate range of 130 km in both scenarios, while the spherical semivariogram presents this problem at ranges from 115 km, in the first case, and 90 km, in the second. However, the spherical semivariogram, adjusted to the points of the empirical semivariogram with network distances, resulted in range parameters greater than 200 km. Therefore, the kriging stage was carried out only with the exponential model, which presented calibrated range parameters smaller than 100 km in both cases with network distances, eliminating the possibility of negative variances in its kriging estimates.

Figure 8 shows the empirical semivariograms of the anisotropic approach in the two scenarios considered, together with the respective exponential model and calibrated parameters. The angles are given in azimuths.

The different semivariograms found for the directions of greater and lesser spatial continuity (Figure 8a and Figure 8b) show that there is, in fact, anisotropy in the spatial distribution of AADT. The main direction in the first and second scenarios was 160º and 0º, respectively, while the secondary direction was 250º and 90º. Figure 9 presents the isotropic semivariograms of the first scenario.

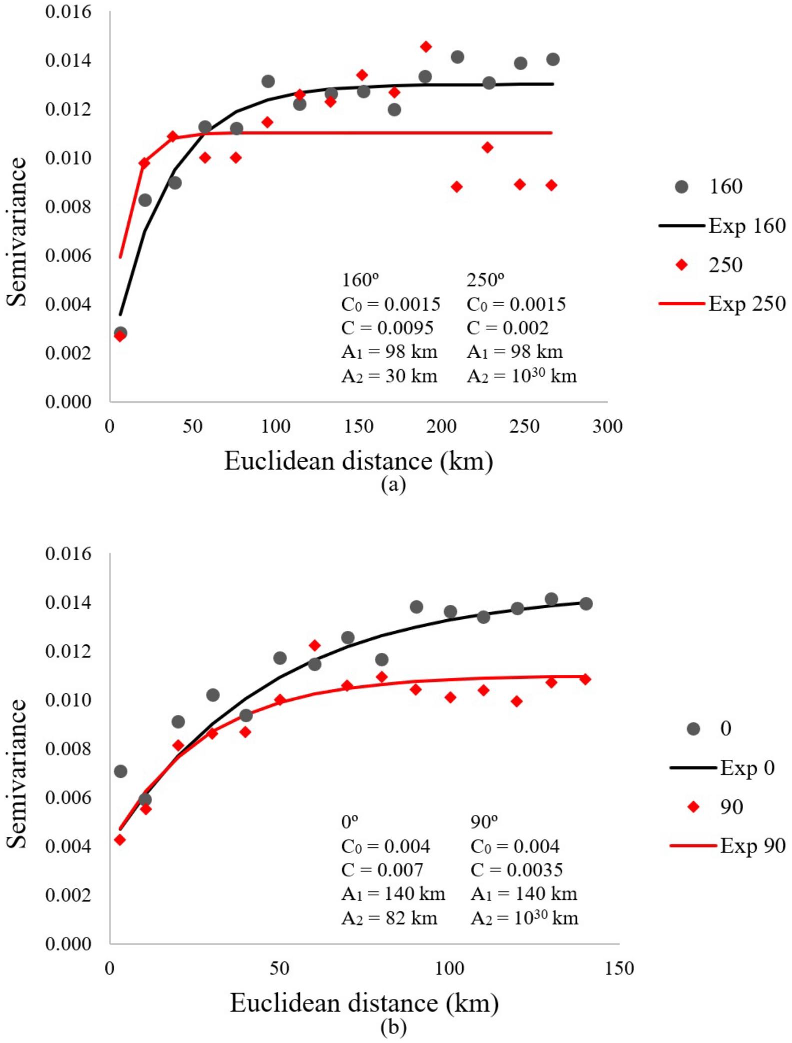
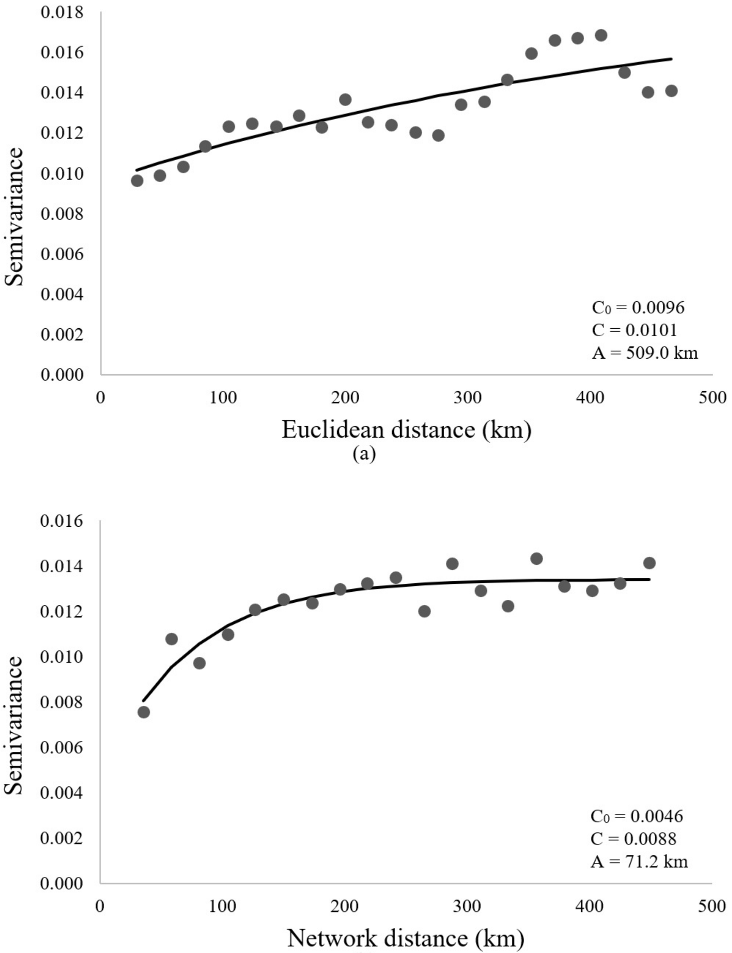


Figure 8. Semivariograms of the transformed AADT in the main and secondary directions of the first (a) and second (b) scenarios



(a)

(b)

Figure 9. Semivariograms of the transformed AADT with Euclidean (a) and network (b) distances in the first scenario

The calibrated exponential model, in the isotropic case with Euclidean distances of the first scenario, resulted in a range greater than 500 km, as a way to minimize the error between estimated and observed values. A range greater than 500 km would not be allowed in network distances, since it would be pointing to the occurrence of negative variances in the kriging estimates. This result may again indicate a better suitability of network distances compared to the traditional one. The isotropic semivariograms of the second scenario are shown in Figure 10.

When comparing the isotropic semivariograms of the second scenario, which turned out to be a little more similar to each other, there is a greater range in the case with network distances. The ratio between both ranges is of 1.66, which is greater than 93% of the ratios between network and Euclidean distances from all pairs of points in the second scenario. A range parameter significantly greater for the network distance case may be revealing that this approach more accurately reflects the spatial continuity of the phenomenon under analysis.

With regard to isotropic approaches, a better-defined spatial structure is clearly perceived in cases that use network distances, which points to a more adequate representation of the phenomenon when network distances are prioritized. Empirical semivariograms with network distances are smoother and exhibit much less fluctuation than those with Euclidean distances. The difference between the two types of distance is more noticeable in the first scenario (with 143 observations). Furthermore, the percentage of nugget effect relative to the sill, which reflects the spatial discontinuity of the variable of interest, is smaller in cases with network distances.

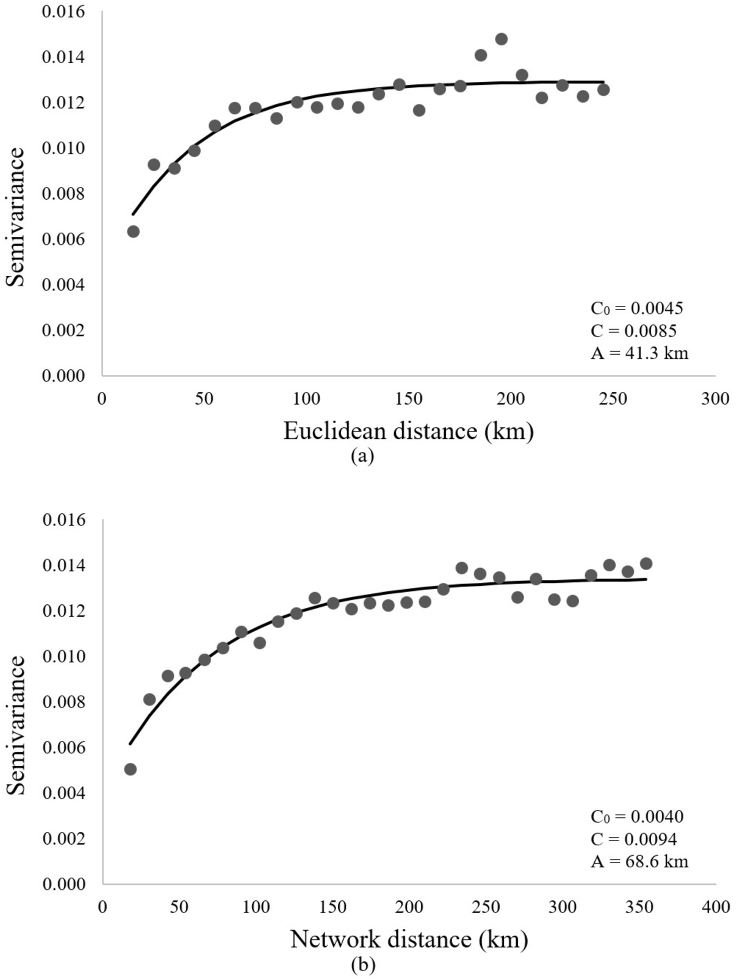


Figure 10. Semivariograms of the transformed AADT with Euclidean (a) and network (b) distances in the second scenario

Table 2 shows the results of the goodness-of-fit measures applied to the comparison between original values and those estimated by cross-validation. Although the semivariograms were calculated and fitted based on the transformed values, the cross-validation estimates were recalculated by applying the inverse transformation. Thus, the estimated values returned to the same scale as the originals.

Table 2: Goodness-of-fit measures

n	Case	Type of distance	MedAPE	RMSE	R
143	Anisotropic	Euclidean	44.750%	14,439.743	0.654*
	Isotropic	Euclidean	46.847%	16,016.041	0.617*
	Isotropic	Network	38.055%	13,986.530	0.709*
339	Anisotropic	Euclidean	27.570%	9,597.729	0.775*
	Isotropic	Euclidean	33.150%	10,112.490	0.793*
	Isotropic	Network	26.691%	9,145.024	0.825*

Note: MedAPE, RMSE and R express, respectively, the Median of Absolute Percentage Error, Root Mean Squared Error and Pearson's linear correlation coefficient between original and estimated values.
* is statistically significant at a 99% confidence level (p < 0.01, one-tailed).

As expected, the anisotropic modeling estimates were better than the isotropic approach with Euclidean distances in both scenarios. However, although not being able to deal with the anisotropy of the phenomenon, the isotropic estimates with network distances were superior to both the anisotropic modeling and the omnidirectional modeling with Euclidean distances. This result shows that when considering the real path traveled between two sample points, the approach with network distances better represents the spatial behavior of the traffic volume than the one that addresses only the directional variation of the AADT.

The most notable difference refers to the median of the absolute error in percentage of the first scenario, which decreases by 15% from the anisotropic case to the network distance case. Reductions in the RMSE of both scenarios and in the MedAPE of the second scenario range between 3% and 5%. However, the better performance of the network distances approach compared to that of Euclidean distances with isotropic semivariogram is emphasized: in both scenarios, the MedAPE is reduced by approximately 20%, while the RMSE is reduced by 13% and 10%, respectively, in the first and second scenarios. Furthermore, Pearson's linear correlation coefficient shows an increase in the proportionality of the original and estimated values. The 137% increase in the number of sampling points, from 143 to 339, also contributes to improving the estimates in all cases analyzed. For example, in the network distance approach, MedAPE and RMSE were reduced by 30% and 35%, respectively, and R increased by 16%.

A MedAPE of 38% in the first scenario and 27% in the second, indicates that half of the database, that is, 71 and 169 points, respectively, had an absolute error in percentage lower than 38% and 27%. Considering that OK requires only the value of the variable of interest in spatial points, with respective geographic coordinates, this result is quite satisfactory. Chi and Zheng (2013), who also used OK with network distances to estimate the carbon footprint as a function of AADT, attained an absolute mean error of 76.11%, while this metric resulted in 41.37% and 54.17% in the present study, both considering network distances for the second and first scenarios, respectively. These errors are also smaller than five of the six case studies carried out by Selby and Kockelman (2013), whose variable of interest was also the AADT. Notwithstanding the explanatory variables, in five of the six situations analyzed and considering a validation sample, the authors found mean absolute errors varying between 55.8% and 63.1% with Euclidean distances, and 55.9% and 62.4% with network distances. This pattern was repeated in the calibration sample in two cases shown by the authors. Furthermore, the use of network distances had little or no positive influence on the results: the greatest reduction in the APE error was of approximately 3%. Zhang and Wang (2014) also compared a multivariate geostatistical approach with network and Euclidean distances, and observed an improvement of only 0.74% in the goodness-of-fit measure used (the adjusted R^2 of the models). In contrast, the results shown in Table 2 point to improvements of 10% up to 20% in the applied error metrics.

The absence, or little difference observed in the results of Selby and Kockelman (2013) and Zhang and Wang (2014), is probably due to the use of several explanatory variables in geostatistical modeling. In this case, it is assumed that the semivariogram is present in the residuals, and the excess of covariates can blur the spatial structure of the errors, making it difficult to accurately model the semivariogram.

Although the exponential semivariogram resulted as the only permissive model for geostatistical modeling, Table 1 showed that most of the studies that compared the fit of several variographic models demonstrated better performance of the exponential model. This result suggests that the exponential model may be the most suitable for modeling AADT.

The maps in Figure 11 show the spatial variation of the Absolute Percentage Error of the network distance approach in both scenarios. Most of points have an error of up to 60%, for the case with 143 points, and 30%, for the scenario with 339 points. The positive asymmetrical profile of the frequency of points by category again confirms the satisfactory results of OK.

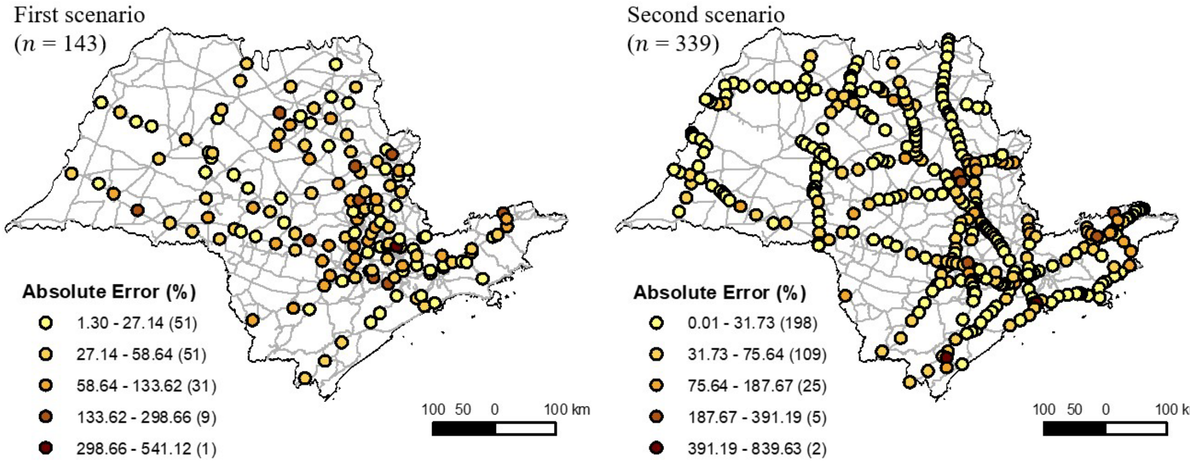


Figure 11. Estimate errors of network distances for both scenarios

However, extreme errors are observed in the last two categories probably due to the great amplitude of AADT in both cases. Since OK assumes a constant local mean of the variable of interest, it fails to adequately model possible database outliers. However, this problem can be addressed by using robust semivariogram estimators (Cressie and Hawkins, 1980), which smooth the elongated tail of skewed distributions; or by including explanatory variables to the geostatistical modeling, such as: highway category, number of lanes, speed limit, population and jobs within a certain distance, among others.

The outlier problem can also be seen in Figure 12, which presents the number of points whose AADT value fell within the defined intervals. In the first scenario, predictions were limited to the upper bound of 70,000 vehicles, but three points had an observed AADT higher than 70,000. In the second scenario, two points had an AADT higher than 90,000 vehicles. However, none of the approaches used predicted values beyond this threshold.

Figure 12 also confirms a tendency already shown in Table 2: results from network distances are clearly better than the Euclidean ones in the first scenario, but, in the second scenario, the anisotropic approach is seen as a competitive alternative. The frequency of estimates from network distances in the first scenario is closer to the real AADT frequency in most intervals up to 70,000. In the second scenario, the anisotropic approach had the best adjustment to the observed AADT frequency.

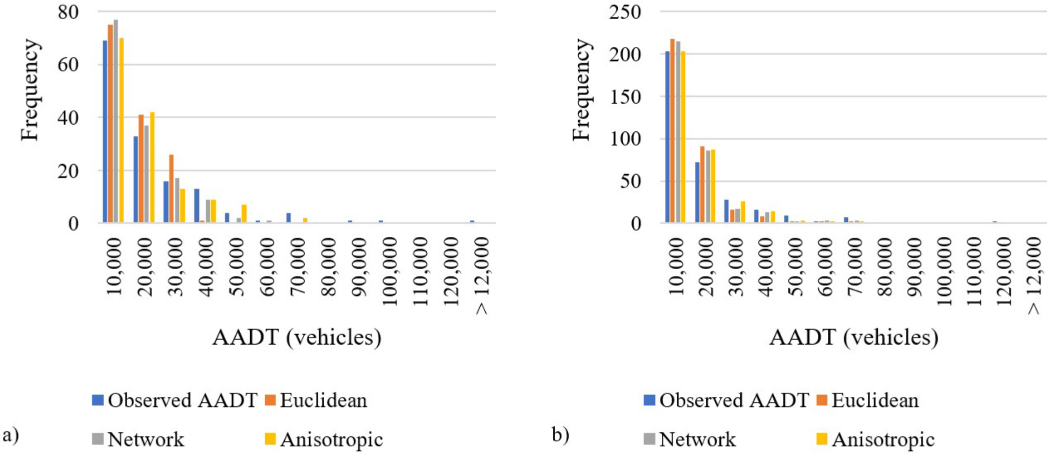


Figure 12. Frequency of points in each AADT interval for scenarios 1 (a) and 2 (b)

In both scenarios, results from the anisotropic approach were very close to the real AADT frequency for volumes up to 10,000 vehicles. This result suggests that addressing the anisotropy of traffic data can be an interesting alternative for modeling and predicting low volume categories.

Figure 13 illustrates the functionality of kriging by showing, on the left, the 339 segments used in the geostatistical modeling of the second scenario (and respective original AADT) and, on the right, the 339 initial segments plus 27 unsampled segments of the National Road Traffic System, all with AADT estimated based on OK with exponential model and network distances. The segments in red could not be estimated due to the absence of neighbors within the range region.

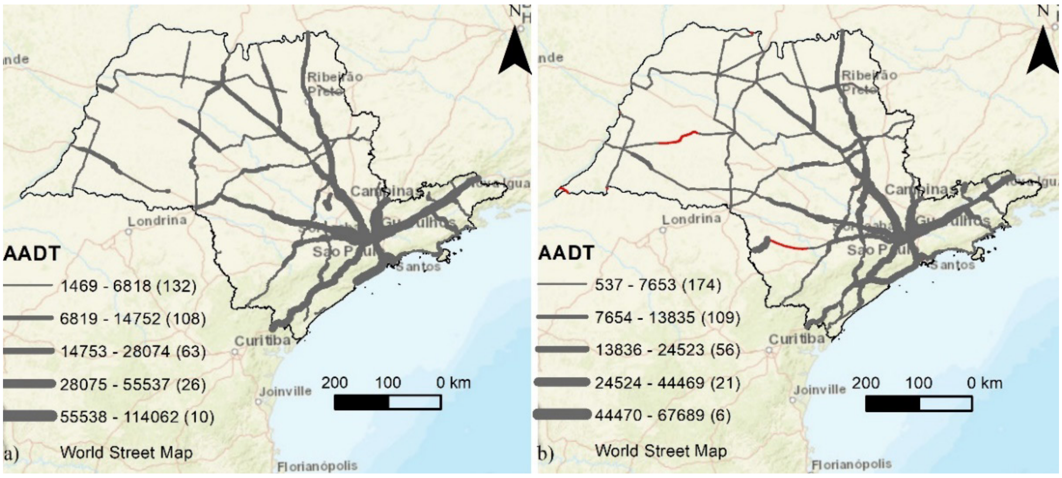


Figure 13. (a) Original AADT in 339 segments; (b) Estimated AADT in 366 segments. World Street Map source: Esri, HERE, Garmin, USGS, Intermpa, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, © OpenStreetMap contributors, and the GIS User Community.

4.1. Validation results

Figure 14 shows the spatial distribution of the calibration and validation samples of the first and second scenarios, together with the resulting empirical and theoretical semivariograms. In the first scenario, the calibration and validation samples had 100 and

43 points, respectively, while in the second one, the calibration and validation samples covered 237 and 102 road segments.

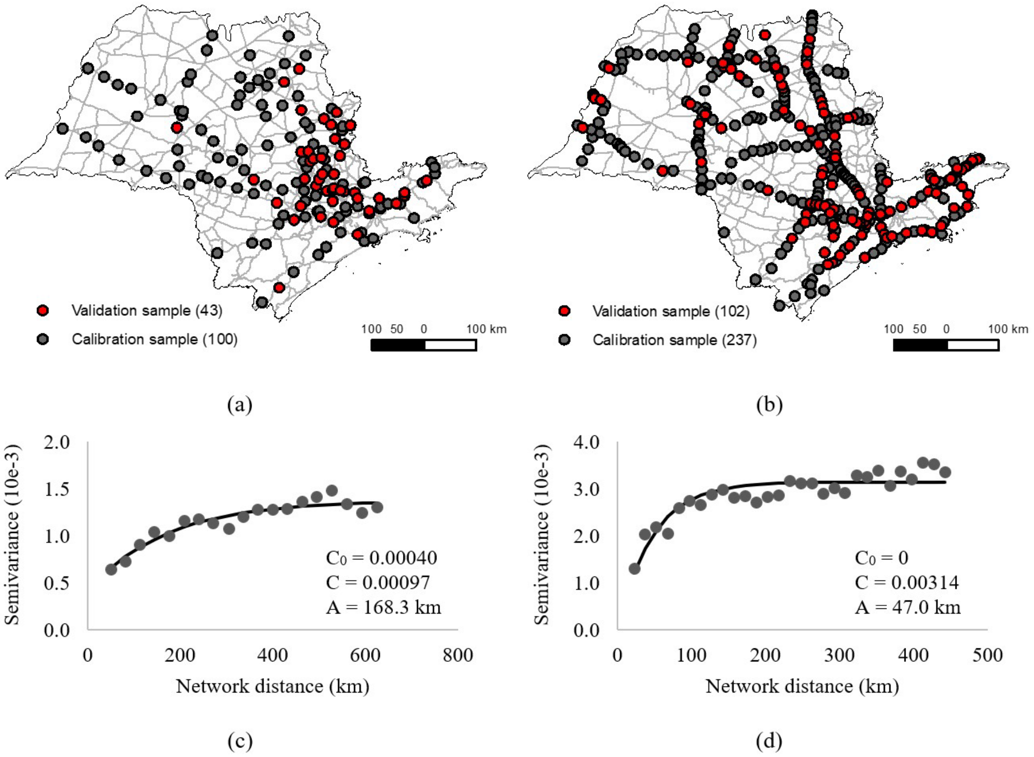


Figure 14. (a) and (b) samples from the first and second scenarios, respectively; (c) and (d) Semivariogram obtained from the calibration sample of the first and the second scenarios

For the validation analysis, each method step of the network case (Figure 5) was applied to the calibration samples of both scenarios and only the exponential model was used. In turn, Ordinary Kriging estimates were calculated through cross-validation for the calibration points. For the validation points, predictions were computed using the models obtained from the calibration samples and Equations 7-9. Table 3 summarizes the validation step results.

Table 3: Results from each method step in the validation analysis

Measure	Sample	1st scenario	2nd scenario
Moran index	Calibration	0.180*	0.228*
Lambda (Box-Cox transformation)		-0.341	-0.298
Maximum range with minimum eigenvalue bigger than zero (km)		432.180	144.500
Actual range (km)		168.323	47.004
MedAPE		33.789%	22.786%
	Validation	42.491%	22.793%
RMSE	Calibration	13,161.542	8,888.296
	Validation	18,383.077	9,535.776
R	Calibration	0.733**	0.820**
	Validation	0.453**	0.772**

Note: MedAPE, RMSE and R express, respectively, the Median of Absolute Percentage Error, Root Mean Squared Error and Pearson's linear correlation coefficient between original and estimated values.* pseudo p-value equals 0. ** is statistically significant at a 99% confidence level (p < 0.01, one-tailed).

Since the range parameters were smaller than the maximum range which obtained a minimum eigenvalue bigger than zero, the result of the exponential model was permissive for both scenarios. Using the calibration samples of 100 and 237 points yielded better goodness-of-fit measures compared to their original database (Table 2), probably due to the higher spatial autocorrelation found in the samples, as shown by the Moran index results.

Results from the validation sample of the second scenario were also consistently better than the results from the first scenario using the complete database, and the anisotropic case of the second scenario (Table 2). The performance of the validation sample of the first scenario was comparable to the Euclidean approaches of the first scenario complete database.

Figure 15 shows the frequency of points in each AADT interval for the validation samples. Good results from the first scenario can be seen in the AADT range from 20,000 to 40,000 vehicles. In the second scenario, the first range (from 0 to 10,000 vehicles) and the range from 30,000 to 40,000 vehicles also showed good results.

For all scenarios considered, using both the complete database and the calibration/validation samples, the p-value from the Mann-Whitney test was greater than 0.05. Thus, the null hypothesis of the same distribution between observed and predicted values is accepted in all cases analyzed. The following subsection provides support for the decision-making on whether to use network distances or anisotropy.

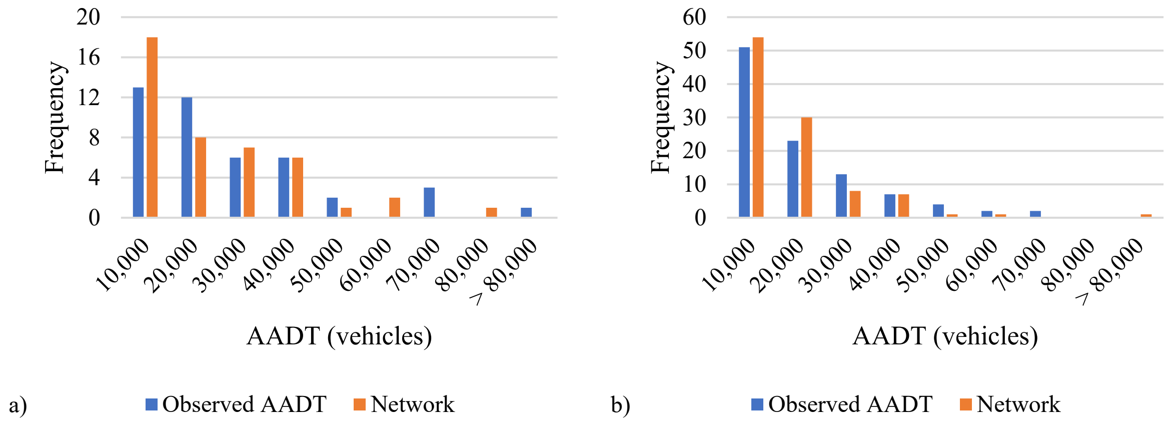


Figure 15. Frequency of points in each AADT interval for the validation samples of the first (a) and second (b) scenarios

4.2. Should we account for network distances or anisotropy in the spatial estimation of AADT?

The boxplots in Figure 16 illustrate the differences in the ratio between network and Euclidean distances in scenarios 1 and 2. Figure 16a was prepared based on all pairs of points in each scenario. In Figure 16b, only the pairs whose distance was within the range region were selected, that is, which presented spatial autocorrelation and had influence on the estimate of the neighboring point in the cross-validation step.

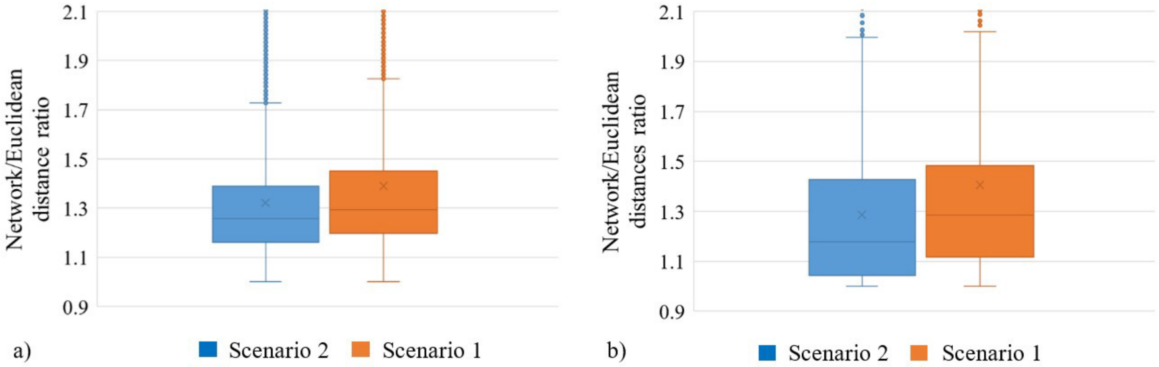


Figure 16. Ratio between network and Euclidean distances considering all pairs of points (a) and only pairs of points less distant than the range parameter (b)

As expected, the descriptive measures are smaller for the second scenario compared to the first, especially with regard to the points that can be selected for kriging. The increase in the density of points available for spatial interpolation tends to reduce network distances and make them closer to Euclidean distances. In the first scenario, there were 7.16 count locations per 1,000 km of highways, while in the second scenario, the segment midpoints represented 16.97 locations/1,000 km.

When approximating the network and Euclidean distances, the kriging estimates obtained by the two approaches also start to be more similar. Case studies using urban travel demand variables (boarding, alighting and loading along bus lines) have resulted in little or no improvement in the use of network distances compared to the traditional Euclidean distance (Marques and Pitombo, 2021b, 2021c). The study by Zhang and Wang (2014), whose variable of interest was the number of boardings at subway stations, also did not show considerable positive differences in the use of network distances. However, the authors used a multivariate interpolator and the independent variables were able to explain almost all the variance of boardings, leaving little variance for the spatial interpolation. Thus, little improvement was observed in the spatial model compared to the non-spatial model. When kriging is applied to the residuals of a multivariate model, it is possible that the explanatory variables can incorporate the spatial dependence of the regionalized phenomenon and account for a significant part of the variance of the dependent variable. Thus, the differences between the use of network and Euclidean distances are in fact subtle. This may also have been one of the reasons responsible for the lack of improvement in the use of network distances in the study of Selby and Kockelman (2013).

As the density of data on traffic volumes along the network is lower than that of urban variables, AADT presents better estimates when using network distances. However, Table 2 shows that the improvement observed in the second scenario is smaller than in the first scenario. Thus, the contribution of anisotropy appears to be competitive with the use of network distances. As reported by some authors (Selby and Kockelman, 2013; Wong and Kwon, 2021), obtaining network distance matrices is computationally expensive, and also requires a perfectly connected network without topological faults to calculate the shortest paths. The results found in the present article, together with those of previous studies, suggest that, as the density of count locations increases, incorporating anisotropy may be a viable alternative to the use of network distances, if the main intention is only to provide accurate estimates. In this context, the greater the number of sampled segments, the greater the order of the network distance matrix, and its calculation can become a barrier to spatial interpolation.

However, as shown in Figure 9 and Figure 10, when considering the spatial continuity along the network where the regionalized phenomenon occurs, the use of network distances better represents the spatial behavior of AADT. Range parameters measured from straight-line distances do not correspond precisely to the real region of influence of AADT at a given point, information that can only be measured by semivariograms with network distances. Thus, if there are resources available to obtain network distances, such approach should be prioritized. Otherwise, one should not neglect the anisotropy of AADT. An alternative to spatial interpolation with network distances using a lower-order matrix can be found in Ver Hoef (2018).

5. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE DEVELOPMENTS

Noting that the traffic volume usually shows spatial dependence, the main objective of this study was to estimate the Annual Average Daily Traffic on São Paulo highways based on a spatial interpolator known as Ordinary Kriging. This technique requires only the values of the variable of interest in segments spatially sampled to estimate the traffic volume in segments without this data. The São Paulo highways were used as a case study with two types of distance: the traditional Euclidean distance, and distances along the road network. A third approach analyzed the effect of anisotropy on the spatial modeling of AADT and, consequently, on the calculated estimates.

The results showed that the use of network distances more adequately represents the spatial behavior of AADT than the traditional approach with straight-line distances, and surpassing the goodness-of-fit measures of the anisotropic case. The gains when using network distances as opposed to straight-line distances are about 10% to 20%. Although the scenario with the highest number of points resulted in the best estimates, the case with only 143 counting locations was also satisfactory when compared to previous studies. In this scenario, half of the estimates had an error of less than 38%, while, for the case with 339 points, the median error was approximately 27%.

Although the accuracy of the interpolator decreases in cases of low density of counting stations, in regions with scarce road network the differences between Euclidean and network distances are even greater and the use of distances along the road network is fundamental to improve the results. As the density of count locations increases, obtaining the network distance matrices is computationally expensive, which may become a limiting factor for spatial interpolation. The results obtained in the present article and in previous studies suggest that if there are no resources to calculate the shortest paths, kriging estimates can still benefit from addressing anisotropy. The directional variation of AADT, overlooked in previous studies, was also shown to influence the results.

In locations with low density of count locations, it is necessary to verify if, due to the low number of samples, the interpolator is searching for neighbors outside the range region (a) to calculate the estimates. This phenomenon can impair the performance of the interpolator. In the border regions of the spatial field, which usually suffer from this problem, it may be necessary to include data from adjacent geographic units. Information collected from other categories of roads can also be tested, especially in cases such as Figure 13b, in which the long length of two segments caused the respective midpoint to be at a distance from the nearest neighbor greater than the range. Adjustments in separating segments also appear as a possible solution to this problem.

It is important to remember that Ordinary Kriging is a univariate interpolator. As it does not depend on explanatory variables, its potential to model variables with a high range of variation is limited. Thus, in the case of AADT, it is customary to divide such databases into categories of traffic volume before advancing to geostatistical modeling. Similarly, OK is also not suitable for estimating AADT in the medium and long term, requiring the application of other models, such as Universal Kriging and Regression-Kriging.

In future work, the sensitivity of kriging estimates to the density of samples per kilometer of network can be analyzed in depth. Based on the use of several scenarios, it can be verified to what extent the use of network distances is more advantageous than anisotropic analysis, taking into account the time and resources for processing the network distance matrices.

In summary, the present article contributes to a quick and economical modeling of AADT to support estimating the traffic volume of not only segments of federal highways, but also other categories, since kriging allows calculating the variable in any point in space. An important contribution refers to the fact that since AADT has a remarkable spatial dependence, confirmed by the Moran index and semivariograms, its modeling can be strongly benefited by spatial approaches, especially those that use network distances. One advantage of using geostatistical interpolators is the possibility of incorporating anisotropy into the modeling. As the geostatistical modeling with network distances or anisotropy is available on open source and/or free computer programs, it can be successfully replicated to other databases.

ACKNOWLEDGEMENTS

This research was supported by the São Paulo Research Foundation (Grant #2019/12054-4, Brazil), the Coordination for the Improvement of Higher Education Personnel (CAPES, Brazil) and the National Council of Technological and Scientific Development (Grant #304345/2019-9, Brazil).

REFERENCES

Apronti, D.; K. Ksaibati; K. Gerow et al. (2016) Estimating traffic volume on Wyoming low volume roads using linear and logistic regression methods. *Journal of Traffic and Transportation Engineering*, v. 3, n. 6, p. 493-506. DOI: 10.1016/j.jtte.2016.02.004.

Box, G.E.P. and D.R. Cox (1964) An analysis of transformations. *Journal of the Royal Statistical Society. Series A (General)*, v. 26, n. 2, p. 211-52.

Carvalho, S.D.P.C.; L.C.E. Rodriguez; L.D. Silva et al. (2015) Predição do volume de árvores integrando Lidar e Geoestatística. *Scientia Forestalis*, v. 43, n. 107, p. 627-37.

Chi, G. and Y. Zheng (2013) Estimating transport footprint along highways at local levels: a combination of network analysis and kriging methods. *International Journal of Sustainable Transportation*, v. 7, n. 3, p. 261-73. DOI: 10.1080/15568318.2013.710150.

Chiles, J. and P. Delfiner (2012) *Geostatistics: Modeling Spatial Uncertainty* (2nd ed). New Jersey: John Wiley & Sons. DOI: 10.1002/9781118136188.

Cressie, N. and D.M. Hawkins (1980) Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, v. 12, n. 2, p. 115-25. DOI: 10.1007/BF01035243.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*. New Jersey: John Wiley & Sons.

Deutsch, C.V. and A.G. Journel (1998) *GSLIB: Geostatistical Software Library and User's Guide* (2nd ed). New York: Oxford University Press.

DNIT (2006) *Manual de Estudos de Tráfego*. Rio de Janeiro: Departamento Nacional de Infraestrutura de Transportes. Available at: <https://www.gov.br/dnit/pt-br/assuntos/planejamento-e-pesquisa/ipr/coletanea-de-manuais/vigentes/723_manual_estudos_trafego.pdf> (accessed 06/03/2023).

Duddu, V.R. and S.S. Pulugurtha (2013) Principle of demographic gravitation to estimate annual average daily traffic: comparison of statistical and neural network models. *Journal of Transportation Engineering*, v. 139, n. 6, p. 585-95. DOI: 10.1061/(ASCE)TE.1943-5436.0000537.

Eom, J.K.; M.S. Park; T. Heo et al. (2006) Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. *Transportation Research Record: Journal of the Transportation Research Board*, v. 1968, n. 1, p. 20-9. DOI: 10.1177/0361198106196800103.

Eriksson, M. and P.P. Siska (2000) Understanding anisotropy computations. *Mathematical Geology*, v. 32, n. 6, p. 683-700. DOI: 10.1023/A:1007590322263.

Evans, J.S. (2021) *_spatialEco_. R package version 1.3-6*. Available at: <https://github.com/jeffreyevans/spatialEco> (accessed 06/03/2023).

Gomes, M.M.; C.S. Pitombo; A. Pirdavani et al. (2018) Geostatistical approach to estimate car occupant fatalities in traffic accidents. *Revista Brasileira de Cartografia*, v. 70, n. 4, p. 1231-56. DOI: 10.14393/rbcv70n4-46140.

Goovaerts, P. (2009) Medical geography: a promising field of application for geostatistics. *Mathematical Geosciences*, v. 41, n. 3, p. 243-64. DOI: 10.1007/s11004-008-9211-3. PMid:19412347.

Hollander, Y. and R. Liu (2008) The principles of calibrating traffic microsimulation models. *Transportation*, v. 35, n. 3, p. 347-62. DOI: 10.1007/s11116-007-9156-2.

IBGE (2021) *IBGE Cidades*. Available at: <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama> (accessed 06/03/2023).

Isaaks, E.H. and R.M. Srivastava (1989) *An Introduction to Applied Geostatistics* (1st ed). New York: Oxford University Press. Available at: <https://books.google.com.br/books?id=t62mtgAACAAJ> (accessed 06/03/2023).

Kerry, R.; P. Goovaerts; D. Giménez et al. (2016) Investigating geostatistical methods to model within-field yield variability of cranberries for potential management zones. *Precision Agriculture*, v. 17, n. 3, p. 247-73. DOI: 10.1007/s11119-015-9408-7.

Khan, S.M.; S. Islam; M.D.Z. Khan et al. (2018) Development of statewide annual average daily traffic estimation model from short-term counts: a comparative study for South Carolina. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2672, n. 43, p. 55-64. DOI: 10.1177/0361198118798979.

Kim, S.; D. Park; T. Heo et al. (2016) Estimating vehicle miles traveled (VMT) in urban areas using regression kriging. *Journal of Advanced Transportation*, v. 50, n. 5, p. 769-85. DOI: 10.1002/atr.1374.

Klatko, T.J.; T.U. Saeed; M. Volovski et al. (2017) Addressing the local-road VMT estimation problem using spatial interpolation techniques. *Journal of Transportation Engineering, Part A: Systems*, v. 143, n. 8, p. 4017038. DOI: 10.1061/JTEPBS.0000064.

Krige, D.G. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, v. 52, n. 6, p. 119-39.

Lindner, A. and C.S. Pitombo (2019) Sequential Gaussian simulation as a promising tool in travel demand modeling. *Journal of Geovisualization and Spatial Analysis*, v. 3, n. 2, p. 15. DOI: 10.1007/s41651-019-0038-x.

Marques, S.F. and C.S. Pitombo (2020) Intersecting geostatistics with transport demand modeling: a bibliographic survey. *Revista Brasileira de Cartografia*, v. 72, p. 1028-50. DOI: 10.14393/rbcv72nespecial50anos-56467.

Marques, S. de F. and C.S. Pitombo (2021a) Applying multivariate Geostatistics for transit ridership modeling at the bus stop level. *Boletim de Ciências Geodésicas*, v. 27, n. 2, p. e2021019. DOI: 10.1590/1982-2170-2020-0069.

Marques, S.F. and C.S. Pitombo (2021b) Ridership estimation along bus transit lines based on kriging: comparative analysis between network and euclidean distances. *Journal of Geovisualization and Spatial Analysis*, v. 5, n. 1, p. 7. DOI: 10.1007/s41651-021-00075-w.

Marques, S.F. and C.S. Pitombo (2021c) Spatial modeling of transit ridership along bus lines with overlapping sections. In: *Anais do 35º Congresso de Pesquisa e Ensino em Transportes*. p. 1568-80. Available at: <https://www.researchgate.net/publication/357517939_SPATIAL_MODELING_OF_TRANSIT_RIDERSHIP_ALONG_BUS_LINES_WITH_OVERLAPPING_SECTIONS> (accessed 06/03/2023).

Matheron, G. (1971) *The Theory of Regionalized Variables and its Applications.* Paris: Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu.

Mathew, S. and S.S. Pulugurtha (2021) Comparative assessment of geospatial and statistical methods to estimate local road annual average daily traffic. *Journal of Transportation Engineering, Part A: Systems*, v. 147, n. 7, p. 04021035. DOI: 10.1061/JTEPBS.0000542.

Millard, S.P. (2013) *EnvStats: an R Package for Environmental Statistics.* New York: Springer.

Moran, P.A.P. (1948) The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B. Methodological*, v. 10, n. 2, p. 243-51. DOI: 10.1111/j.2517-6161.1948.tb00012.x.

Olea, R.A. (2006) A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, v. 20, n. 5, p. 307-18. DOI: 10.1007/s00477-005-0026-1.

Oliver, M.A. and R. Webster (2015) *Basic Steps in Geostatistics: The Variogram and Kriging*. Cham: Springer. DOI: 10.1007/978-3-319-15865-5.

Ortúzar, J. de D. and L.G. Willumsen (2011) *Modelling Transport*. Oxford: John Wiley & Sons. DOI: 10.1002/9781119993308.

Paradis, E.; J. Claude and K. Strimmer (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, v. 20, n. 2, p. 289-90. DOI: 10.1093/bioinformatics/btg412. PMid:14734327.

Pebesma, E.J. (2004) Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, v. 30, n. 7, p. 683-91. DOI: 10.1016/j.cageo.2004.03.012.

Pulugurtha, S.S. and P.R. Kusam (2012) Modeling annual average daily traffic with integrated spatial data from multiple network buffer bandwidths. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2291, n. 1, p. 53-60. DOI: 10.3141/2291-07.

Pulugurtha, S.S. and S. Mathew (2021) Modeling AADT on local functionally classified roads using land use, road density, and nearest nonlocal road data. *Journal of Transport Geography*, v. 93, p. 103071. DOI: 10.1016/j.jtrangeo.2021.103071.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.r-project.org/> (accessed 06/03/2023).

Remy, N.; A. Boucher and J. Wu (2009) *Applied Geostatistics with Sgems: A User's Guide*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139150019.

Ribeiro Jr, P.J. and P.J. Diggle (2016). *geoR: Analysis of Geostatistical Data. R package version 1.7-5.2*. Available at: <https://CRAN.R-project.org/package=geoR> (accessed 06/03/2023).

Sarlas, G. and K. W. Axhausen (2015) Prediction of AADT on a nationwide network based on an accessibility-weighted centrality measure. *Arbeitsberichte Verkehrs- und Raumplanung, 1094*, 1-21.

Selby, B. and K.M. Kockelman (2013) Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, v. 29, p. 24-32. DOI: 10.1016/j.jtrangeo.2012.12.009.

Shamo, B.; E. Asa and J. Membah (2015) Linear spatial interpolation and analysis of annual average daily traffic data. *Journal of Computing in Civil Engineering*, v. 29, n. 1, p. 4014022. DOI: 10.1061/(ASCE)CP.1943-5487.0000281.

Sharma, S.; P. Lingras; F. Xu et al. (2001) Application of neural networks to estimate AADT on low-volume roads. *Journal of Transportation Engineering*, v. 127, n. 5, p. 426-32. DOI: 10.1061/(ASCE)0733-947X(2001)127:5(426).

Song, I. and D. Kim (2022) Three common machine learning algorithms neither enhance prediction accuracy nor reduce spatial autocorrelation in residuals: an analysis of twenty-five socioeconomic data sets. *Geographical Analysis*, p. gean.12351. DOI: 10.1111/gean.12351.

Song, Y.; X. Wang; G. Wright et al. (2019) Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles. *IEEE Transactions on Intelligent Transportation Systems*, v. 20, n. 1, p. 232-43. DOI: 10.1109/TITS.2018.2805817.

Souza, V.H.P. and M.R. Silveira (2009) Aspectos econômicos e infraestrutura rodoviária no estado de São Paulo: uma relação solidária. In: *XII Encuentro de Geógrafos da América Latina - Caminando a una América Latina en Trasnformación* (p. 1–16). Montevideo, Uruguay: Universidad de la República.

Stelzenmüller, V.; S. Ehrich and G.P. Zauke (2005) Impact of additional small-scale survey data on the geostatistical analyses of demersal fish species in the North Sea. *Scientia Marina*, v. 69, n. 4, p. 587-602. DOI: 10.3989/scimar.2005.69n4587.

Tobler, W.R. (1970) A computer movie simulating urban growth in the detroit region. *Economic Geography*, v. 46, p. 234-40. DOI: 10.2307/143141.

UFRJ (2018) *Nota Técnica Nº 003/2018/DE: Síntese do Desenvolvimento Técnico-Científico da Metodologia para Estimativa do Volume Médio Diário Anual - VMDa em Toda a Malha Rodoviária Federal Pavimentada*. Rio de Janeiro: Universidade Federal do Rio de Janeiro.

Ver Hoef, J.M. (2018) Kriging models for linear networks and non-Euclidean distances: cautions and solutions. *Methods in Ecology and Evolution*, v. 9, n. 6, p. 1600-13. DOI: 10.1111/2041-210X.12979.

Wang, T.; A. Gan and P. Alluri (2013) Estimating annual average daily traffic for local roads for highway safety analysis. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2398, n. 1, p. 60-6. DOI: 10.3141/2398-07.

Wang, X. and K. Kockelman (2009) Forecasting network data. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2105, n. 1, p. 100-8. DOI: 10.3141/2105-13.

Wong, A.H. and T.J. Kwon (2021) Advances in regression kriging-based methods for estimating statewide winter weather collisions: an empirical investigation. *Future Transportation*, v. 1, n. 3, p. 570-89. DOI: 10.3390/futuretransp1030030.

Yang, H.; J. Yang; L.D. Han et al. (2018) A Kriging based spatiotemporal approach for traffic volume data imputation. *PLoS One*, v. 13, n. 4, e0195957. DOI: 10.1371/journal.pone.0195957. PMid:29664928.

Zhang, D. and X.C. Wang (2014) Transit ridership estimation with network Kriging: a case study of Second Avenue Subway, NYC. *Journal of Transport Geography*, v. 41, p. 107-15. DOI: 10.1016/j.jtrangeo.2014.08.021.

Zou, H.; Y. Yue; Q. Li et al. (2012) An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, v. 26, n. 4, p. 667-89. DOI: 10.1080/13658816.2011.609488.