Intersecting Geostatistics with Transport Demand Modeling: a Bibliographic Survey

Intersectando Geoestatística com Modelagem da Demanda por Transportes: um Levantamento Bibliográfico

Samuel de França Marques [1] and Cira Souza Pitombo [2]

1 University of São Paulo, Department of Transportation Engineering, São Carlos-SP, Brazil. samuelmarques@usp.br.
  ORCID: https://orcid.org/0000-0001-5602-3277

2 University of São Paulo, Department of Transportation Engineering, São Carlos-SP, Brazil. cirapitombo@gmail.com.
  ORCID: https://orcid.org/0000-0001-9864-3175

Abstract: Transport planning depends on modeling variables, and because their collection usually requires high resources, they have limited sampling. However, since they are spatially dependent, the use of Geostatistics in transport demand modeling has proved to be especially convenient, as this tool allows obtaining estimates in non-sampled locations. In this context, the research line related to applying Geostatistics to travel demand forecasting takes place within the scope of three of the four steps of the traditional planning model (trip generation, modal choice and traffic assignment), covering studies that can be divided according to the support, or geographic scale adopted, and type of model used. Thus, in order to establish the state-of-the-art of this research line, the present study proposed surveying and discussing articles within the scope of traffic analysis zones, regular areas, road segments, metro stations, bus stops, bus line segments and household/individual analysis, which used Simple, Ordinary, Indicator, Universal and Spatio-temporal Kriging geostatistical interpolators, in addition to Gaussian Sequential Simulation. The detailed analysis of the studies allowed identifying research gaps in the models' validation stage, comparison with other spatial and non-spatial approaches, use of network distances, applying Universal Kriging (UK) to modal choice variables and the selection of predictors for UK. Special attention should be given to Sequential Gaussian Simulation and Spatio-temporal Kriging, models that could dictate the evolution of the research line in the coming years.
Keywords: Transport demand. Geostatistics. Urban mobility. Kriging. Geographic support.

Resumo: O planejamento de transportes depende da modelagem de variáveis que, em função de usualmente exigirem recursos elevados para a sua coleta, dispõem de uma amostragem limitada. Entretanto, uma vez que apresentam dependência espacial, a utilização da Geoestatística na modelagem da demanda por transportes se mostrou bastante conveniente, já que esse ferramental permite a obtenção de estimativas em locais não amostrados. Nesse contexto, a linha de pesquisa voltada a aplicações da Geoestatística na previsão da demanda por transportes se dá no âmbito de três das quatro etapas do modelo sequencial de planejamento (geração de viagens, escolha modal e alocação de fluxo), abrangendo trabalhos que podem ser divididos conforme o suporte, ou escala geográfica adotada, e tipo de modelo utilizado. Dessa forma, no intuito de estabelecer o estado da arte dessa linha de pesquisa, o presente trabalho propôs o levantamento e discussão de estudos no âmbito de zonas de tráfego, áreas regulares, segmentos viários, estações de metrô, pontos de parada, trechos de linha de ônibus e domicílios/indivíduos, os quais se valeram dos interpoladores geoestatísticos Krigagem Simples, Ordinária, Indicativa, Universal e Espaço-temporal, além de Simulação Sequencial Gaussiana. A análise detalhada dos trabalhos permitiu a identificação de lacunas de pesquisa nas etapas de validação dos modelos, comparação com outras abordagens espaciais e não espaciais, utilização de distâncias em rede, aplicações da Krigagem Universal (KU) a variáveis de escolha modal e seleção de preditores para a KU. Atenção especial deverá ser dada à Simulação Sequencial Gaussiana e à Krigagem Espaço-temporal, modelos que devem ditar a evolução da linha de pesquisa nos próximos anos.
Palavras-chave: Demanda por transportes. Geoestatística. Mobilidade urbana. Krigagem. Suporte geográfico.

1   INTRODUCTION

The problem of data scarcity in modeling, in general, is and has always been a recurring subject in scientific research in the most different study areas. Regarding spatial variables, several methods have been

developed, over the years, in order to circumvent this limitation and to understand the behavior of the variable of interest throughout the space in which it operates.

However, the first methods for data imputation consist of deterministic functions, a condition that contradicts the real nature of spatial variables, which are spatially dependent, making it impossible to perform traditional statistical inference in the estimates provided by them. In this context, in the mid-50s and 60s, the Theory of Regionalized Variables (MATHERON, 1963; 1971) was created, on which Geostatistics is based. Geostatistics is a tool that includes a set of spatial interpolation techniques whose algorithms are based on a probabilistic approach of those spatial variables that have an almost well-defined pattern of variation in space.

Among the advantages of Geostatistics, what stands out is the fact that it is a tool that uses the maximum available information about the variable of interest, normally difficult to collect/acquire, that is, scarce data, in order to estimate its values in unsampled locations, based on the generation of a continuous surface of estimated points.

The emergence of Geostatistics dates back to Krige's pioneering work (1951), focused on the resolution of problems in Mining Engineering, which was initially conceived to model continuous spatial variables, that is, that can assume a value anywhere in the space where they operate, as for example gold content in a land mine, temperature, precipitation, CO_2 concentration in the atmosphere, among others. However, based on the observation that some spatially discrete variables, with values assigned only to certain points in space, also had an almost well-defined spatial structure, Geostatistics was expanded to several study areas.

One of the areas highlighted in this context refers to Epidemiology, where Geostatistics has an important role in estimating disease risks, detection of areas that pose substantially greater risks, and analysis of the influence of risk factors (GOOVAERTS, 2009a). Thus, Goovaerts (2009a) applied Geostatistics to the risk of cervical cancer mortality and to assess the incidence of breast cancer (GOOVAERTS, 2009b). On the other hand, Stelzenmüller, Ehrich and Zauke (2005) successfully applied Ordinary Kriging and Universal Kriging to the biomass of two fish species from the German Bight. And, in agriculture, Kerry et al. (2016) highlighted the better performance of Kriging in the classification of cranberry cultivars compared to another method based on satellite images, and reinforced the cost reduction in fruit counting. In forest sciences, Carvalho et al. (2015) also applied two geostatistical techniques to estimate the circumference of trees. In all these studies, the variables of interest are not spatially continuous, since they refer to individuals in the first two examples, and to fish, fruits and trees, in the antepenultimate, penultimate and last cases, respectively.
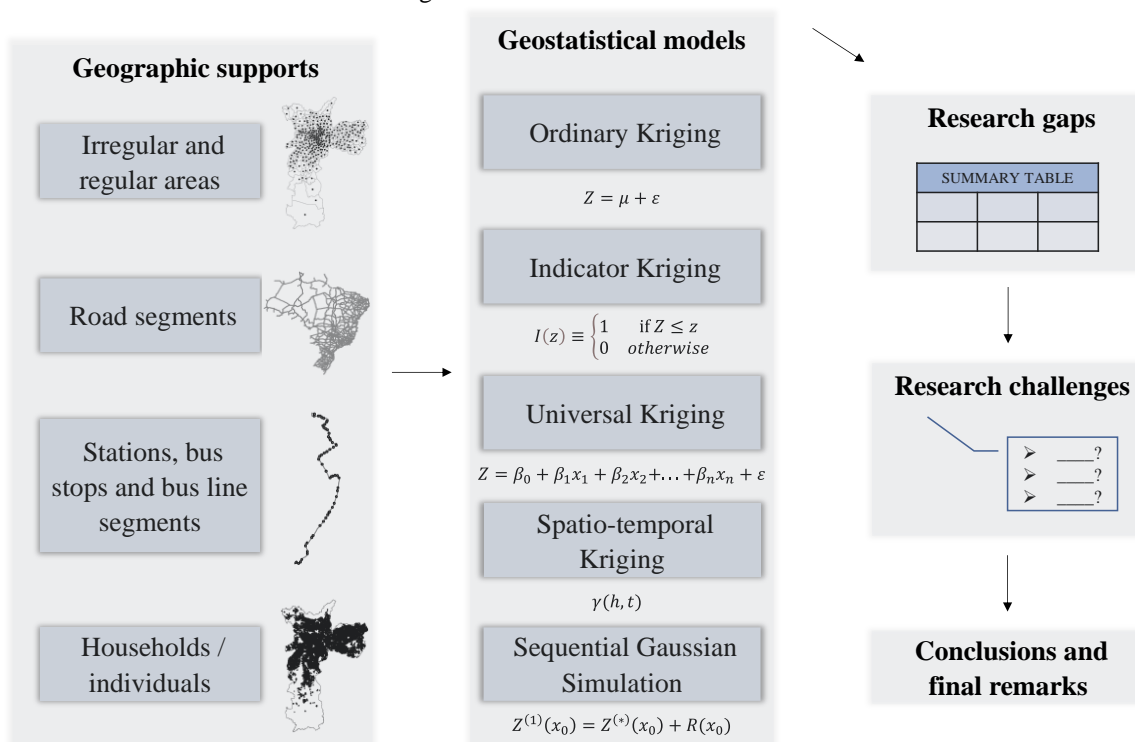
The convenience of applying Geostatistics to spatially discrete variables was also well employed by Transportation Engineering: the use of this tool has been observed in accident/road safety modeling (GOMES et al., 2018; GUNDOGDU; FACULTY, 2014; MAJUMDAR; NOLAND; OCHIENG, 2004; MANEPALLI; BHAM, 2011) and, more recently, in the estimation of transportation demand variables (KLATKO et al., 2017; LINDNER; PITOMBO, 2017; SELBY; KOCKELMAN, 2013; ZHANG; WANG, 2014). In this context, the studies focus on the estimation of trips with origin and/or destination in traffic analysis zones, choice of travel mode (assigned to the geographic coordinate of the household), passengers loading in sections of the public transport network, number of boarding passengers at bus stops, etc. Within this scope, the applications of Geostatistics to problems related to mobility are differentiated by the aggregation unit used in the studies: traffic analysis zones, regular grid squares, homogeneous road sections, subway stations, bus stops, bus line segments and households.

While it is considered that all approaches represent, each in its own way, a contribution to strengthening this line of research, an evolution, not always chronological, can be observed, regarding the supports and models used. Although still in consolidation, since the first studies carried out, there has been a substantial increase in studies related to this theme. The current stage of full development of this line of research shows it is as an opportune moment to survey and schematize publications, as well as pointing out the existing gaps in the Area, in order to guide the state-of-the-art progress and strengthen the line of research. Thus, the main objective of the present study is to address the problem of choosing the best scale for transport demand variables, as well as the utilization of simulation, univariate and multivariate geostatistical modeling.

This article is divided as follows: in addition to this brief introduction, Section 2 describes the classification of Geostatistical approaches to transport demand according to the basic aggregation element used, also known as support. Section 3 emphasizes the evolution of the type of model used in the studies. And,

Section 4 summarizes the comments made in the previous sections, thus deducing the gaps that still exist in this line of research and which may serve as the basis for future research. Based on the comments in Section 4, Section 5 lists some challenging questions that must be answered by future research. Section 6 concludes the study, summarizing the main conclusions reached. Figure 1 illustrates the structure of the article.

Figure 1 – Structure of the article.

Geographic supports

Irregular and regular areas

Road segments

Stations, bus stops and bus line segments

Households / individuals

Geostatistical models

Ordinary Kriging

$Z = \mu + \varepsilon$

Indicator Kriging

$I(z) \equiv \begin{cases} 1 & \text{if } Z \leq z \\ 0 & otherwise \end{cases}$

Universal Kriging

$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

Spatio-temporal Kriging

$\gamma(h,t)$

Sequential Gaussian Simulation

$Z^{(1)}(x_0) = Z^{(*)}(x_0) + R(x_0)$

Research gaps

SUMMARY TABLE

Research challenges

➢ ____?
➢ ____?
➢ ____?

Conclusions and final remarks

Source: The authors (2020).

2   GEOSTATISTICAL SUPPORT IN TRANSPORT DEMAND APPROACHES

The basic unit of aggregation used in Geostatistical approaches is known as support. Bearing in mind that the main objective of Geostatistics is, based on available sample data about a given variable, to interpolate the values of that variable for the entire field of interest, geostatistical support can be defined as the spatial element from which the sample information is collected and where the value of the variable of interest will be estimated (JOURNEL; HUIJBREGTS, 1978).

The choice of support is usually based on the geographic scale for which there is available sample data, depending on the variable of interest, the type of study, methods already consolidated in the literature, among other factors. In the context of mobility variables, the following supports have been observed: areas (irregular and regular ones), road sections, metro stations, bus stops and bus line segments, and households and individuals, starting consecutively from the most aggregated support to the most disaggregated. Below, the studies found using these units are detailed.

2.1   Areas

The application of Geostatistics to transport demand variables at area level is relatively recent and restricted to Brazil, considering the studies found. In this case, the main contributions refer to the adequate treatment of transport demand variables through the inclusion of spatial autocorrelation to modeling, estimation in non-sampled locations and visualization of the spatial behavior of the variable of interest throughout the space in which it operates.

The support of areas, which can refer to both irregular and regular areas, contains studies focused on urban transport demand. In this context, Lindner et al. (2016) used Geostatistics to estimate the number of public transport trips generated, that is, originating in 320 traffic analysis zones in the city of São Paulo, Brazil.

The data included in the modeling were collected from the 2007 Household Origin and Destination Survey, which, according to the authors, covered a total of 30,000 households.

Traffic analysis zones are the basic units of aggregation of the sequential transport planning model (ORTÚZAR; WILLUMSEN, 2011) and their division derives from several aspects, such as: homogeneity in socioeconomic characteristics and the transport system, physical barriers, relief, among others. Considering that trip production consists of the trip generation stage of the sequential model, the use of zones as support in the study by Lindner and Pitombo (2017) proved to be useful.

Notwithstanding this, Geostatistics assumes that the support used must be regular, both in size and in shape (GOOVAERTS, 2008). Since the design of traffic analysis zones is not based on this criterion, the resulting areas exhibit varying sizes and shapes. In addition, Geostatistics also assumes that the aggregation elements must be considerably small compared to the mesh to be interpolated (JOURNEL; HUIJBREGTS, 1978), a condition easily violated in the case of traffic analysis zones.

In order to circumvent these problems, the studies that address the transport demand within areas proposed methods to move information from traffic analysis zones to regular grid squares (LINDNER; PITOMBO, 2019; ROCHA; PITOMBO; SALGUEIRO, 2016; ROCHA, LINDNER; PITOMBO, 2017). In the three cases, it was assumed that the variable of interest, public transport trip generation, was related to a socioeconomic characteristic of the traffic analysis zone, with the first study using the São Paulo metropolitan region as a case study, and the last two, the metropolitan region of Salvador. Due to the impossibility of performing the modeling at the household level, the authors used different procedures to disaggregate the information of the areas in regular squares and perform the geostatistical treatment based on the most refined mesh.

Although the solution proposed by previous studies represents an improvement of the geostatistical modeling applied to transport demand variables, in the scope of areas, it is emphasized that Geostatistics uses only available information in the form of points to generate the surface of estimated values. Regarding the areas, it is considered that the data related to the traffic analysis zones or regular squares are concentrated in the centroid of these units. Thus, as the zones are broken down into smaller and regular area units, the Geostatistical information used to perform the modeling will be increasingly closer to the point representation.

However, since the data also comes from areas, the simplification adopted when using the centroid as a reference point brings the problem of loss of intrazonal variation, since the centroid carries the average value of the variable for the whole zone, that is, it assumes that all points within the zone behave similarly. This unfavorable condition limits the analysis of results and can lead to ecological fallacy. In addition, although variograms from areas associated to centroids have a good structure and less nugget effect, when compared to variograms from more disaggregated units, there is greater heterogeneity and variance associated with the values of the variables. In this sense, other aggregation units, slightly less aggregated, were found in the literature. The transport demand, at the road section level, is commented below.

2.2   Road sections

The application of Geostatistics to road sections generally addresses the variable of Annual Average Daily Traffic (AADT) on rural highways, and comprises most of the studies found in the literature within the scope of this article. The AADT variable is fundamentally important for dimensioning pavements, preparation, monitoring and updating of projects, level of service analysis, implantation of road tolls, etc. However, it is only available for the sections that have traffic counting stations and adjacent segments that are not affected by road access. In this context, based on the AADT collected at the counting stations, Geostatistics has the primary function of estimating this variable in the sections that do not have it, which is even more relevant in the case of rural highways, given the impossibility of installing and maintaining traffic counting stations along the entire length of the road network.

In this context, Klatko et al. (2017) applied four different spatial interpolation methods, including one from Geostatistics, in order to estimate the AADT and, consequently, the total number of miles traveled by vehicles (VMT, Vehicle Miles Traveled) for local roads without traffic count data. Furthermore, Shamo, Asa and Membah (2015) proposed the application of Geostatistics to AADT data for the years 2008, 2009 and 2010

in the state of Washington.

Selby and Kockelman (2013) used Geostatistics to predict/estimate the traffic volume for roads in the state of Texas, in the United States. Additionally, Eom et al. (2006), who claim to be the first authors to publish a study within this scope, used Geostatistics to estimate/predict the AADT for 5 types of roads in Wake County, USA. Wang and Kockelman (2009) also used geostatistical tools, but in this case, the emphasis was given to the Average Daily Volume of two different classes of highways in the state of Texas, USA.

Indirect geostatistical approaches of transport demand, in which the geostatistical interpolation method is applied to a variable directly proportional to the traffic volume, have also been observed in the literature. In this context, in order to assess the impact of traffic volume and its composition on the individual speed of vehicles and on the demand for passenger cars (PCU, Passenger Car Unit), Biswas et al. (2017) used Geostatistics to model/estimate the speed of vehicles by category as a function of traffic volume in the same category and in the other categories. Subsequently, the estimated speed, together with the projected rectangular area of each type of vehicle, the PCU calculation was applied. Another similar application can be found in Zou et al. (2012).

In addition, Chi and Zheng (2013) interpolated the variable transport footprint (carbon footprint due to transportation) associated with each section of the road system in Houghton County, Michigan, USA. The authors put the variable of interest as a function of the AADT and performed geostatistical modeling based on data from counting stations in both Houghton and three adjacent municipalities. Other approaches, based on Geostatistics, for traffic data imputation can be found in Offor, Vaci and Mihaylova (2019), Wang and Mao (2019) and Yang et al. (2018).

Returning to the discussion initiated in subsection 2.1, it is important to comment that, in the case of the AADT of road sections, in contrast to public transport trip production per traffic analysis zone, the aggregation unit is substantially smaller than the mesh to be interpolated and presents relative regularity in its size and shape. In addition, considering that the traffic volume is unique over a given segment, the road sections do not suffer from the loss of intrazonal variation observed in the zones and regular squares. However, in this case, the simplification of assigning the value of the variable of interest to a point along the section remains, since it is a necessary condition for carrying out the geostatistical treatment.

Despite the advantages related to road sections, in comparison with the area units, some limitations, observed and/or commented in the aforementioned studies, can be mentioned. The accuracy of the estimates provided by Geostatistics strongly depends on the amount of information available and the spatial arrangement of the sample points. In the case of road segments, the traffic counting stations, where the information used in geostatistical modeling is collected, is obviously irregularly distributed within the space to be interpolated. Thus, in order to improve the accuracy of the estimates, it is often necessary to use data from geographic units neighboring those of interest. Furthermore, the resulting errors vary substantially, with the smallest errors being attributed to sections located in regions with high density of traffic counting stations, and the largest to segments located in areas with a shortage of stations. Bearing in mind that the process of installing such equipment does not necessarily follow any regular spatial arrangement criteria that would benefit systematic sampling, the geostatistical approach at the road section level is impacted from such restrictions, which are overcome by the supports detailed below.

2.3 Stations, bus stops and bus line sections

The geostatistical approach related to the travel demand at subway stations, bus stops and bus line sections is quite recent and, for this reason, few studies were found within this scope. In this case, the variables of interest (boarding, alighting and loading), in general, also comprise the trip generation stage of the sequential model, since they are the trips produced and attracted by public transport at the station or bus stop level. The loading variable, although similar to the AADT, was not addressed in the previous subsection because the bus line segments are smaller than the road sections, and because the loading is derived from the boarding and alighting volume, information derived from the bus stop aggregation unit.

Regarding the stations, only the study of Zhang and Wang (2014) has been found so far. In this study, the authors proposed applying Geostatistics to the boarding volume at subway stations in the city of New York,

USA. Considering that the number of passengers boarding at metro stations can be easily collected from the electronic ticketing system, the objective of this work was not to interpolate the demand variable, but to extrapolate the boarding volume to metro stations in a new line, which would be implemented in the vicinity of the first metro line.

Concerning the bus stops and bus line segments, the variables of interest are the boarding and alighting volume per bus stop, and the volume of bus passengers in the line segment between two consecutive bus stops, also known as loading. In this context, Marques and Pitombo (2019) and Marques (2019) applied Geostatistics to these three demand variables along a bus line in the city of São Paulo. Such an approach is justified by the high financial resources required to carry out the survey for boarding and alighting counts, a survey that supports the collection of variables of interest and, consequently, the bus public transport network planning. Thus, the authors' intention was to show that Geostatistics could be used to make carrying out the survey along the entire bus network unnecessary, since, with only a percentage of sampled points, the geostatistical tool would be able to interpolate the travel demand for the unsampled bus stop and line segments.

The geostatistical treatment of the public transport demand at the stations, bus stop and bus line segment levels have the following advantages: 1) the variables of interest refer to the points themselves, in contrast to the previous approaches (only the loading variable, given by bus line segment, is similar to the case of road segments); 2) the variables of interest are available at an aggregation level favorable to the application of Geostatistics. In the approach by areas, there is a high degree of aggregation, which is detrimental to disclosing variations that occur within the support used and can lead to ecological fallacy. In the case of households/individuals, as will be seen below, which are also points, there is the problem of high randomness in the immediate neighborhood.

2.4   Households / individuals

The most detailed geostatistical treatment can be attributed to the household or individual supports. While the previous approaches focused on the trip generating stages and network flow assignment, the use of Geostatistics at the individual or household level is related to the modal choice, with only one exception among the studies found.

In terms of choosing the travel mode, Geostatistics is extremely significant. The data collected to carry out modeling come from household surveys or questionnaires. However, it is very difficult to cover the entire population involved in the study. The visits required for applying the questionnaires are considerably expensive, thus such data is considered scarce. On the other hand, knowing the travel mode for the entire population and not just the sampled points is fundamentally relevant for the development of urban projects, implementation of new travel modes, changing the characteristics of existing modes, information on the factors that affect the modal choice, etc.

In this context, Pitombo et al. (2015), based on a household survey, applied Geostatistics to the probability of choosing a car as a travel mode, choosing a non-motorized mode and choosing public transport in the city of São Carlos, SP, Brazil. As these three variables did not show any apparent spatial structure considering all households interviewed, the authors proposed dividing the city into six homogeneous regions according to income and used only two of them in the geostatistical modeling.

Using 110 sampled points from the city of São Carlos, Gomes et al. (2016a, 2016b) interpolated, based on Geostatistics, the probability of choosing the individual motorized mode (car or motorcycle) per household. In this case, the disaggregation of the database into six homogeneous regions, according to income, was initially adopted. However, the authors identified the existence of an almost well-defined spatial pattern in the variable of interest in only one of the regions, which was the only one used in the case study.

Also related to Brazil, the study of Lindner and Pitombo (2017) can be found. Using the 2007 household origin and destination survey of the São Paulo Metropolitan Area, the authors generated maps of the probability of choosing the public transport mode instead of the motorized individual. Despite having errors that were larger than the logistic regression traditionally used in the sequential model (ORTÚZAR; WILLUMSEN, 2011), the authors emphasized that the continuous surface of interpolated values, provided by Geostatistics, is an important contribution, that is, Geostatistics allows knowing the travel mode choice in non-

sampled households, and considering the presence of spatial dependence in the travel mode choice.

While the modal choice was treated in previous studies, in general, without worrying about the destination of the trips, Chica-Olmo, Rodríguez-López and Chillón (2018) proposed applying Geostatistics to the selection of the travel mode to or from four high schools in the city of Granada, Spain. In this article, the objective of the authors was to find the type of distance that best explained the walking mode used for trips to schools and, recognizing that not only did distance influence this modal choice, but also other difficult to measure variables, they modeled the probability of choosing the walking mode instead of the motorized one, using Geostatistics adapted to binary variables of interest. Thus, in addition to knowing the probability of choosing the walking mode in households not interviewed, the continuous surface of estimated values allowed delimiting the schools' area of influence, that is, the regions in the space whose probability of using walking for the home and school trajectory and/or vice versa is closer to 1 (greater than 50%).

Despite the indisputable contributions of geostatistical approaches at the household level, some limitations can be observed in the studies: when using information that has a high level of disaggregation, the occurrence of an apparent spatial pattern in the variable of interest is difficult. Even if there is some spatial dependence in the choice of travel mode, the household level will eventually show high randomness in the immediate neighborhood involved, as human behavior is admittedly complex. Bearing in mind that Geostatistics assumes spatial dependence on the variables of interest, that is, that points close to each other in space exhibit similar behavior, the results of the modeling per household will eventually show substantial errors. In addition, the space to be interpolated may include regions with results of the continuous surface that do not make sense, such as green areas. Operationally, this type of analysis generates well-structured experimental and theoretical variograms, but with a high nugget effect.

A solution to these restrictions was proposed by Rocha (2019). Based on the 2007 origin and destination survey of the São Paulo Metropolitan Area, the author presented a method of aggregating household information (percentage of trips by car) in regular squares. In this study, which included only the city center as a case study, again due to the absence of an apparent spatial structure when the entire city was used, Rocha (2019) presented an optimization model that found the most appropriate regular grid (or support) for the geostatistical treatment undertaken. In other words, the optimal grid would not be as aggregated as the traffic analysis zones, nor as disaggregated as the household level, and the map resulting from the interpolation should reproduce a pattern similar to that of the more detailed level, but with minor errors. The aggregation of the demand for activities in regular cells, based on a household survey, is also found in Yoon, Ravulaparthy and Goulias (2014), with a case study focused on a city in Pennsylvania, USA.

In summary, Section 2 focused on discussing geostatistical approaches of transport demand at different geographic scales. Such geostatistical supports are illustrated in Figure 2.

Figure 2 – From left to right: traffic analysis zones of São Paulo with respective centroids; sampled households in the 2017 origin and destination survey in São Paulo; road sections of the National Road System; and line 856R-10-1 in the city of São Paulo, with respective bus stops and line segments.
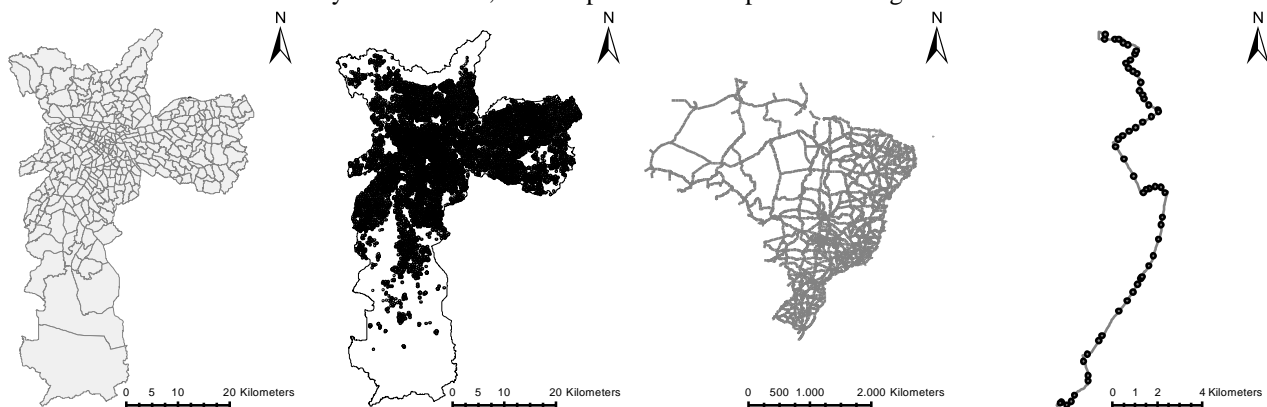
Source: The authors (2020).

In view of the limitations imposed by the assumptions of Geostatistics and by the level of aggregation of the variables, better adequacy of the geostatistical treatment was observed at the bus stop and bus line

segment levels, as well as stations and road segment levels. Such supports are at an intermediate level of aggregation between more aggregated traffic analysis zones, and more disaggregated households, and, in the case of both zones and households, the use of regular grids is shown as a possible solution. The following section classifies the studies according to the geostatistical interpolation used, commenting, as in this section, on its advantages and limitations.

3   GEOSTATISTICAL MODELS USED IN THE APPROACHES

The traditional Geostatistical interpolation methods are known as Kriging, named after Krige (1951), a mining engineer who pioneered this tool. As shown in Eq. (1), such estimators basically comprise a linear combination of neighboring values of the variable of interest, at sampled points, associated with optimal weights, which depend on the spatial arrangement of the database and the theoretical variograms obtained in a previous stage.

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

where $Z^*(x_0)$ is the estimated value of the regionalized variable Z in geographical position x_0; and λ_i is the optimal weight assigned by kriging to the neighbor i observation. Thus, Kriging's main objective is to calculate these weights, taking into account that sampling points closer to the one where the estimate will be carried out should receive a greater weight than the points located further away, following the logic that spatial dependence is more intense in the immediate neighborhood. The evolution of geostatistical modeling of transport demand, using different types of Kriging, is commented below.

3.1   Ordinary Kriging

Ordinary Kriging (OK) is one of the simplest geostatistical interpolation methods and is used in almost all approaches described in Section 2. Its simplicity is due to the fact it is a univariate technique, that is, it depends only on the values of the variable of interest in spatial location points and respective geographic coordinates. OK also assumes that the variable of interest has a constant local mean (μ) and that the values of this variable oscillate around this mean according to the distance between the points in the database (CRESSIE, 1993; MATHERON, 1971). Ordinary kriging representation is shown in Eq. (2).

$$Z = \mu + \varepsilon \tag{2}$$

where ε are the residuals of the model. In order to estimate the Annual Average Daily Traffic (AADT) volume and, consequently, the total miles traveled by vehicles to local roads without traffic count data, Klatko et al. (2017) applied OK to the AADT by dividing the roads into categories according to their characteristics (urban or rural and with heavy or low traffic). In this case, 90% of the database was used in the modeling stage, while the remaining 10% was reserved for validation. The OK results were compared to three other interpolation methods and, although none of the techniques showed predominantly superior performance in the validation, the authors claimed that the spatial interpolation, such as OK, can lead to more reliable and accurate estimates than other existing methods.

Marques and Pitombo (2019) applied OK to the travel demand variables from public transport at the bus stop level (boarding and alighting) and line section level (loading). In order to show the full potential of the OK in estimating such variables, the authors preferred to calculate the goodness-of-fit measures based on cross-validation instead of the validation itself. In Geostatistics, cross-validation is performed through the fictitious point test (CRESSIE, 1993), which uses the entire database in the modeling stage. The OK presented satisfactory performance, in this case, highlighting the loading variable, which, because it demonstrated a smooth spatial variation along the bus line, it was the one that stood out the most when compared to the boarding and alighting variables. The authors stated that using explanatory variables could improve boarding

and alighting performance.

Regarding variables along the transport network, Geostatistics allows expanding the traditional approach, based on the Euclidean distance (in a straight line) between the points in the database, for the modeling along the network. Since the transport activity occurs along a road network, authors such as Wang and Kockelman (2009) stated that using network distances to replace the traditional Euclidean ones could improve the interpolation results. Based on this suggestion, Marques (2019) applied OK, with cross-validation, to the same variables used in Marques and Pitombo (2019), but this time, using also the distances along the bus line. However, comparing the results of this approach with those of traditional OK, the author noted little or no significant improvement.

Chi and Zheng (2013) also used OK with cross-validation and network distances to spatially interpolate the transport footprint variable associated with each section of the road system in Houghton County, Michigan, USA, placing the variable of interest as a function of the AADT. In addition to verifying that the percentage error varied according to the magnitude of the real values, that is, that the errors increased as the value of the variable of interest increased, the authors pointed out, as a limitation, the fact that Ordinary Kriging assumes that the response variable has a constant mean, which may have affected the performance of the technique. In this case, the traditional approach with Euclidean distances was not carried out.

OK was also used by Pitombo et al. (2015) for the interpolation of choosing three different travel modes at the household level for two regions of São Carlos, SP, Brazil. The validation step, for which 30% of the database was reserved, indicated a hit rate below 50% in only one of the six cases analyzed. However, considering that Ordinary Kriging was developed to address continuous variables, the fact that the mode choice is usually of a discrete/categorical nature requires that the original data be transformed before being used in the OK. In this case, Pitombo et al. (2015) used Decision Trees (DT) to convert the dichotomous variable into continuous. As the results of the DTs bring some errors, the estimates of Ordinary Kriging show accumulated errors, adding the uncertainty of the DTs to the OK itself.

In addition, as it is univariate, it is not possible to use the OK for future predictions of the variable of interest (precisely due to the absence of explanatory variables). Thus, the obtained surface can be used for operational planning purposes for short-term periods or to decrease sampled locations. The limitations of the OK are overcome by interpolators detailed in the following subsections.

3.2   Indicator Kriging

Unlike Ordinary Kriging, Indicator Kriging (IK) was developed to deal with categorical and, more specifically, binary variables. As shown in Eq. (3), the dichotomous variable I varies according to a cut-off z, which divides the regionalized variable Z into two groups that can assume only the 0 or 1 values (CRESSIE, 1993; JOURNEL, 1983).

$$I(z) \equiv \begin{cases} 1 & if\ Z \leq z \\ 0 & otherwise \end{cases}$$

(3)

Thus, the geostatistical modeling of the modal choice does not depend on previous steps and the continuous surface of estimated values has only two indications: one for the points where the probability of choosing a particular mode is greater than 50% (or another specified level); and another for the points where the probability is lower than this established cut-off, which would be the region in which other travel mode(s) is(are) prioritized.

In this context, Lindner and Pitombo (2017) compared the performance of IK to that of OK to estimate the preference for public transport instead of the individual motorized mode, in the city of São Paulo for 2007, reserving 30% of the database for validation. Similarly to that adopted by Pitombo et al. (2015), the data used in the OK were derived from the traditional discrete choice model (logit, logistic regression), which uses information from socioeconomic variables to explain the modal choice. Despite showing error propagation, OK results were similar to IK. However, in this case, IK can only generate maps for the analyzed situation, while OK, associated with the logit model, is able to make estimates of the choice probability in alternative/future scenarios. However, additional information is required, which is not the case with IK.

In the case of Gomes et al. (2016a, 2016b), instead of using the logit model, the discrete choice variable was transformed into continuous based on Decision Trees. Thus, the OK was applied to the probabilities of choosing the individual motorized mode, subsequently compared to the IK, which uses the original data in its raw form. In contrast to the results of Lindner and Pitombo (2017), the simplicity of IK performed better in cross-validation than the two-stage method involving DTs and Ordinary Kriging, which suffered losses due to the propagation of errors.

The association of OK to the logit model, as verified in Lindner and Pitombo (2017), basically seeks to overcome the limitation of OK and IK to model only the situation under scrutiny, which prevents its application to medium and long-term planning. In addition, as these univariate interpolation techniques do not depend on explanatory variables, they also prevent knowing the relationships between the variables of interest and the factors that affect them. Such restrictions are circumvented by the Kriging Universal interpolation method, as commented below.

3.3 Universal Kriging

Resuming the introduction of Section 3.1, Ordinary Kriging assumes that the variable of interest has a constant local mean and that the estimates at unsampled points depend on this central value and on the distances between such points and their neighbors (CRESSIE, 1993; MATHERON, 1971). This assumption can generate more accurate results only when the amplitude of local variation of the variable of interest is not high. As sudden changes begin to occur in the variable of interest, especially in the close neighborhood, the OK estimator becomes inefficient, as is the case of Marques and Pitombo (2019) and Chi and Zheng (2013), previously commented.

However, the Universal Kriging (UK) estimator relaxes this assumption by assuming there is a long-range variation in the variable of interest, which occurs due to the influence of external variables. Thus, as shown in Eq. (4), the constant local mean is replaced, in most cases, by a linear combination of explanatory variables x_k and parameters β to be estimated (CRESSIE, 1993).

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{4}$$

where ε are the residuals of the model, in which a spatial structure is assumed to exist. Due to this flexibility, but with the additional cost of the explanatory variables, UK usually achieves better results than OK and also allows predicting the values of interest in other scenarios, in addition to the one analyzed.

However, to avoid the need to collect secondary data for geostatistical modeling, it is possible to use, as explanatory variables, only the geographical coordinates of the points, as performed by Wang and Kockelman (2009), referring to applications of UK to traffic volume. Wang and Kockelman (2009) found that the method used returned reasonably reliable results for intermediate traffic volumes, whereas, for low AADT and high AADT, the technique overestimated and underestimated, respectively, the real values. According to the authors, these results could be improved with data acquisition about explanatory variables and using network distances.

One of the simplest forms of Universal Kriging, which actually uses an external explanatory variable, refers to Kriging with External Drift (KED). KED requires a drift model, deduced from a linear function of the secondary variable (CRESSIE, 1993; HENGL; HEUVELINK; ROSSITER, 2007). Thus, a secondary variable, which must be highly correlated with the variable of interest, is selected to assist in the modeling. Thus, KED was used by Rocha, Pitombo and Salgueiro (2016) and Rocha, Lindner and Pitombo (2017) to estimate the public transport trip production, with population as a secondary variable. In the validation, which used 40% of the database, errors close to zero and high adherence between real and estimated values were observed.

Lindner et al. (2016) also applied KED to the public transport trip production, but this time, the authors used, as a secondary variable, one of the factors resulting from the aggregation of multiple socioeconomic variables from a Principal Component Analysis. Comparing the results of this approach with those of simple linear regression (SLR), the authors verified similar performance of both techniques, but reinforced that only KED, based on the spatial autocorrelation of the data, is able to generate maps with estimated values, including

at unsampled points. Linear regression, in turn, does not account for the spatial dependence of the travel demand variable.

Using multiple explanatory variables, Eom et al. (2006) applied UK for the AADT estimate/prediction for five types of roads in Wake County, USA. Based on a sample with only 200 of 1154 counting stations, the variance explained by the explanatory variables alone was of approximately 65%, leaving a contribution of 55% to the spatial interpolation. Compared to multiple linear regression (MLR), the UK results showed superior performance. Biswas et al. (2017) also applied UK for indirect modeling of traffic volume, in passenger car units, based on vehicle speed. In this case, Universal Kriging was validated using data collected from an arterial route with characteristics similar to those of the route whose data were used in the model training. The authors also compared the UK results with those obtained from regressions based on traffic volume and density, with the UK showing better performance in speed prediction for the five vehicle categories considered.

As suggested by Wang and Kockelman (2009), UK with network distances was compared to traditional UK (with straight-line distances) and geographically weighted regression (GWR) in the AADT prediction/estimate for state roads of Texas in the United States (SELBY; KOCKELMAN, 2013). Explanatory variables such as speed, number of lanes, population accessibility, job density and functional class were included in the modeling. Based on error metrics applied in the validation stage, the results showed a slightly better performance of the Universal Kriging over the GWR, and both techniques were superior to a third non-spatial model that considered only the presence of heteroscedasticity in the variable of interest. However, the network distance alternative, in addition to requiring significant computational resources and processing time, did not demonstrate a positive effect on the results, although the authors claim that models based on count data densely distributed in space could be benefited by such a feature. However, they stressed that the use of non-Euclidean distances can compromise achieving a positive covariance matrix, necessary for the application of Kriging.

Network distances were also used by Zhang and Wang (2014) in spatial modeling of the number of boardings at subway stations in New York City. In this case, the traditional UK approach with Euclidean distances, together with the classic linear regression, was compared to the UK with network distances. Although the latter approach provided parameters estimates with statistics t considerably higher than the others, the performance of the three models did not vary significantly. The reason for this may have been the high variation percentage already explained by the long-range trend (the linear combination of parameters), which left little variability for the spatial structure of the variable, expressed by the semivariogram function. In turn, the small performance difference of the UK with and without network distances has again questioned the real effectiveness of using network distances.

Finally, Shamo et al. (2015) proposed the adjustment of three types of Kriging (Simple, Ordinary and Universal) associated with three variogram models (exponential, spherical and Gaussian) to the AADT data for the years 2008, 2009 and 2010 in the state of Washington. Simple Kriging is based on an even more rigorous assumption than that of OK: in addition to being univariate, this interpolation method considers that the variable of interest has a constant and well-known global mean (YAMAMOTO; LANDIM, 2015). In general, based on a validation process in which several error metrics were applied, an optimal combination for the database was not identified. Thus, the authors reinforced that the best Kriging technique and best variogram can only be determined based on the structure present in the available information about the variable of interest.

Although the equation describing the long-range trend in Universal Kriging is usually similar to linear regression, Chica-Olmo, Rodríguez-López and Chillón (2018) modeled the likelihood of choosing the walking mode from Universal Kriging (called regression-kriging by the authors) adapted for binary variables of interest, in which, instead of using the linear combination of explanatory variables, it used logistic regression. Thus, the authors assigned the spatial dependence of the probability of choosing the mode to the residuals, since it covered the locational explanatory variables not included in the model. This application, which was the only one of its kind found, would overcome the limitations of Indicator Kriging, which cannot predict the choice mode in hypothetical/future scenarios, and Ordinary Kriging, which depends on other techniques to interpolate the modal choice, displaying error propagation.

Through the studies in which Kriging is applied, it can be observed that the variables of interest are

given, in some cases, in time series. Geostatistics can also be used in databases that contain a temporal and a spatial aspect, as shown below.

3.4   Spatio-temporal Kriging

Spatio-temporal Kriging refers to the interpolation in which the covariance matrix takes into account not only spatial, but also temporal lags. Although several studies use cross-section data, that is, they cover only one period of time, in some cases the availability of exhaustive travel demand information over time is observed. In this context, it is essential to recognize the dynamic nature of trip variables and the potential dependence between present, past and/or future values (serial / temporal autocorrelation), a situation where Spatio-temporal Kriging is very convenient.

Within the scope of this article, until now, few studies have been found that apply Spatio-temporal Kriging to a travel demand variable. In order to provide a taxonomy of urban environments based on their dynamics and social aspects, Yoon, Ravulaparthy and Goulias (2014) interpolated the activity intensity in regular space cells in a small city in Pennsylvania, USA. Based on an activity diary, the interpolated variable corresponded to the number of people performing a given activity (shopping, food) in space and time, with hourly intervals. Since the sample of respondents corresponded, approximately, to only 1% of the population of the municipality, the use of the Spatio-temporal Kriging, in this case, was essential to understand the spatio-temporal patterns of daily activities in the entire city considered.

Yang et al. (2018) applied Spatio-temporal Kriging to impute traffic volume data based on recordings from six sensors installed consecutively along a road in Nashville, Tennessee, USA. A panel database was consolidated based on the sensor recordings for 33 days, at 30-second intervals, and different analysis scenarios were proposed, gradually increasing the percentage of missing data. The comparison with two other data imputation methods (historical average and k-nearest neighborhood), using one of the sensors as a case study, revealed that the Spatio-temporal Kriging was the method with the best accuracy in ten of the eleven scenarios considered. This approach has the advantage of including, in the modeling, the fact that, not only the traffic volume of the sensor chosen for the analysis depends on the volume recorded in the adjacent sensors, but also there is a dependence between the current, past and/or future traffic volumes, recorded by the studied sensor.

3.5   Sequential Gaussian Simulation

Like the Spatio-temporal Kriging, the Sequential Gaussian Simulation (SGS) is also a geostatistical tool whose application to travel demand variables is quite recent and still incipient. In a simplified way, while Kriging involves the analysis of only one scenario of the variable of interest, whether current, hypothetical or future, SGS performs the randomization of the regionalized phenomenon (REMY; BOUCHER; WU, 2009). In this context, instead of producing only one surface of estimated values, this technique generates several simulations of Kriging results, from which it is possible to map uncertainty measures, obtain the average of realizations, confidence intervals and analyze critical scenarios.

Moreover, the simulations allow to generate a disaggregated synthetic population, from an aggregated database, which reproduces, with the greatest possible similarity, the profile of the true population (PIANUCCI et al., 2019), which represents an important advance in the search for solutions to the limited sampling of household surveys. This feature also allows SGS to handle changes in the support.

In this context, the only application of Sequential Gaussian Simulation to a transport demand variable, found so far, is due to the study by Lindner and Pitombo (2019). The authors simulated public transport trip production in the 460 traffic analysis zones in the São Paulo Metropolitan Area, changing the support to a regular grid. Thus, it was possible to plot maps of the mean values of the realizations, variance, minimum and maximum values in each cell.

After detailing the geostatistical models already used in the treatment of transport demand, as well as their advantages and limitations, the following section summarizes the comments made in this section and in Section 2, thus deducing the existing research gaps within this scope. To the reader interested in the mathematical deepening of the interpolations commented in this section, the following publications are

recommended: for Ordinary Kriging, Indicator Kriging and Universal Kriging (CRESSIE, 1993; JOURNEL, 1983; MATHERON, 1971); for Spatio-temporal Kriging (GRÄLER; PEBESMA; HEUVELINK, 2016); and for Sequential Gaussian Simulation (DEUTSCH; JOURNEL, 1998). Several of the articles cited in sections 2 and 3 also offer an objective demonstration of the statistical formalism imbued in the estimator applied.

4   RESEARCH GAPS

Table 1 consolidates the studies commented in sections 2 and 3, and summarizes their main characteristics.

Table 1 – Summary of the studies analyzed. IK, OK, UK, SLR, MLR and VMT refer, respectively, to Indicator Kriging, Ordinary Kriging, Universal Kriging, Simple Linear Regression, Multiple Linear Regression and Vehicle Miles Traveled.

Source	Support	Variable(s)	Methods used (*software*)	Advantages (+) and Limitations (-)
Lindner et al. (2016)	Traffic analysis zones	Public transport trip production	UK with only 1 covariate and SLR (*GeoMS 1.0*)	(+) Multivariate spatial analysis; future projections. (-) Secondary variable derived from Principal Component Analysis; Modifiable Area Unit problem.
Yoon, Ravulaparthy and Goulias (2014)	Regular squares	Number of people performing a given activity	Spatio-temporal Kriging without validation (R)	(+) Mapping how urban spaces attract different social interactions at different times of the day. (-) Does not consider the interdependence of activities between individuals; it does not include covariates or comparison with other methods.
Rocha, Pitombo and Salgueiro (2016) e Rocha, Lindner and Pitombo (2017)	Regular squares	Public transport trip production	UK with only 1 covariate and validation (*GeoMS 1.0)*	(+) Disaggregation in regular mesh; multivariate spatial analysis. (-) Propagation of errors because the dependent variable is an estimate.
Rocha (2019)	Regular squares	Percentage of car trips	OK (*R Studio, SGeMS and GeoMS 10.1*)	(+) Disaggregation in optimal regular mesh; (-) Time of execution of the algorithm; low spatial dependence; occurrence of negative semi-variances.
Lindner and Pitombo (2019)	Regular squares	Public transport trip production	Sequential Gaussian Simulation (*R Studio and SGeMS 3.0*)	(+) Mapping of critical scenarios; calculation of uncertainty parameters; change of support. (-) Deals only with short-term planning and Gaussian variables.
Eom et al. (2006)	Road segments	Annual average daily volume	UK and MLR with validation	(+) Comparison between different methods of parameter calibration; UK performed best. (-) Variation of errors according to the density of counting stations.
Wang and Kockelman (2009)	Road segments	Annual average daily volume	UK only with coordinates and validation (*ArcGIS*)	(+) Initial attempt to model the long-term trend through the geographical coordinates of the points. (-) Reasonably reliable results only for intermediate traffic volumes; absence of covariates and the use of network distances; potential error propagation as the dependent variable is an estimate.
Chi and Zheng (2013)	Road segments	Transport carbon footprint / Annual average daily volume	OK with network distances (*R 2.12.2*)	(+) Use of robust estimator of the semivariogram and network distances. (-) Variation of the percentage of error according to the magnitude of the real values; reduced number of points for calculating the semivariogram; absence of covariates.
Selby and Kockelman (2013)	Road segments	Annual average daily volume	UK, with network and Euclidean distances, GWR and non-spatial model, with validation (*MATLAB*)	(+) Use of Box-Cox transformation and robust estimator; comparison between spatial and non-spatial models; UK performed best. (-) Kriging's limitation regarding the heteroscedasticity of the variable of interest; variation of errors according to the density of counting stations.
Shamo, Asa and Membah (2015)	Road segments	Annual average daily volume	SK, OK and UK with validation (*ArcGIS*)	(+) Comparison of combinations between interpolators and theoretical semivariogram models. (-) Non-modeling of the long-term trend based on covariates.

(To be continued)

(Conclusion)

Source	Support	Variable(s)	Methods used (*software*)	Advantages (+) and Limitations (-)
Klatko et al. (2017)	Road segments	VMT / Annual average daily volume	OK with validation (*ArcGIS*)	(+) Does not require additional information; division of roads into classes. (-) Requires a comprehensive inventory of local roads.
Biswas et al. (2017)	Road segments	Vehicle speed / Passenger car units	UK and MLR with validation	(+) Transferability of results to other countries, either through the prepared graphs, or through replication of the experiment to new data; Better performance of UK compared to the non-spatial model; division of vehicles by category. (-) Only the volume of traffic for five categories of vehicles is considered as covariate.
Yang et al. (2018)	Road segments	Traffic volume	Spatio-temporal Kriging with validation (*R Studio*)	(+) Validation with gradual increase in the percentage of missing data (sensitivity analysis); more accurate and flexible method than the historical average and k-nearest neighborhood; the method does not assume any probability distribution for the data. (-) Absence of comparison with other geostatistical interpolators and non-spatial models.
Zhang and Wang (2014)	Subway stations	Number of boardings	UK, with Euclidean and network distances, and MLR without validation (*MATLAB*)	(+) Use of covariates and network distances; UK with network distances performed better than the traditional UK and a non-spatial model. (-) Covariates explain a significant part of the boarding variation and leave little to the semivariogram function.
Marques and Pitombo (2019)	Bus stops and bus line segments	Boarding, alighting and loading	OK (*ArcGIS 10.1*)	(+) Favorable level of aggregation; resource savings of boarding and alighting counts survey. (-) Absence of covariates, validation and comparison with other methods.
Marques (2019)	Bus stops and bus line segments	Boarding, alighting and loading	OK with Euclidean and network distances (*R Studio*)	(+) Favorable level of aggregation; use of network distances; resource savings of boarding and alighting counts survey. (-) Absence of covariates, validation and comparison with other methods; distances along the road network very similar to distances in a straight line.
Pitombo et al. (2015)	Households	Probability of choosing the car mode, choosing a non-motorized mode and choosing public transport	OK with Decision Trees and validation (*GeoMS*)	(+) Estimation of modal choice based on geographical position and socioeconomic attributes; variable of interest generated from a non-parametric technique. (-) Propagation of errors; only short-term planning.
Gomes et al. (2016a, 2016b)	Households	Probability of choosing the individual motorized mode	OK with Decision Trees and IK (*GeoMS*)	(+) Estimation of modal choice based on geographical position and socioeconomic attributes; variable of interest generated from a non-parametric technique; comparison with direct method; IK performed better than OK. (-) Propagation of errors; only short-term planning.
Lindner and Pitombo (2017)	Households	Probability of choosing the public transport mode	OK with logistic regression and IK, both with validation (*GeoMS*)	(+) OK medium and long-term planning with logit trend; estimation of modal choice based on geographic position and socioeconomic attributes; comparison with direct method; similar performance of OK and IK. (-) Error propagation in OK.
Chica-Olmo, Rodríguez-López and Chillón (2018)	Individuals	Probability of choosing the walking mode	UK with logit trend without validation	(+) Medium and long-term planning; analysis of the effect of covariates; no error propagation. (-) Absence of validation and comparison with classical non-spatial methods.

Source: The authors (2020).

Although, in Geostatistics, both validation and cross-validation are performed by the fictitious point test, in the present study, validation was considered only in cases where the authors reserved a part of the database specifically for this purpose. When the entire database is used for both model calibration and validation, this validation is called cross-validation. In Table 1, the methods column discriminates the studies according to the inclusion of a validation stage or absence thereof. For the cases that used cross-validation, this information was omitted in the cell.

Table 1 shows that the line of research aimed at Geostatistics applications for transport demand is relatively recent: the first study, among those analyzed, refers to the year 2006. However, in view of the wide variety of studies thus far undertaken, this research line has shown considerable progress. In addition, the articles cited are mostly limited to Brazil and the United States, with one representative from Spain. Regarding Brazil, the case studies focus on the trip generation and modal choice stages, while in the USA the emphasis is on the traffic volume in road sections, that is, the flow assignment step.

Although Geostatistics has assumptions that are not verified in all studies, it is emphasized that the modeling of the aforementioned stages depends strictly on certain supports and techniques. In this context, it is recommended to continue looking for solutions that converge with geostatistical restrictions, but that do not limit the potential applications, since the occurrence of data scarcity is recurrent in transport demand.

Table 1 identifies some research gaps. It should be noted that the Spatio-temporal Kriging was used only twice in the articles covered. It is noteworthy that the collection of travel demand data over time is not uncommon in transport engineering: traffic counting stations usually record the number of vehicles at 5-minute intervals. In addition, the origin and destination survey, on which modal choice and traffic analysis zones studies are based on, also has a report of daily activities practiced by the interviewee the day before the visit. Smartphones to collect trip information, in turn, also allow creating a panel database. Considering that the travel demand information over time has, to a lesser or greater degree, temporal dependency, the application of Geostatistical spatio-temporal interpolation is strongly recommended, hence providing spatial missing data in historical series.

This suggestion can be extended to Sequential Gaussian Simulation. Transport planning usually involves proposing solutions to previously diagnosed mobility problems, based on the analysis of hypothetical scenarios. In this context, projecting the current transport condition into the future considers at least three approaches: realistic, pessimistic and optimistic. Since SGS can generate several simulated scenarios, including numerous combinations of results, its application to other travel demand variables and case studies is extremely convenient. However, it is noteworthy that, both for SGS and Spatio-temporal Kriging, validation and comparison with other traditional methods are essential for strengthening and consolidating the technique.

In this context, there is a lack of validation, or presence of only cross-validation, in several studies contained in Table 1. In the case of univariate interpolation (Indicator Kriging and Ordinary Kriging), a plausible justification would be that such techniques depend only on the variable of interest to carry out the modeling. Thus, when a part of the database is reserved only for validation, the loss of information in the calibration of the models can negatively impact the results, which is inconvenient if the purpose of the investigation is to highlight the maximum potential of the estimator. However, if the study does not include a validation stage, the ability of the geostatistical tool to treat scarce data is not properly proven.

The potential for using validation is increased through Universal Kriging, which uses information about explanatory variables, in addition to the variable of interest. The ideal situation would be to validate the method based on the gradual removal of data and verify the extent to which interpolation can still generate acceptable estimates. In this case, one could attest to the effective cost reduction resulting from not having to carry out a survey in the entire field where the variable of interest operates, as for instance, the boarding and alighting counts survey, from which the variables addressed by Marques and Pitombo (2019) and Marques (2019) originate. Among the studies in Table 1, validation with gradual increase in the percentage of missing data was observed only in the study by Yang et al. (2018).

However, it should be noted that, although the Universal Kriging interpolation technique would allow better validation results than those of OK and IK, the need to collect additional data may represent a trade-off in this technique. Thus, it is appropriate to compare the multivariate estimator with the univariate ones whenever possible, in order to question whether the resulting UK gains outweigh the additional cost of auxiliary information. The comparison of Geostatistics with other methods, both spatial and non-spatial, is also absent in several studies of Table 1, which is fundamentally relevant for advancing the state-of-the-art of the line of research under analysis.

On the other hand, Ordinary Kriging and Indicator Kriging can generate estimates only for the scenario under analysis, while Universal Kriging, in addition to allowing to know how and how much the external variables affect the transport demand, also allows predicting the variable of interest for hypothetical and future

scenarios. In other words, UK is able to contribute to medium and long-term planning, while OK and IK are restricted to the short term, for which it is assumed there is no significant variation in the intervening factors of transport demand.

It is emphasized, however, that although UK has been used successfully in several of the studies shown in Table 1, its modeling can also be improved. Some studies (SELBY; KOCKELMAN, 2013; ZHANG; WANG, 2014) showed a high percentage of variation, in the variable of interest, explained by the explanatory variables. In UK, the spatial structure of the model is assigned to the residuals, which is used to calculate the variogram function. If only a small part of the variation of the variable of interest is assigned to the spatial structure, little or no improvement can be provided by the UK, when compared to classical linear regression. In this context, it is necessary to develop methods that can predict how many and which are the best predictors to be included in Universal Kriging, as a way to optimize the performance of its estimates and considering the trade-off between predictive power and number of explanatory variables, as previously commented.

It is also highlighted that the traditional OK and UK assume normality of the variables of interest. Considering that transport demand information usually occurs in the form of counts and presents asymmetry, solutions to bring such data closer to the normal distribution include the Box-Cox transformation, used by Eom et al. (2006), Selby and Kockelman (2013), Marques (2019) and Marques and Pitombo (2019). However, it cannot be said that this method is sufficient to improve the estimates, since, when calculating the goodness-of-fit measures, the real values of the variable of interest are compared to the estimated values already submitted to the inverse transformation, whose results may be similar to those that would be obtained if the transformation had not been used.

Another way to smooth the asymmetry of the count data would be through the application of robust variogram estimators, as proposed by Cressie and Hawkins (1980), applicable for cases where the distribution is normal in the central region, but flatter than the normal in the tails. Among the studies cited, only Selby and Kockelman (2013) and Chi and Zheng (2013) used the aforementioned estimator. Thus, there is an opportunity to prove the better performance of the estimates coming from robust variogram estimators in order to popularize their use. The elaboration of methods able to imbue the nature of count data in geostatistical interpolation, just as the IK was created to deal with binary variables, is also opportune.

Regarding demand along the transport network (road system, subway or bus line), the better performance of interpolation using network distances, instead of the Euclidean ones, has not yet reached a consensus conclusion. In the case of Marques (2019), the author attributed the similarity of results of the two distance approaches, among other factors, to the fact that the bus line used in the study has a more or less straight route, with network distances being similar to the distances in a straight line. Thus, applying this method to bus lines with a more curvilinear and/or irregular layout, and subsequently comparing it to the traditional estimator is indicated.

However, it should be noted that, as pointed out by Selby and Kockelman (2013), the use of non-Euclidean distances can compromise the achievement of a positive covariance matrix, necessary to calculate Kriging estimates. In order to circumvent this problem, Ver Hoef (2018) developed computational routines for prior verification, based on the spatial arrangement of the transport network, of the possible non-occurrence of the aforementioned phenomenon. This author also provided a method for resolving this obstacle, if confirmed. However, among the works that used network distances, only Marques (2019), as it is the most recent study, used the inspection proposed by Ver Hoef (2018), a fact that points to the importance of adapting the method to this update in following research to be carried out within this scope.

Attention is also drawn to the use of Universal Kriging, as proposed by Chica-Olmo, Rodríguez-López and Chillón (2018). Considering that the variable of interest in the study was binary (modal choice), instead of the linear equation usually imbued in the UK estimator, the authors applied the logit function. This approach should be further developed regarding its calibration process, as well as compared to other modal choice approaches within the scope of Geostatistics: Indicator Kriging and Ordinary Kriging with Decision Trees or logistic regression. Comparing the performance of this approach with that of the sequential planning model, which uses non-spatial logistic regression, is also necessary to better understand its contribution. In addition, the calculation of error metrics, based on a validation sample, was not included in the abovementioned study and represents an important tool to prove the accuracy and reliability of the model. If its applicability and

effectiveness is confirmed, the use of the logit trend, in Universal Kriging applications for modal choice, may serve as an important tool for predicting modal choice in scenarios other than the one analyzed, which is not possible in Indicator Kriging, and without depending on previous treatments that can lead to the propagation of errors, as is the case of Ordinary Kriging.

Finally, in addition to suggesting the application of Spatio-temporal Kriging to data in a transport demand panel, it is also recommended to include explanatory variables to this interpolator, whose original formulation requires only the value of the variable of interest in space and in time. Just as Universal Kriging usually demonstrates better performance than Ordinary Kriging, a Spatio-temporal interpolation that accounts for the influence of external variables can make significant contributions to the geostatistical modeling of transport demand. Also, with regard to external variables, the search for interpolators that can discriminate the spatial heterogeneity of the covariates estimated parameters is encouraged, which often vary at the local level. Synthesizing and complementing the comments made in this section, Figure 3 shows the evolution of the research line aimed at Geostatistical applications for modeling the transport demand. This figure summarizes the advances in using the most appropriate support, as well as the interpolation evolution, highlighting the cases that have been little or not yet explored.

Figure 3 – Summarized evolution of the intersection of Geostatistics with transport demand modeling.
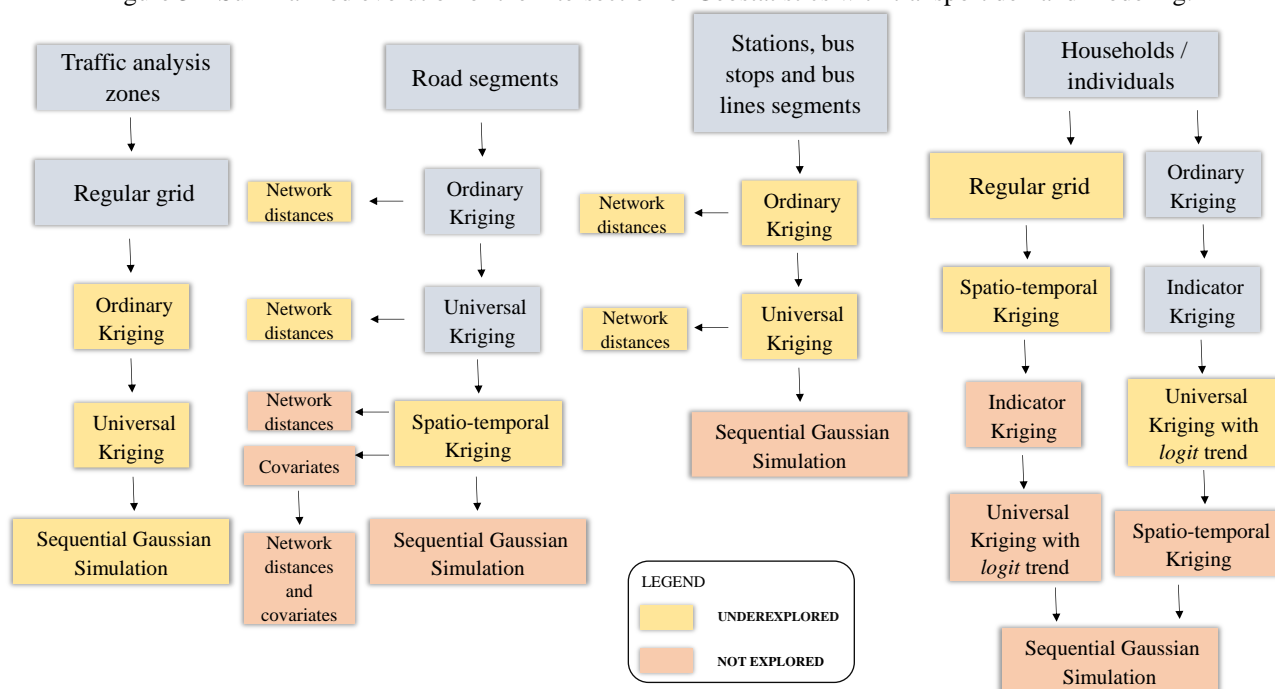
Source: The authors (2020).

5 FUTURE CHALLENGES

Based on the research gaps consolidated in Section 4, some research challenges are synthesized as questions:

a) What improvements can geostatistical modeling bring to urban and transport system planning, both in terms of method as well as in the practical aspect?

b) How can the selection of predictors for multivariate geostatistical modeling be optimized?

c) In the case of travel demand variables that occur along the road network, can the use of network distances bring better results than distances in a straight line? If so, under what circumstances and how significant are the gains?

d) How to apply Geostatistics to the case of bus lines with overlapping sections?

e) Can Universal Kriging, with logit trend equation, really be successfully used to model modal choice variables?

f) Could the inclusion of the non-stationarity of the parameters, in the multivariate geostatistical

1044

modeling, improve travel demand estimates?

g)   What is the adequate minimum percentage of bus stops, to be sampled in a boarding and alighting counts survey, to generate acceptable travel demand estimates at the non-sampled points?

h)   Can the univariate and multivariate Spatio-temporal Kriging be successfully applied to longitudinal transport demand databases? If so, are these models competitive against traditional regressions for panel data?

i)   How can the use of Gaussian Sequential Simulation be useful for the operational planning of the public transport system?

Answers to these questions are expected to be part of the scope in future research intended to the application of Geostatistics in transport demand modeling.

6   CONCLUSIONS AND FINAL REMARKS

Transport planning usually depends on the modeling of transport demand variables, which are difficult to collect, that is, they require high financial resources to be acquired. In view of the spatial dependence normally seen in these variables, geostatistical applications in transport demand began to be found in the second half of the 2000s, in the scientific literature. The approaches utilize the capacity of the geostatistical tool to use as much information about a spatial variable to estimate its values in non-sampled locations.

In this context, the studies, which focus on three of the four stages of the sequential model of transport planning (trip generation, modal choice and traffic assignment), were analyzed based on two main characteristics: the geographic support and the geostatistical model used. With regard to support, the research carried out at the bus stops, bus lines sections, subway station and road segment levels are the ones that stand out the most, in terms of the suitability of the level of aggregation of the variables and convergence with the assumptions of Geostatistics. The use of regular optimized grids, either by genetic algorithms or Gaussian Sequential Simulation, replacing the original information given in traffic analysis zones and households, is one of the improvements in the search for the most appropriate geographic support for geostatistical applications to transport demand.

The type of support is directly associated with the contribution of studies to transport demand modeling. The studies carried out in the context of traffic analysis zones, represented by their centroid, include the spatial dependence of trip observations in the process to estimate the variable of interest, given that this characteristic is not accounted for in the classical regression models. This contribution extends to all other supports, however, in the case of regular grids, the process of disaggregating travel demand data, when derived from traffic zones, or aggregation, when derived from households, complements the already mentioned contribution, considering the progress made in the search for an aggregation unit that meets the assumptions of Geostatistics and also demonstrates satisfactory results. In general, these studies cover the trip generation stage of the sequential transport planning model.

As representatives of the flow assignment stage, the Geostatistical applications to the modeling of Annual Average Daily Traffic represents an important contribution to understanding the traffic demand in segments where this data is not available. It is worth remembering that the service level analysis, pavement projects, road safety, road maintenance, among others, usually depends on the traffic volume, however, the installation of traffic counting stations in all segments of a road network is not viable. In the case of a bus line, the estimate of passenger loading in non-sampled sections is a strong contribution to identifying the critical occupation of the vehicle, commonly used to calculate the size of the bus fleet.

Regarding the volume of passengers along a bus line, Geostatistics has the primary function of estimating the boarding and alighting numbers per bus stop. Concerning this situation, most cities do not have automatic counters and this important information for planning the public transport network depends on very expensive field surveys. In order to overcome this obstacle, Geostatistics allows the municipalities not to carry out such surveys in the entire bus network, but only in a sample of it, which would also guarantee results close to those of the complete survey.

The household approach, which generally focuses on the modal choice stage, is justified by the

impossibility of conducting a household survey that reaches the entire population of a given urban area. On the other hand, understanding how the use of the various transport modes varies throughout the city is fundamentally important for the development of public policies and projects related to mobility. Despite the various types of support already used in the context of transport demand, the application of Geostatistics to new aggregation units is encouraged, in order to stimulate the continuous evolution and scope of the research line under analysis.

On the other hand, the models used are conditioned, among other factors, to the type of variable of interest and purpose of the study. In the modal choice, the dependent variable is categorical, which requires the application of Indicator Kriging or Ordinary Kriging (with data derived from a previous treatment). A Universal Kriging approach to the choice of transport mode was also found. In general, as UK includes explanatory variables in its formulation, it is able to achieve better results than the previous ones, and it presents medium and long-term planning contributions. This might be the reason why UK is the most used model, together with Ordinary Kriging, in applications to travel demand variables in the three steps mentioned above.

The popularity of Ordinary Kriging is due to the fact it is a univariate interpolator, that is, it can be easily applied to the cases in which the only information available refers to the variable of interest, and because it assumes that the data average is unknown and varies locally. In this context, however, it is necessary to assume that the intervening factors of transport demand do not vary or that they demonstrate only a negligible change in the period of time for which the results are valid. Thus, OK applications are restricted to short-term planning only.

In turn, the Spatio-temporal Kriging, whose applications found also do not include explanatory variables, adds to OK the temporal aspect in panel databases, which provide a more improved modeling of spatial variables that also have temporal dependence. Although several transport demand variables undergo significant fluctuations over time, for many of them the difficulty to obtain a panel database makes it impossible to use the Spatio-temporal Kriging. However, when such information is available, the temporal aspect should not be overlooked. Finally, it is noteworthy that Geostatistics, as it generates a continuous surface of estimated values, in addition to providing information of the variable in non-sampled points, also allows visualizing the spatial variation of the variable of interest, facilitating the precise identification of critical regions, spatial patterns and, in the case of multivariate interpolation, associating these patterns with various intervening factors.

In summary, research gaps were identified in the stages of models' validation, comparison with other spatial and non-spatial approaches, use of network distances, applications of Universal Kriging to modal choice variables and selection of predictors for UK. Special attention should be given to the Sequential Gaussian Simulation and the Spatio-temporal Kriging, approaches that should dictate the evolution of the research line under analysis in the coming years.

Some geostatistical models, not yet found in the context of transport demand modeling, also deserve to be highlighted and can/should compose the new evolutionary stage of this line of research. Sequential Indicator Simulation is important to obtain different scenarios when dealing with dichotomous problems (or related to continuous variables divided into classes) effectively useful in the area of Transportation, such as those referring to modal choice variables. Co-kriging can also be a future application field, in order to make comparisons with other multivariate interpolations such as Universal Kriging, for example. Block Kriging and Sequential Direct Simulation are other techniques whose exploration is also timely.

Moreover, with the evolution of technology, implementing Geostatistical algorithms in computer programs has made the application of such interpolations feasible, to the extent that there is no longer any justification for ignoring them. In this context, Table 1 also shows that, in addition to a paid software, such as ArcGIS, Geostatistics has been found in other free and open programs, such as the R programming interface. The easy access to computational resources for using Geostatistics contributes to definitively consolidate the range this tool has in transport demand modeling, with the possibility of continuous expansion.

Acknowledgments

/ 12054-4) and the National Council for Scientific and Technological Development (CNPq, Brazil - Process 304345 / 2019-9).

Authors' Contribution

The first author (Samuel de França Marques) was responsible for Conceptualization, Research, Visualization, Writing - initial draft and Writing - review and editing; the second author (Cira Souza Pitombo) was responsible for Conceptualization, Supervision and Writing - revision and editing.

Conflict of interest

The authors declare no conflict of interest.

References

BISWAS, S.; CHAKRABORTY S.; CHANDRA, S.; GHOSH, I. Kriging-based approach for estimation of vehicular speed and passenger car units on an urban arterial. **Journal of Transportation Engineering, Part A: Systems**, v. 143, n. 3, p. 4016013, 2017. American Society of Civil Engineers. DOI.: 10.1061/JTEPBS.0000031.

CARVALHO, S. D. P. C. E; RODRIGUEZ, L. C. E.; SILVA, L. D.; CARVALHO, L. M. T; CALEGARIO, N.; LIMA, M. P.; SILVA, C. A.; MENDONÇA, A. R.; NICOLETTI, M. F. Predição do volume de árvores integrando Lidar e Geoestatística. **Scientia Forestalis/Forest Sciences**, v. 43, n. 107, p. 627–637, 2015.

CHI, G.; ZHENG, Y. 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. 261–273, 2013. DOI.: 10.1080/15568318.2013.710150.

CHICA-OLMO, J.; RODRÍGUEZ-LÓPEZ, C.; CHILLÓN, P. Effect of distance from home to school and spatial dependence between homes on mode of commuting to school. **Journal of Transport Geography**, v. 72, p. 1–12, 2018. Available in: <http://www.sciencedirect.com/science/article/pii/S0966692316307013>. Access in: jul. 2020.

CRESSIE, N. A. C. **Statistics for spatial data**. John Wiley & Sons, Inc., 1993.

CRESSIE, N.; HAWKINS, D. M. Robust estimation of the variogram: I. **Journal of the International Association for Mathematical Geology**, v. 12, n. 2, p. 115–125, 1980. DOI.: 10.1007/BF01035243.

DEUTSCH, C. V.; JOURNEL, A. G. **GSLIB: Geostatistical Software Library and User's Guide**, Oxford University Press, 1998.

EOM, J. K.; PARK, M. S.; HEO, T.-Y.; HUNTSINGER, L. F. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. **Transportation Research Record**, , n. 1968, p. 22–29, 2006. DOI.: 10.1177/0361198106196800103.

GOMES, M. M.; PITOMBO, C. S.; PIRDAVANI, A.; BRIJS, T. Geostatistical approach to estimate car occupant fatalities in traffic accidents. **Revista Brasileira de Cartografia**, v. 70, n. 4, p. 1231–1256, 2018.

GOMES, V. A.; PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R. Kriging geostatistical methods for travel mode choice: a spatial data analysis to travel demand forecasting. **Open Journal of Statistics**, v. 06, n. 03, p. 514–527, 2016a. DOI.: 10.4236/ojs.2016.63044.

GOMES, V. A.; PITOMBO, C. S.; ROCHA, S. S.; SALGUEIRO, A. R. Estudo comparativo do uso de krigagem para estimação da escolha do modo particular motorizado. **Revista Brasileira de Cartografia**, v. 68, n. 4 SE-Artigos, 2016b. Available in: <http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/44279>. Access in: jul. 2020.

GOOVAERTS, P. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. **Mathematical geology**, v. 40, n. 1, p. 101–128, 2008. Available in: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2518693/>. Access in: jul. 2020.

GOOVAERTS, P. Medical geography: A promising field of application for geostatistics. **Mathematical Geosciences**, v. 41, p. 243-264, 2009a.

GOOVAERTS, P. Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. **Spatial and Spatio-temporal Epidemiology**, v. 1, n. 1, p. 61-71, 2009b.

GRÄLER, B.; PEBESMA, E.; HEUVELINK, G. Spatio-temporal interpolation using gstat. **RFID Journal**, v. 8, n. 1, p. 204-218, 2016. Available in: <https://cran.r-project.org/web/packages/gstat/vignettes/spatio-temporal-kriging.pdf>. Access in: jul. 2020.

GUNDOGDU, I. B.; FACULTY, E. Risk governance for traffic accidents by Geostatistical Analyst methods. **International Journal of Research in Engineering and Science (IJRES)**, v. 2, n. 9, p. 35–40, 2014.

HENGL, T.; HEUVELINK, G. B. M.; ROSSITER, D. G. About regression-kriging: from equations to case studies. **Computers & Geosciences**, v. 33, n. 10, p. 1301–1315, 2007. Available in: <http://www.sciencedirect.com/science/article/pii/S0098300407001008>. Access in: jul. 2020.

VER HOEF, J. M. Kriging models for linear networks and non-Euclidean distances: Cautions and solutions. **Methods in Ecology and Evolution**, v. 9, n. 6, p. 1600–1613, 2018. John Wiley & Sons, Ltd (10.1111). DOI.: 10.1111/2041-210X.12979.

JOURNEL, A. G. Nonparametric estimation of spatial distributions. **Journal of the International Association for Mathematical Geology**, v. 15, p. 445-468, 1983. DOI.: 10.1007/BF01031292.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. Academic press, 1978.

KERRY, R.; GOOVAERTS, P.; GIMÉNEZ, D.; OUDEMANS, P.; MUÑIZ, E. Investigating geostatistical methods to model within-field yield variability of cranberries for potential management zones. **Precision Agriculture**, vol. 17, p. 247-273, 2016.

KLATKO, T. J.; USMAN, S. T.; MATTHEW, V.; et al. Addressing the local-road VMT estimation problem using spatial interpolation techniques. **Journal of Transportation Engineering, Part A: Systems**, v. 143, n. 8, p. 4017038, 2017. American Society of Civil Engineers. DOI.: 10.1061/JTEPBS.0000064.

KRIGE, D. G. 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. 119–139, 1951. Southern African Institute of Mining and Metallurgy.

LINDNER, A.; PITOMBO, C. S. A conjoint approach of spatial statistics and a traditional method for travel mode choice issues. **Journal of Geovisualization and Spatial Analysis**, v. 2, n. 1, 2017. DOI.: 10.1007/s41651-017-0008-0.

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

LINDNER, A.; PITOMBO, C. S.; ROCHA, S. S.; QUINTANILHA, J. A. Estimation of transit trip production using Factorial Kriging with External Drift: an aggregated data case study. **Geo-spatial Information Science**, v. 19, n. 4, p. 245–254, 2016. Taylor & Francis. DOI.: 10.1080/10095020.2016.1260811.

MAJUMDAR, A.; NOLAND, R. B.; OCHIENG, W. Y. A spatial and temporal analysis of safety-belt usage and safety-belt laws. **Accident Analysis & Prevention**, v. 36, n. 4, p. 551–560, 2004. Available in: <http://www.sciencedirect.com/science/article/pii/S0001457503000617>. Access in: jul. 2020.

MANEPALLI, U. R. R.; BHAM, G. H. Crash prediction: evaluation of empirical Bayes and kriging Methods. In: International Conference on Road Safety and Simulation, 3., 2011, Indianapolis. **Proceedings...** Transportation Research Board, 2011.

MARQUES, S. F. **Estimativa do volume de passageiros ao longo de uma linha de transporte público por ônibus a partir da Geoestatística**. 98 p. Dissertação (Mestrado em Engenharia de Transportes) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2019. DOI.: 10.11606/D.18.2019.tde-26042019-110232.

MARQUES, S. F.; PITOMBO, C. S. Estimativa do volume de passageiros ao longo de uma linha de transporte público por ônibus a partir da Geoestatística. **Transportes**, v. 27, n. 3, 2019. DOI:

10.14295/transportes.v27i3.2007.

MATHERON, G. Principles of geostatistics. **Economic Geology**, v. 58, n. 8, p. 1246–1266, 1963.

MATHERON, G. **The theory of regionalized variables and its applications**. Paris: Les Cahiers du Centre de Morphologie Mathematique in Fontainebleu, 1971.

OFFOR, K. J.; VACI, L.; MIHAYLOVA, L. S. Traffic estimation for large urban road network with high missing data ratio. **Sensors**, v. 19, n. 12, p. 2813, 2019. DOI.: 10.3390/s19122813.

ORTÚZAR, J. DE D.; WILLUMSEN, L. G. **Modelling Transport**. John Wiley & Sons, 2011.

PIANUCCI, M. N.; PITOMBO, C. S.; CUNHA, A. L.; SEGANTINE, P. C. L. Previsão da demanda por viagens domiciliares através de método sequencial baseado em população sintética e redes neurais artificiais. **Transportes**, v. 27, n. 4, p. 1–23, 2019. DOI:10.14295/transportes.v27i4.1406.

PITOMBO, C. S.; SALGUEIRO, A. R.; DA COSTA, A. S. G.; ISLER, C. A. A two-step method for mode choice estimation with socioeconomic and spatial information. **Spatial Statistics**, v. 11, p. 45–64, 2015. Available in: <http://www.sciencedirect.com/science/article/pii/S2211675314000645>. Access in: jul. 2020.

REMY, N.; BOUCHER, A.; WU, J. **Applied geostatistics with SGeMS: a user's guide**. Cambridge University Press, 2009.

ROCHA, S. S. **Uso de Algoritmos Genéticos para otimização de modelagem geoestatística aplicada à demanda por transportes**, 2019. Escola de Engenharia de São Carlos - Universidade de São Paulo. DOI.: 10.11606/T.18.2019.tde-23052019-114238.

ROCHA, S. S.; LINDNER, A.; PITOMBO, C. S. Proposal of a geostatistical procedure for transportation planning field. **Boletim de Ciências Geodésicas**, v. 23, p. 636–653, 2017. DOI: 10.1590/s1982-21702017000400042.

ROCHA, S. S.; PITOMBO, C. S.; SALGUEIRO, A. R. Interpolação espacial de viagens urbanas por transporte coletivo através de amostra sistemática desagregada artificial. **Revista Brasileira de Cartografia**, v. 68, n. 4, 2016. Available in: <http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/44277>. Access in: jul. 2020.

SELBY, B.; KOCKELMAN, K. M. 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, 2013. DOI.: 10.1016/j.jtrangeo.2012.12.009.

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

STELZENMÜLLER, V.; EHRICH, S.; ZAUKE, G. P. Impact of additional small-scale survey data on the geostatistical analyses of demersal fish species in the North Sea. **Scientia Marina**, vol. 69, n. 4, p. 587-602, 2005.

WANG, S.; MAO, G. Missing data estimation for traffic volume by searching an optimum closed cut in urban networks. **IEEE Transactions on Intelligent Transportation Systems**, v. 20, n. 1, p. 75–86, 2019. DOI.: 10.1109/TITS.2018.2801808.

WANG, X.; KOCKELMAN, K. Forecasting network data. **Transportation Research Record: Journal of the Transportation Research Board**, v. 2105, p. 100–108, 2009. Transportation Research Board. DOI.: 10.3141/2105-13.

YAMAMOTO, J. K.; LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. Oficina de textos, 2015.

YANG, H.; YANG, J.; HAN, L. D.; LIU, X; PU, L.; CHIN, S.; HWANG, H. A Kriging based spatiotemporal approach for traffic volume data imputation. **PloS one**, v. 13, n. 4, p. e0195957, 2018. Public Library of Science. Available in: <https://pubmed.ncbi.nlm.nih.gov/29664928>. Access in: jul. 2020.

YOON, S. Y.; RAVULAPARTHY, S. K.; GOULIAS, K. G. Dynamic diurnal social taxonomy of urban environments using data from a geocoded time use activity-travel diary and point-based business establishment inventory. **Transportation Research Part A: Policy and Practice**, v. 68, p. 3–17, 2014.

Available in: <http://www.sciencedirect.com/science/article/pii/S0965856414000056>. Access in: jul. 2020

ZHANG, D.; WANG, X. C. Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC. **Journal of Transport Geography**, v. 41, p. 107–115, 2014. DOI.: 10.1016/j.jtrangeo.2014.08.021.

ZOU, H.; YUE, Y.; LI, Q.; YEH, A. G. O. 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. 667–689, 2012. Taylor & Francis. DOI.: 10.1080/13658816.2011.609488.

Authors' biography

Samuel de França Marques is from Porto Nacional–TO, Brazil. He holds a degree in Civil Engineering from the Federal University of Tocantins (2016), he has a master's degree in sciences (2019) and is currently a Ph.D. student in the Graduate Program in Transportation Engineering at the São Carlos School of Engineering, University of São Paulo. He lives in the city of São Carlos–SP, Brazil, where he is dedicated to the application of spatial statistics in modeling/estimating the transit ridership along bus lines. In 2018, he was awarded the ANPET Scientific Production Award. His research interests include models and techniques for transportation planning and road safety.

Associate Professor of the Department of Transportation Engineering of São Carlos School of Engineering, University of São Paulo, Brazil. She carried out her Post-doctoral studies at the Technical University of Lisbon, Portugal (2007-2009) and at the Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, EPFL, Lausanne, Switzerland (2016-2017). Her research interests include: Geostatistics applied to Transportation Planning, New data collection techniques for trip demand forecasting, and traffic accident modeling. She has supervised 18 post-graduate students and has authored 36 publications in international and national journals.