Estimation of Non-Technical Loss Rates by Regions

Lucas Ventura[a], Gustavo E. Felix[b], Renzo Vargas[c], Lucas Teles Faria[b], Joel D. Melo[a]

[a]*The Engineering, Modeling and Applied Social Sciences Center, Federal University of ABC (UFABC), Santo Andre, 09210-580, São Paulo, Brazil*
[b]*Department of Energy Engineering, São Paulo State University (UNESP), Rosana, 19274-000, São Paulo, Brazil*
[c]*Department of Electrical Engineering, São Paulo State University (UNESP), Ilha Solteira, 15385-000, São Paulo, Brazil*

Abstract

Identifying vulnerable regions to non-technical losses allows more assertive combat against them. In this context, this paper presents a spatiotemporal methodology composed of two modules, spatial and temporal, to assist distribution companies in action planning to decrease the rates of non-technical losses by region. The spatial module contains a neighborhood structure based on the similarity among small regions named ”neighborhood by the similarity of attributes,” which improves the characterization of non-technical losses actions performed by end-consumers. That neighborhood structure is incorporated as an input parameter into a hierarchical spatial autoregressive regression model to represent the relationships between inhabitants. On the other hand, the temporal module uses a linear mixed-effects model to consider future values that are subject to the actions of consumers or distribution companies. The proposed methodology is applied to a medium-sized city with approximately 200,000 inhabitants, considering the inspections carried out by a Brazilian distribution utility. The proposal identified the future non-technical loss state in all the city's regions with values greater than 69% of the success rate in identifying NTL to residential, commercial, and industrial consumer classes.

Keywords: Electric power distribution, non-technical losses, spatial data analysis.

Email addresses: lucas.ventura@aluno.ufabc.edu.br (Lucas Ventura), gustavo.felix@unesp.br (Gustavo E. Felix), renzo.vargas@ufabc.edu.br (Renzo Vargas), lucas.teles@unesp.br (Lucas Teles Faria), joel.melo@ufabc.edu.br (Joel D. Melo)

1. Introduction

The non-technical losses (NTL) or commercial losses are related to illegal connections, fraud, energy theft, issues with energy meters such as delay in installing or reading errors, contaminated, defective or non-adapted measuring equipment, low valid estimates, faulty connections, and disregarded customers [1, 2, 3]. Thus, NTL are referred to any electrical energy consumed and not invoiced, resulting in a decrease in distribution companies' revenues. For example, in Brazil, the NTL index was at 5%, and the country lost over US$ 2.4 billion in 2015 [4]. India loses about US$ 4.5 billion every year due to electricity theft, and their loss rate is estimated at 10-40% of revenue [5]. In European Union countries, annual losses are estimated at 2-10% [6]. The annual worldwide financial losses due to NTL are estimated to be around US$ 100 billion [7]. The high unemployment rate, the economic recession, high energy bills, and low human development index create a favorable scenario for NTL growth in several countries worldwide [8].

In many countries, various actions have been taken to reduce the NTL [8]. Among these actions, visits or inspections are the most common actions taken by several distribution companies [9]. Such inspections are carried out considering visiting teams' availability and the most appropriate dates to visit consumers suspected of carrying out some activity classified as NTL. Several methodologies have been proposed to identify the consumers to be visited [10]. Although these methodologies' identification rates can be very high, the dispersion of NTL distribution in urban areas can result in long trips, making it necessary for the visits to be carried out over more extended periods. Also, the heterogeneous socioeconomic distribution in the urban area can make it difficult to correctly identify the consumers who carry out an NTL action [11]. Thus, studies that incorporate geographic space and socioeconomic characteristics in the identification of NTL may be more appropriate in scheduling visits. These studies could also complement other actions such as installing smart meters or using artificial intelligence algorithms to identify the consumer who performs an NTL activity [11, 12].

This work's main objective is to propose a spatiotemporal methodology composed of two

modules for estimating the NTL rate by urban regions. In the first module, a spatial regression is used to characterize the influence of socioeconomic information on NTL's rate. In the second module, a linear mixed-effected model predicts the NTL rate's future value by regions.

The spatial regression helps in decision-making problems when the behavior of a dependent variable (response variable) is correlated with independent variables (explanatory variables) [13, 14]. In the NTL performed modeling, the response variable is represented by the loss rate in each region, while the explanatory variables are the regional and socioeconomic characteristics. The loss rate is defined as the total NTL detected by each consumer unit (CU) to the total number of CU visited or inspected. Thus, this kind of regression allows determining the spatial distribution of NTL in the study zone from each year's visits. In this work, we call this spatial distribution the present state of the NTL.

In the NTL problem using the regression technique, one of the challenging tasks is to choose the explanatory variables that best characterize the loss rate. Among the available regression models, the hierarchical spatial autoregressive (HSAR) model can better estimate NTL's spatial distribution because the degree of influence of the explanatory variables' set is separated into several levels to adjust their parameters. Such separation reduces the weighting of the variables' coefficients in the spatial estimation of the loss rate. However, an adequate calibration of parameters of HSAR that allow characterizing the influence among the regions must be performed. For this, a weighting matrix by the similarity of attributes (WMSA) can model the neighborhood in the urban zone. Thus, in the proposed methodology, the WMSA is used as an input for HSAR, seeking a better characterization of the neighborhood structures formed to carry out NTL activities. Another advantage of the proposed methodology is its capacity to estimate the NTL rate of the regions that have not been inspected or visited. In this way, it will be possible to cover all city areas, estimating each region's NTL rate in the present state. Likewise, the WMSA that considers the influence of socioeconomic variables and the HSAR multilevel structure that seeks to identify the spatial pattern for NTL actions helps to correct low values of the loss rate in regions with a high number

of visits and a low number of detected frauds.

Linear mixed-effects models are recommended when working with fixed and random effects in collected data at several time points [15]. This recommendation is because these models estimate future values, considering the dynamics from a previous period to a later period within the prediction intervals [16]. In the current digitalization scenario in several government agencies, the explanatory variables are available in public databases because several countries generally conduct demographic census research to obtain information to help public policies [17]. Likewise, some municipalities carry out surveys at the zone level to direct master plans for the sustainable development of cities [18]. Also, as shown in [19], there can be a high correlation between the variables and NTL actions for short periods. Therefore, the spatial regression results can be inserted in a linear mixed-effects model to predict the future state of the NTL. The incorporation of geographic space to estimate NTL helps distribution companies to plan actions to decrease NTL rates by region. Also, this estimate's results can be used as an input database for other soft computing techniques to improve NTL identification performance in areas with high NTL values.

1.1. Literature Review

Several artificial intelligence techniques have been proposed to assist distribution companies in planning visits to detect NTL [20]. In recent years, these techniques have been used according to some objectives of the distribution companies. For instance, reference [21] presents a methodology based on a Bayesian risk structure, random forest, support vector machine, and artificial neural network (ANN) to recover the revenue lost due to NTL. This methodology detects potential irregular consumers based on their consumption profile.

On the other hand, hybrid techniques have been proposed to improve the success rate in identifying NTL. For example, in [22], the authors have used a hybrid neural network model from sequential and non-sequential data. This proposed architecture consists of an extended short-term memory network that analyzes the daily history of energy consumption to treat sequential data. Also, for non-sequential data, there is a multi-layer perceptron network that integrates data such as

contracted energy or geographic information. These hybrid neural network results helped schedule inspections for network users and showed 47% accuracy in detecting irregular consumers.

The methodology in [23] involves two coordinated modules. The first performs consumer filtering based on text and data mining and an ANN. The second makes use of data-mining techniques and their contract information. The result of these models is a list of consumers selected for inspection. In addition, consumers are grouped according to similar characteristics that describe them, such as economic activity, geographic location, contracted power, but no further study on the influence of these variables on NTL has been carried out.

In [24] the authors have used data mining techniques in the NTL problem with machine-learning classifiers. The ensemble methods and the ANN methods showed the best results in NTL detecting. Also, 71 data elements from the company's database about customers were evaluated to be implemented in the classifier algorithms. This classification allows for a 77% improvement in the NTL forecast.

H. Long et al. [25] use a data-based method to identify NTL, including at which feeder the loss was recorded, the position, and the time at which it occurred. This method is used based on the daily data of energy supply and sales of electricity and the analysis of the characteristics of energy consumers' load curves. The results of this method suggested that they can effectively detect abnormal power losses in the distribution network.

Due to the modernization of electrical networks, smart meters can be used to detect NTL. In this modernization scenario, the NTL occurs in the form of cyber-attacks on companies' digital databases and the digital breach of smart meters is presented in [26]. The authors have proposed a strategy to detect NTL using a multivariate control chart that establishes a reliable region for monitoring the measured variance. After detecting NTL, a path search procedure based on the A-Star algorithm can locate the consumption point that targets the NTL's cyber-attack.

The methodologies mentioned above seek to identify the CU that carries out fraud. Although the success rate can be satisfactory for the distribution company's goals, the risk of misidentifi-

cation can cause legal problems for the distribution company. In this way, some companies visit regions, check electrical infrastructures and consumption bases, improve the success rate, and reduce fraud consumers' false identification. Thus, as shown in the methodology [11], socioeconomic variables can help direct teams to the most likely NTL regions, improving their success rates.

1.2. Identification of NTL by Regions

To exemplify the relevance of identifying NTL by region, consider that NTL in a city is concentrated in an urban zone with 100 consumer units (CUs) divided into five areas. This zone presents 15 CUs where each region has consumers carrying out NTL. Using a high-efficiency technique shown in [17] for this zone, the distribution company identified 12 CUs that carried out NTL. However, suppose the three unidentified CUs are located in one region. In that case, the value of losses in this region is high compared to the other areas, resulting in urban segregation that could lead to other social problems. In Table 1, we place the values of the loss rate by region (LRR) before and after identification. LRR is the ratio between CU with NTL detected and total CU in an area. We can see that there is a risk that fraud will be concentrated in region C, resulting in 30% of LRR. Moreover, as shown in [23], in some countries, the concentration of NTL has brought other social problems and reduced the quality of electricity in the region with a high concentration of NTL. However, discussing these problems is outside the scope of this work.

Suppose the distribution company performs region identification before applying an identification methodology, the risk of concentrating NTL decreases because of the success rate depending on pattern recognition dispersion. In this way, the purpose of our proposed method is to incorporate socioeconomic characteristics through a spatiotemporal analysis by regions, correlating these characteristics with the frauds found by distribution transformers. The proposal estimates each region's possibility of having NTLs in the present and future states according to the degree of correlation between variables or the influence that certain variables have.

Table 1: Example of NTL assessment by region

Regions	**Total CU**	**CU with NTL Detected**	**CU with Undetected NTL**	**LRR Before Detection**	**LRR After Detection**
A	20	3	0	15.0%	0.0%
B	15	2	0	13.3%	0.0%
C	10	0	3	30.0%	30.0%
D	15	1	0	6.7%	0.0%
E	40	6	0	15.0%	0.0%

1.3. Contributions

The main contributions of this work are explained in the following:

1. The proposed methodology models the relations between the inhabitants to carry out NTL actions by neighborhood structure based on the similarity of attributes among areas. This modeling improves the characterization of the influence of NTL actions in several regions that are close or not close within urban zones.
2. The HSAR model allows estimating the spatial distribution of NTL in the present, even in unvisited regions or without recorded data, due to similarity with other areas with CUs that carried out NTL actions.
3. A linear mixed-effects model is used to estimate NTL's distribution in the future and allows power utilities may plan actions to combat the NTL beforehand. Moreover, this proposed model is suitable when some collected databases are not extensive for the NTL problem.

1.4. Paper structure

The rest of the paper is as follows. Section 2 describes the methods and models used in the proposal to estimate NTL. Section 3 explains and discusses the obtained results. Finally, Section 4 presents the conclusions.

2. The Proposed Spatiotemporal Estimation for Non-Technical Losses

The proposed methodology uses as input data: NTL values registered by smart meters or by inspections carried out in recent years by the distribution company and their geographic location; the geographic information layer with the demographic census information; and the parameters related to HSAR and linear mixed-effects model to characterize the spatial relationships and temporal evolution of the loss rates in the urban zone.

2.1. Spatial Module: Hierarchical Spatial Autoregressive Regression using Neighborhood by Similarity of Attributes

The neighborhood structure among regions [27] is necessary for a study involving spatial data analysis with aggregated data by subarea. For example, the spatial variability in the data aggregated in n regions is represented by the weighting matrix $\mathbf{W}_{(n\times n)}$. The elements w_{ij} of the weighting matrix represent a weighting measure between the regions S_i and S_j , which are directly influenced by the chosen neighborhood structure.

In the spatial data analysis [28], Tobler's first law is used to characterize spatial dependence in all directions, considering the criteria of the Euclidean distance-based weighting matrix among regions [29]. However, there is no evidence that these criteria adequately represent NTL's problem because the neighborhood of inhabitants who can carry out NTL activities does not necessarily need to have common boundaries. In this context, the neighborhood among regions called the neighborhood by the similarity of attributes (NSA) could be more suitable to characterize the neighborhoods. This neighborhood is based on the degree of similarity among regions concerning their characteristics represented by socioeconomic variables and distribution networks, which is one of this work's contributions. The NSA expands the traditional neighborhood notion based on the Euclidean distance among the centroids of regions. It allows distant regions to exert more significant influence than nearby regions and vice versa.

2.1.1. Self-Organizing Maps

Self-Organizing Maps (SOM) [30, 31] are used to build an NSA. SOM define clusters that contain regions with similar attributes. All regions that are included in the same group belong to the same neighborhood. SOM is an ANN with lattice architecture, competitive unsupervised learning, which identifies patterns in multivariate data vectors [32]. Thus, SOM is a mapping that provides a structural representation of input data through neuron weight vectors. It is characterized by forming a topographic map of input patterns where neuron locations map the intrinsic statistical characteristics of input patterns. The $\mathbf{x} \in \mathbb{R}$ input pattern is mapped in the two-dimensional output space by the weight vector $\mathbf{w}$ whose location is a function of $\mathbf{x} - \mathbb{R}^n \Rightarrow \mathbb{R}^2$.

2.1.2. Context Map

After SOM training and operation, neurons are grouped into more extensive and more representative clusters of the problem for constructing the context map. Our proposal uses a criterion inspired by the SOM training algorithm to build the context maps. Thus, it is not necessary to specify the number of clusters. In NTL, the exact number of clusters that must be formed *a priori* is unknown; therefore, SOM is an appropriate tool for creating clusters of regions with similar attributes. The algorithm for training and operation of the SOM is described in [32]. An intermediate step for the construction of an NSA is the definition of an interneuron neighborhood. Two-dimensional topological maps commonly have rectangular or hexagonal arrangements. Fig. 1 presents an illustrative example of a rectangular two-dimensional grid of $4 \times 4 - 16$ neurons. In the proposal, each city region is associated with a two-dimensional map neuron. For example, in Fig. 1, the regions belonging to *cluster* A are associated with one of the neurons in a $\Omega^{(R=1)}$ set, which is part of the same NSA.

There are several strategies for constructing these context maps as statistical techniques and expert knowledge [32]. We show the steps for the *cluster A* construction to clarify the criteria to build the context map in Fig. 1. *Step 1)* The first neuron (neuron 1) enters *cluster A*. *Step 2)*

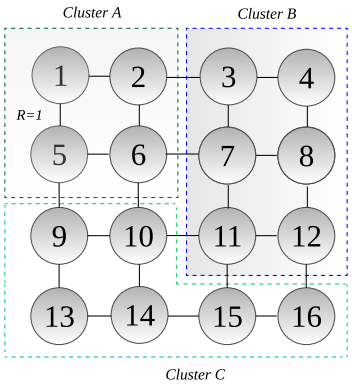

Figure 1: Two-dimensional rectangular map (4x4 neural grid with 16 neurons) with three clusters of neurons A, B, and C highlighted by dotted polygons.

The last neuron that entered *cluster A* is declared current neuron (*cn*). *Step 3)* The neurons in the *cn*'s neighboring compete to enter *cluster A*. The candidate neuron that enters *cluster A* is the *cn*'s neighbor (cnn*) whose weight matrix is closest to the *cn*'s weight matrix for the Euclidean norm using (1). *Step 4)* If the candidate neuron that enters *cluster A* was added to the cluster previously, then the cluster construction is finished; otherwise, the candidate neighbor neuron is added to *cluster A*. Go to *Step 2)*.

$$cnn^{*} = \underset{k \in \Omega_{cn}^{(R=1)}}{\text{minimize}} \; \| \mathbf{w}_{cn} - \mathbf{w}_{k} \| \tag{1}$$

2.1.3. Weighting Matrix by Similarity of Attributes

In the proposed methodology, the weighting matrix is used to estimate the spatial variability of the data. Considering a set of n regions $\{S_1, \cdots, S_n\}$, the $\mathbf{W}_{(n\times n)}$ weighting matrix is constructed where the elements w_{ij} represent a measure of weighting measure between the regions S_i and S_j. To illustrate the obtaining of the elements w_{ij} of the WMSA $\mathbf{W}_{(n\times n)}$, consider the context map of Fig. 1. After the SOM training, each input sample $\mathbf{x}_k$ is associated with $k = 1, \cdots, n_s$ for each of the 16 neurons in the two-dimensional map where the number of city regions is n_s . Each neuron in the neural grid $n = 1, \cdots, 16$ is associated with a weight vector $\mathbf{w}_n$. *Clusters A, B,* and *C* are represented by the winning neuron's weight vectors $\mathbf{w}_{n*}$, where $n*$ is the winning neuron. This

neuron is associated with a more significant number of input samples $\mathbf{x}_k$ when compared to other neurons in the same cluster. The element w_{ij} of WMSA $\mathbf{W}_{(n\times n)}$ represents the weighting between regions i and j that are represented by the vector of variables $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Suppose that the regions i and j belong to the same *cluster A*; therefore, they belong to the same NSA, then $w_{ij} \neq 0$; otherwise, $w_{ij} = 0$.

In Fig. 1, we assumed that regions i and j are in the same neighborhood; therefore, they are associated with any of the four *cluster A* neurons. Thus, w_{ij} represents a degree of similarity of regions i and j with *cluster A* that is represented by the weight vector of their respective winning neuron $\mathbf{w}_{n*}$. In (2), w_{ij} is the complement of the arithmetic mean between the Euclidean norm of the difference between the variables $\mathbf{x}_i$ and $\mathbf{x}_j$ and the weight vector of the winning neuron representing *cluster A* and designated by $\mathbf{w}_{n*}$. The greater the similarity between the inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ with the weight vector of the winning neuron $\mathbf{w}_{n*}$ *cluster A* representative to which both samples belong, the more significant will be $w_{ij} \in [0, 1]$, and $w_{ij} = w_{ji}$, as $\mathbf{W}_{(n\times n)}$ is a symmetrical matrix.

$$w_{ij} = 1 - \frac{1}{2} \sum_{k\in\{i,j\}} \|\mathbf{x}_k - \mathbf{w}_{n*}\| \tag{2}$$

2.1.4. Hierarchical Spatial Autoregressive Regression

Spatial regression models consider the geographic data on a single structure or only on one observation level [33, 34]. However, many data sets have multilevel structures. For example, the relationships between inhabitants within a neighborhood can be characteristic at one level. Likewise, neighborhood relationships can be analyzed at another level, with information aggregated at each observation level. The term relationship is associated with the degree of interaction between the variables and their influence on identifying urban dynamics [35].

On the other hand, in multilevel modeling, this observation level can be understood as a hierarchical structure since the effects and influences of variables on an object of interest can differ depending on the disposition of the objects and the observation level at which they occurred [35, 36].

In the spatial estimation of NTL, a hierarchical data structure allows characterizing the CU relationships within the same region and how the city zones are related. Several HSAR applications [37, 38] show that such a division into levels allows for characterizing the spatial pattern of urban dynamics for databases with correlated variables. In the proposal, there are two hierarchical levels. The low level characterizes CU relationships, while the highest level models the relationships between the city's different zones. The HSAR model estimates the loss rate (Y_{ij}) in each sector or region of the location ij of the city by:

$$Y_{ij} = \rho w_{ij} Y_{\text{rate}_{ij}} + x_{ijk}\boldsymbol{\beta}_k + z_{jk}\boldsymbol{\gamma}_k + \boldsymbol{\Delta}_j\theta_j + \boldsymbol{\varepsilon}_{ij} \tag{3}$$

$$\theta_j = \lambda \mathbf{M}_j\theta + \boldsymbol{\mu}_j \tag{4}$$

where $Y_{\text{rate}_{ij}}$ is a loss rate calculated for the location ij; ρ is an autoregressive spatial parameter that indicates the strength of spatial interactions at the low level; w_{ij} is an element of the WMSA or spatial matrix of low-level weights; x_{ijk} is an element of the matrix with the values of the explanatory k-variables; $\boldsymbol{\beta}_k$ is a vector of regression coefficients to be estimated; z_{jk} is an element of the matrix with the values of the explanatory k-variables in the high level; $\boldsymbol{\gamma}_k$ is a vector of coefficients related to the high level to be estimated; $\boldsymbol{\Delta}_j$ is the random effect matrix that links the low and high-level random effect vector to the response variable; λ is an autoregressive spatial parameter that indicates the strength of spatial interactions at the high level; $\mathbf{M}_j$ is the matrix of high-level spatial weights; θ_j is an element of the vector of random regional effects; $\boldsymbol{\varepsilon}$ and $\boldsymbol{\mu}$ are vectors of random errors, or residuals, that follow a normal distribution, at the high and low levels, respectively. The Y_{ij} value calculated by HSAR considers consumer interactions within a sector (low level) and interactions between regions (high level). From the modeling of these interactions, it is possible to find a value of Y_{ij} in regions that the distributor's teams have not visited.

Fig. 2 shows the methodology proposed in the Spatial Module for NTL estimation in the present state.

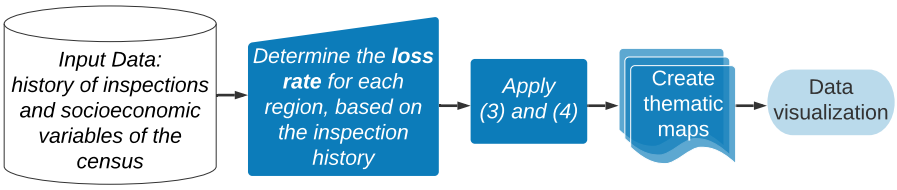

Figure 2: HSAR methodology outline for estimating NTL in the present state.

To estimate the future state of losses, the value of Y_{ij} is inserted as input data for the linear mixed-effects model.

2.2. Temporal Module: Linear Mixed-Effects Model

Linear mixed-effects models have been used in several areas of knowledge to estimate and predict a system's behavior [39, 40]. For example, in statistical analysis, these models are used when working with hierarchical [41], longitudinal [42], and non-independent data [43], also when the value observed at one time depends on the one(s) observed before [44].

The techniques involved in the linear mixed-effects model use historical data to describe and understand the system parameters and make predictions. The application of this model is extended to several statistical analysis software. Two data sets are considered in the modeling performed to characterize the NTL rate's future value for each region. The first is formed by the values estimated by the HSAR of the loss rates of the last three years. The second set considers the values defined as a goal by the distribution company for the same years, being called expected values. In general, distribution utilities expect the NTL level to reach a goal in some regions. Thus, the proposed methodology generates three random numbers that follow a normal distribution with an average equal to the expected NTL level and a deviation that can be calibrated according to the success of the values defined as the desired level. These random numbers seek to characterize variations that may occur in the value designated as a goal for the NTL due to actions taken by the distribution companies or consumers, called disturbances. Likewise, this generation of random numbers is controlled [45] because the numbers follow a normal distribution with information

from the distribution companies' target and the maximum deviation in previous years. The three years were chosen after several observations in the temporal analysis of NTL's evolution [19, 46]. However, the period can be changed according to the planners' experience.

Additionally, we considered that future values could be modified with increments or decreases because of the dynamics of consumers' actions or the distribution companies' combat actions, respectively. In this way, a function that modifies these rates' expected values, producing disturbance values with lags and leads, is considered. This function is available in several statistical packages, such as R [47], Minitab [48], to analyze the sensitivity of random disturbances in the response variable. Thus, after applying this function, we have a database for each region with a size equal to [3 × (2 + P)], with P being the number of disturbances considered. This database is a matrix divided into columns, the first column being the loss rates estimated by the HSAR. The second column being the expected values of NTL for the same years of HSAR, and the random disturbances form the other columns. The linear mixed-effects equation is shown in expression (5).

For example, $\mathbf{X}_{ij}$ corresponds to a matrix for a region at location (i, j). An example of this division in columns of attributes for $\mathbf{X}_{1,2}$ is shown in Fig. 3. It is important to note that Fig. 3 shows the ideal data arrangement for applying the linear mixed-effect model. As such, two sets of input data must be considered. The first set represents the estimated values by the HSAR model for the years 2009 until 2011. The second set represents the expected values of non-technical losses in each sector that were defined as a goal by the distribution company in each sector for the same years. A correlation analysis is carried out looking between the first column and the others to find the most correlated column with the loss rates estimated by the HSAR.

	LOSS RATES	EXPECTED VALUES	DISTURBANCES	
2009	0.2593	0.2957	0.2127	0.2127
2010	0.0573	0.3028	0.2957	0.2127
2011	0.0000	0.0395	0.3028	0.2127

Figure 3: An example of the matrix for the one sector X at a location (1,2).

By considering $P = 2$, we have for each sector X located in position i, j (X_{ij}) a matrix of dimension [3 × 4]. For example, consider that the region (1,2) of a city has a matrix, as shown in Fig. 3. Considering the above, two sets of input data must be placed in the matrix of Fig. 3. The first set represents the estimated values by the HSAR model for the years 2009 until 2011. The second set represents the expected values of non-technical losses that were defined as a goal by the distribution company in each sector for the same years. After placing these databases, we add other columns to represent the increments or decreases that can be had in the expected value that was placed in the second column. Such added columns allow for characterizing the dynamics of actions by distribution companies or consumers that can modify the expected values of non-technical losses. As a result, we have a matrix of size [3 × (2 + P)]. By considering P=2, we have for each region X located in position (i,j) a matrix (X_{ij}) of dimension [3×4]. After filling in all the columns of the [3x4] dimension matrix, a correlation analysis is performed between the first column and the others to find the column most correlated with the loss rates estimated by HSAR. The most correlated column is considered in the linear mixed-effects equation, as shown in expression (5).

$$\hat{Y}_{ij}(tk) = \beta_{ij} x_{ij} + \varepsilon_{ij} \tag{5}$$

The average value of $\hat{Y}_{ij}(tk)$ from Equation (5) for location (i, j) in the map represents the NTL's future state. This value considers the disturbances due to variations in consumer actions and the combat actions of distribution companies, and have the most significant correlation with the spatial influence characterized by HSAR; x_{ij} is the most correlated column with the actual losses data for the location ij; β_{ij} is the weighted parameter associated with x_{ij}; ε_{ij} is the error term on the NTL state for the location ij; tk is an integer representing the forecast horizon to be defined by the planner. In (5), we assumed that $\hat{Y}_{ij}(tk) \sim \mathcal{N}(\mu_{ij}, \sigma)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, where μ_{ij} is the conditional mean calculated for each time point t using the maximizing likelihood of:

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_{ij,t} - \mu_{ij,t}\right) \tag{6}$$

where σ^2 and σ are the variance and the standard deviation, respectively, estimated based on likelihood; $y_{ij,t}$ is the present state of NTL rate calculated for the location ij and time point t; T is the database size of losses data for the location ij.

Fig. 4 presents a summary of the methodology applied to the temporal module to estimate NTL's future values.

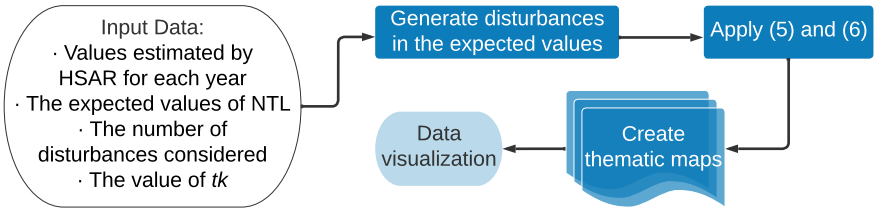

Figure 4: Linear mixed-effects model framework overview.

3. Test and Results

The proposed methodology was applied using socioeconomic information from a medium-sized city in São Paulo, Brazil. Census sectors group the data. For this, each region of the city corresponds to a census sector. The socioeconomic data used in the application is shown in Table 2, which is available in the IBGE's Household Census by Subareas [49]. On the other hand, the inspection carried out by the teams of a distribution company in the years 2009 to 2012 was used.

Table 2: Variables for Construction of the Neighborhood by the Similarity of Attributes

Variables	**Description**
Average Income	Nominal average monthly income of persons ten years old or older (with and without income)
%Rented Residency	Percentage of rented private households
Load Density (KVA/km²)	The ratio between the sum of the rated power of the transformers and the total area in km^2

In the inspection history, there is information about each CU visited, including the NTL identification record. There were 2444, 1080, 3781, and 1920 CUs inspected in 2009, 2010, 2011, and 2012, respectively. Additionally, 164, 86, 454, and 422 NTL were identified in 2009, 2010, 2011 and 2012, respectively. In that same period, according to the demographic census, the city had a population of approximately 200,000 inhabitants, totalizing almost 80,000 CUs. NTL actions were mostly identified at the low voltage network. Irregularities registered in CUs connected to the medium voltage grid represented 0.54% of the total NTL activities, with the remainder registered in CUs connected to the low voltage network. There is no record of the number of smart meters in this period in the city.

As input data for the spatial regression, we use the NTL rate per sector, which is calculated using the ratio between the total number of irregular CU per sector and the total inspections carried out per sector. All of the parameters of the proposed methodology are calculated using the HSAR [50] and the grey-box package [51], available in software R [47]. In this way, the HSAR was applied to estimate NTL's present state for each year with information about inspections by the utility. The response variable Y_{ij} represents the present state of the loss rate values of NTL expected in each sector at a location (i, j) of the city. Finally, the linear mixed-effects model was applied to estimate the future state of losses rate. The simulations were made using R 3.5.3 on a 64-bit Windows Server laptop with a 2.20-GHz Intel Core i5 processor and 4 GB RAM. The HSAR model application for the years 2009 to 2011 is considered as input data for the linear mixed-effects model to estimate the future state of NTL, so we define the number of disturbances as equal to 3. Lastly, we want to estimate the probability of losses for the next year, 2012, regarding the known information about the inspections, so we consider the constant tk equals 1 in (5).

3.1. Construction of the Neighborhood by Similarity of Attributes

The NSA is built from sectors based on the similarity of attributes: *Average Income*, *Load Density* (kVA/km2), and *%Rented* CU. These variables are described in Table 2. T. B. Smith [8] has associated the NTLs with socioeconomic vulnerability. This socioeconomic characteristic is

represented in this study by the variable of *Average Income*. *Load Density* is related to NTLs because it is more likely to find irregular CUs in sectors with higher *Load Density*. Finally, the variable *%Rented Residency* is related to NTLs, because malicious individuals residing in rented CUs may implement meter fraud and be more easily covered.

3.2. Construction of Weighting Matrix by Similarity of Attributes

The WMSA $\mathbf{W}_{(n\times n)}$ is a symmetrical matrix 301×301 with a unitary main diagonal. WMSA's dimension is a function of the number of sectors in the city under study. In Fig. 5 shows the city's urban area map with neighborhood structure based on the similarity of attributes after SOM execution with 6×6 hexagonal topology. In Fig. 5, there are eleven sector clusters with similar characteristics, and the number of sectors in each cluster in parentheses. Cluster 4 (in green) includes 64 sectors in the same NSA, for example. It can be seen from Fig. 5 that the NSA expands the notion of traditional neighborhoods based on proximity among sectors (Euclidean distance). Moreover, it allows distant sectors to belong to the same neighborhood. The description of parameters for SOM training and operation is presented in Table 3. SOM was developed based on the recommendations in [38], [35] and with a computational time of 467s.

Table 3: Parameters for Self-Organizing Maps (Som)

Parameters	**Descriptions**
Neural network training	Competitive learning. Competition among neurons for each entry sample
Training Type	Unsupervised
Network Topology	hexagonal lattice (36 neurons)
Convergence Criteria	Variation of the magnitude of neural net weight vectors among consecutive training epoch (precision 10-6)
Training Ratio	0.1 (exponential decay with increased training ratio)
Neighboring Neurons Adjustment Function	Gaussian
Interneuron Neighborhood Topology	Unit radius circle with fixed neighborhood

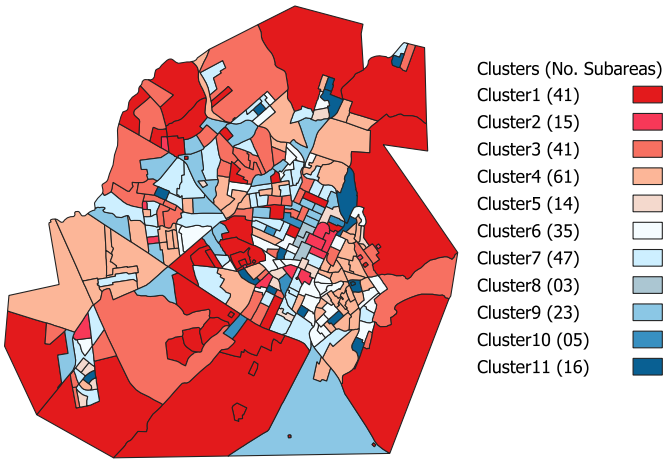

Figure 5: Neighborhood by the similarity of attributes from the hexagonal grid SOM 6×6 – 36 neurons for urban sectors in the city evaluated.

3.3. Results on NTL Estimation

Considering the definition explained in the previous subsections, it was possible to estimate the NTL values from 2009 to 2011, from the HSAR Model, which configure the present state, and the year 2012, from the Linear Mixed-Effects Model, which corresponds to the future state. Each census sector of the urban zone is characterized within a range of loss states that varies in the range from 0 to 1 according to the results obtained in the spatiotemporal estimation methodology. Sectors whose estimation is close to 0 have less chances of finding NTL, in contrast to sectors whose estimation gives values close to 1, that is, more vulnerable to NTL. Probability intervals were used to classify the states of the city's sectors into: regular [0, 0.15], attention (0.15, 0.35], and critical (0.35, 1]. The upper and lower limits for each interval are the same limits explained and used in [11]. In this way, it is possible to count the number of sectors found in each state range.

Fig. 6 presents, through thematic maps of the city, the characterization of each census sector of the city within a range of states, comprising the regular, attention, and critical states, for all consumption classes (residential, commercial, and industrial). The maps (a), (b), and (c) represent

the results for the years 2009, 2010 and 2011, respectively; and map (d) represents the result for the year 2012. In [11], a proposal for spatiotemporal estimation for NTL was presented using the Generalized Additive Model (GAM) [52] method to determine the present state of losses for 2011 and the Markov Chains to estimate the future state for 2012. For comparison purposes, the methodologies for all consumption classes presented in this paper are compared with those of [11]. Table 4 and Table 5 show the percentages obtained in each state for each method, in the present and future states.

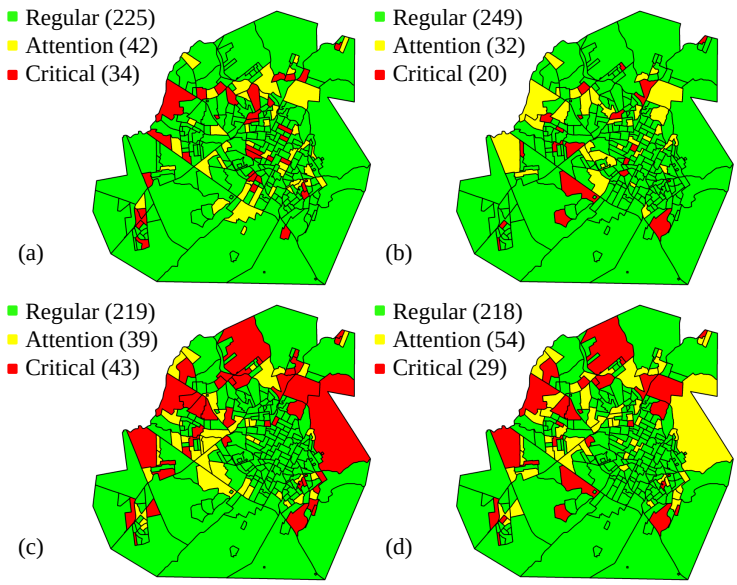

Figure 6: HSAR results for all the classes of consumers for the years 2009, 2010, and 2011, which are (a), (b), and (c), respectively. (d) Result of the linear mixed-effects model application using the same neighborhood matrix for the year 2012.

Table 4 and Table 5 show the comparison between the results obtained by the method proposed in this work with the methods used by [11]. In both studies, methods were used to separately estimate the probability values for the present state (years 2009 to 2011) and the future state (the year 2012), considering a range of values that represent the state of each census sector as regular,

Table 4: Comparison Between the Proposals in the Present State (HSAR vs . GAM)

State of Losses	HSAR	GAM
Regular [0, 0.15]	72.76%	86.71%
Attention (0.15, 0.35]	12.96%	3.98%
Critical (0.35, 1]	14.28%	2.33%

Table 5: Comparison Between the Proposals in the Future State (Linear Mixed-Effects Model vs. Markov Chains)

State of Losses	Linear Mixed-Effects Model	Markov Chain
Regular [0, 0.15]	72.43%	87.71%
Attention (0.15, 0.35]	17.94%	3.65%
Critical (0.35, 1]	9.63%	1.66%

attention, or critical. In [11], the GAM method was used to estimate the current state, whereas the HSAR model is used with the same objective in this work. One of the advantages of HSAR over GAM is that the first one allows estimating values for all sectors, even for unvisited regions. In contrast, GAM estimates values only for sectors that are visited (in total, 21 sectors were not inspected in the 2009- 2012 period). Thus, Table 4 shows that the sum of the GAM percentages does not account for 100% (total number of sectors). The remaining amount, equal to 6.98%, corresponds to 21 sectors from which it was impossible to characterize using GAM. Consequently, these same sectors were not analyzed in the Markov Chain methodology.

Another fundamental difference is the use of the traditional neighborhood by proximity to estimate NTL's future state via Markov Chains in [11]. Therefore, in that reference, the NTL estimation was more conservative with fewer sectors in attention and critical states of NTL, as shown in Table 5. On the other hand, the methods presented in this paper were also compared with the NTL values obtained from the concessionaire's inspection history for the year 2012. In this comparison, we used the success rate in identifying NTL belonging to regular, attention, or critical

states. The success rate is calculated as the ratio between the number of sectors correctly identified in each state by the total census sectors in the urban zone. Since each electricity consumer sector has its particular dynamic, the industrial class relationships are different from those of residential and commercial consumers. From these relationships, we considered making an analysis separately for each consumer sector. This analysis is possible because, in the concessionaire's inspection history, each CU has a code according to its consumption class. Thus, the calculation of the success rate for each sector was carried out separately. When we calculated the residential, commercial, and industrial classes' success rate, referring to 2012, the values obtained were 69%, 88%, and 80%, respectively. Thus, the proposal has a more effective success rate for commercial and industrial consumers since these consumers account for a relatively small portion of the city. In general, distribution companies are more interested in identifying commercial and industrial consumers because the economic return values will be higher because of the high energy losses.

4. Conclusion

The study of geographic space was incorporated into detecting non-technical losses (NTL) in this work. It was incorporated into an efficient method to estimate the present state (via a hierarchical spatial autoregressive model, HSAR) and the future state (via a linear mixed-effect model) of the NTL by subareas. A neighborhood structure based on the similarity of attributes by subareas was proposed. Neighborhood by the similarity of attributes (NSA) expands the traditional neighborhood notion and allows distant subareas to belong to the same neighborhood. The HSAR model allowed the estimation of NTL in the present in all subareas, even in subareas without inspected CUs, classifying more than 9% of the city's regions as critical areas in the present state. The linear mixed-effect model is suitable for small databases; therefore, it is ideal for this problem and allows estimating NTL distribution in the future. Also, compared to another study, it was revealed that the application of the methods separately for different classes of consumers reached higher values of effectiveness. The proposal identified the future NTL state in all city's regions,

classifying them as regular, attention, and critical with values greater than 69% of the success rate in identifying NTL to each consumer class. Therefore, the incorporation of geographical space to the problem of NTL provides a powerful and complementary tool to combat them in conjunction with detection tools that involve soft computing techniques.

References

[1] F. d. S. Savian, J. C. M. Siluk, T. B. Garlet, F. M. do Nascimento, J. R. Pinheiro, Z. Vale, Non-technical losses: A systematic contemporary article review, Renewable and Sustainable Energy Reviews 147 (2021) 111205. doi:10.1016/j.rser.2021.111205.
URL https://linkinghub.elsevier.com/retrieve/pii/S1364032121004937

[2] D. Carr, M. Thomson, Non-technical electricity losses, Energies 15 (6) (2022). doi:10.3390/en15062218.
URL https://www.mdpi.com/1996-1073/15/6/2218

[3] X. Lu, Y. Zhou, Z. Wang, Y. Yi, L. Feng, F. Wang, Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid, Energies 12 (18) (2019). doi:10.3390/en12183452.
URL https://www.mdpi.com/1996-1073/12/18/3452

[4] Instituto Acende Brasil, Perdas comerciais e inadimplência no setor elétrico (18) (2017) 1 – 40.
URL http://www.acendebrasil.com.br

[5] S. Kumar V., J. Prasad, R. Samikannu, Overview, issues and prevention of energy theft in smart grids and virtual power plants in Indian context, Energy Policy 110 (2017) 365–374. doi:10.1016/j.enpol.2017.08.032.
URL https://doi.org/10.1016/j.enpol.2017.08.032

[6] CIRED: Working Group on Losses Reduction CIRED WG CC-2015-2, Reduction of technical and non-technical losses in distribution networks, in: Proc. 2017 Int. Conf. on electricity distribution, 2017, p. 114.

[7] K. Dasgupta, M. Padmanaban, J. Hazra, Power theft localisation using voltage measurements from distribution feeder nodes, IET Generation, Transmission & Distribution 11 (11) (2017) 2831–2839. doi:10.1049/iet-gtd.2016.2011.
URL https://onlinelibrary.wiley.com/doi/10.1049/iet-gtd.2016.2011

[8] T. B. Smith, Electricity theft: A comparative analysis, Energy Policy 32 (18) (2004) 2067–2076. doi:10.1016/S0301-4215(03)00182-4.
URL https://linkinghub.elsevier.com/retrieve/pii/S0301421503001824

[9] J. L. Viegas, P. R. Esteves, R. Melício, V. M. Mendes, S. M. Vieira, Solutions for detection of non-technical

losses in the electricity grid: A review, Renewable and Sustainable Energy Reviews 80 (2017) 1256–1268. doi:10.1016/j.rser.2017.05.193.

[10] T. Ahmad, H. Chen, J. Wang, Y. Guo, Review of various modeling techniques for the detection of electricity theft in smart grid environment, Renewable and Sustainable Energy Reviews 82 (2018) 2916–2933. doi:10.1016/j.rser.2017.10.040.
URL https://doi.org/10.1016/j.rser.2017.10.040

[11] L. T. Faria, J. D. Melo, A. Padilha-Feltrin, Spatial-Temporal Estimation for Nontechnical Losses, IEEE Transactions on Power Delivery 31 (1) (2016) 362–369. doi:10.1109/TPWRD.2015.2469135.

[12] K. Shahzad, S. U. Bajwa, R. B. Ansted, D. Mamoon, K. ur Rehman, Evaluating human resource management capacity for effective implementation of advanced metering infrastructure by electricity distribution companies in pakistan, Utilities Policy 41 (2016) 107–117. doi:https://doi.org/10.1016/j.jup.2016.06.011.
URL https://www.sciencedirect.com/science/article/pii/S0957178716301606

[13] J. L. Rodrigues, I. Morro-Mello, J. D. Melo, A. Padilha-Feltrin, Estimation of electric demand from electric vehicles using spatial regressions, in: Proc. 2019 IEEE PES Innovative Smart Grid Technologies Conf. - Latin America (ISGT Latin America), IEEE, 2019, pp. 1–6. doi:10.1109/ISGT-LA.2019.8895367.

[14] C. Foroni, F. Ravazzolo, L. Rossini, Are low frequency macroeconomic variables important for high frequency electricity prices?, Economic Modelling 120 (2023) 106160. doi:https://doi.org/10.1016/j.econmod.2022.106160.
URL https://www.sciencedirect.com/science/article/pii/S0264999322003972

[15] X. Xu, S. K. Sinha, Robust designs for generalized linear mixed models with possible model misspecification, Journal of Statistical Planning and Inference 210 (2021) 20–41. doi:10.1016/j.jspi.2020.04.006.
URL https://doi.org/10.1016/j.jspi.2020.04.006

[16] G. Verbeke, G. Molenberghs, Linear Mixed Models for Longitudinal Data, 1st Edition, Springer Series in Statistics, Springer, New York, 2000.

[17] L. C. Hui, C. Jim, Urban-greenery demands are affected by perceptions of ecosystem services and disservices, and socio-demographic and environmental-cultural factors, Land Use Policy 120 (2022) 106254. doi:https://doi.org/10.1016/j.landusepol.2022.106254.
URL https://www.sciencedirect.com/science/article/pii/S0264837722002812

[18] R. Camboni, A. Corsini, R. Miniaci, P. Valbonesi, Mapping fuel poverty risk at the municipal level. a small-scale analysis of italian energy performance certificate, census and survey data, Energy Policy 155 (2021) 112324. doi:https://doi.org/10.1016/j.enpol.2021.112324.

URL https://www.sciencedirect.com/science/article/pii/S0301421521001932

[19] L. O. Ventura, J. D. Melo, A. Padilha-Feltrin, J. P. Fernández-Gutiérrez, C. C. S. Zuleta, C. C. P. Escobar, A new way for comparing solutions to non-technical electricity losses in South America, Utilities Policy 67 (2020) 101113. doi:10.1016/j.jup.2020.101113.
URL https://linkinghub.elsevier.com/retrieve/pii/S0957178720301077

[20] G. M. Messinis, N. D. Hatziargyriou, Review of non-technical loss detection methods, Electric Power Systems Research 158 (2018) 250–266. doi:10.1016/j.epsr.2018.01.005.
URL http://dx.doi.org/10.1016/j.epsr.2018.01.005

[21] P. Massaferro, J. M. D. Martino, A. Fernandez, Fraud detection in electric power distribution: An approach that maximizes the economic return, IEEE Transactions on Power Systems 35 (1) (2020) 703–710. doi:10.1109/TPWRS.2019.2928276.

[22] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, A. Gómez-Expósito, Hybrid deep neural networks for detection of non-technical losses in electricity smart meters, IEEE Transactions on Power Systems 35 (2) (2020) 1254–1263. doi:10.1109/TPWRS.2019.2943115.

[23] J. I. Guerrero, I. Monedero, F. Biscarri, J. Biscarri, R. Millan, C. Leon, Non-technical losses reduction by improving the inspections accuracy in a power utility, IEEE Transactions on Power Systems 33 (2) (2018) 1209–1218. doi:10.1109/TPWRS.2017.2721435.

[24] K. M. Ghori, R. A. Abbasi, M. Awais, M. Imran, A. Ullah, L. Szathmary, Performance analysis of different types of machine learning classifiers for non-technical loss detection, IEEE Access 8 (2020) 16033–16048. doi:10.1109/ACCESS.2019.2962510.

[25] H. Long, C. Chen, W. Gu, J. Xie, Z. Wang, G. Li, A data-driven combined algorithm for abnormal power loss detection in the distribution network, IEEE Access 8 (2020) 24675–24686. doi:10.1109/ACCESS.2020.2970548.

[26] J. B. Leite, J. R. S. Mantovani, Detecting and locating non-technical losses in modern distribution networks, IEEE Transactions on Smart Grid 9 (2) (2018) 1023–1032. doi:10.1109/TSG.2016.2574714.

[27] F. Heymann, F. Vom Scheidt, F. J. Soares, P. Duenas, V. Miranda, Forecasting energy technology diffusion in space and time: model design, parameter choice and calibration, IEEE Transactions on Sustainable Energy 12 (2) (2021) 802–809. doi:10.1109/TSTE.2020.3020426.

[28] R. S. Bivand, E. Pebesma, V. Gómez-Rubio, Applied spatial data analysis with R, Use R, Springer, New York, 2008.

[29] W. R. Tobler, A computer movie simulating urban growth in the Detroit region, Economic Geography 46 (1970) 234–240. doi:10.2307/143141.

URL https://www.tandfonline.com/doi/abs/10.2307/143141

[30] Y. R. Gahrooei, A. Khodabakhshian, R.-A. Hooshmand, A new pseudo load profile determination approach in low voltage distribution networks, IEEE Transactions on Power Systems 33 (1) (2018) 463–472. doi:10.1109/TPWRS.2017.2696050.
URL http://ieeexplore.ieee.org/document/7908996/

[31] S. Wang, R. Ferrus, Extracting cell patterns from high-dimensional radio network performance datasets using self-organizing maps and k-means clustering, IEEE Access 9 (2021) 42045–42058. doi:10.1109/ACCESS.2021.3065820.

[32] S. Haykin, Neural Networks: A Comprehensive Foundation, International edition, Prentice-Hall, Upper Saddle River, 1999.

[33] Y. Hao, Y. M. Liu, The influential factors of urban PM2.5 concentrations in China: A spatial econometric analysis, Journal of Cleaner Production 112 (2016) 1443–1453. doi:10.1016/j.jclepro.2015.05.005.

[34] Q. Zhang, J. Yang, Z. Sun, F. Wu, Analyzing the impact factors of energy-related CO2 emissions in China: What can spatial panel regressions tell us?, Journal of Cleaner Production 161 (2017) 1085–1093. doi:10.1016/j.jclepro.2017.05.071.
URL http://dx.doi.org/10.1016/j.jclepro.2017.05.071

[35] G. Dong, R. Harris, Spatial Autoregressive Models for Geographically Hierarchical Data Structures, Geographical Analysis 47 (2) (2015) 173–191. doi:10.1111/gean.12049.

[36] H. L. Willis, Spatial electric load forecasting, Power Engineering (Willis), CRC Press, 2002.

[37] J. L. Rodrigues, H. M. Bolognesi, J. D. Melo, F. Heymann, F. Soares, Spatiotemporal model for estimating electric vehicles adopters, Energy 183 (2019) 788–802. doi:https://doi.org/10.1016/j.energy.2019.06.117.
URL https://www.sciencedirect.com/science/article/pii/S0360544219312496

[38] G. Dong, L. Wolf, A. Alexiou, D. Arribas-Bel, Inferring neighbourhood quality with property transaction records by using a locally adaptive spatial multi-level model, Computers, Environment and Urban Systems 73 (2019) 118–125. doi:https://doi.org/10.1016/j.compenvurbsys.2018.09.003.
URL https://www.sciencedirect.com/science/article/pii/S0198971518301042

[39] Z. Chen, S. Zhu, Q. Niu, T. Zuo, Knowledge discovery and recommendation with linear mixed model, IEEE Access 8 (2020) 38304–38317. doi:10.1109/ACCESS.2020.2973170.

[40] D. Hong, N. Yokoya, J. Chanussot, X. X. Zhu, An augmented linear mixing model to address spectral variability for hyperspectral unmixing, IEEE Transactions on Image Processing 28 (4) (2019) 1923–1938. arXiv:1810.12000, doi:10.1109/TIP.2018.2878958.

[41] Y. Wang, K. Zhang, C. Tang, Q. Cao, Y. Tian, Y. Zhu, W. Cao, X. Liu, Estimation of rice growth parameters based on linear mixed-effect model using multispectral images from fixed-wing unmanned aerial vehicles, Remote Sensing 11 (11) (2019). doi:10.3390/rs11111371.
URL https://www.mdpi.com/2072-4292/11/11/1371

[42] Z. Wang, H. Wang, S. Wang, S. Lu, G. Saporta, Linear mixed-effects model for longitudinal complex data with diversified characteristics, Journal of Management Science and Engineering 5 (2) (2020) 105–124. doi:https://doi.org/10.1016/j.jmse.2019.11.001.
URL https://www.sciencedirect.com/science/article/pii/S2096232019300897

[43] L. Meteyard, R. A. Davies, Best practice guidance for linear mixed-effects models in psychological science, Journal of Memory and Language 112 (2020) 104092. doi:https://doi.org/10.1016/j.jml.2020.104092.
URL https://www.sciencedirect.com/science/article/pii/S0749596X20300061

[44] Y.-c. Yang, T.-i. Lin, L. M. Castro, W.-l. Wang, Extending finite mixtures of t linear mixed-effects models with concomitant covariates, Computational Statistics and Data Analysis 148 (2020) 106961. doi:10.1016/j.csda.2020.106961.
URL https://doi.org/10.1016/j.csda.2020.106961

[45] M. H. DeGroot, M. J. Schervish, Probability and statistics, Pearson Education, 2012.

[46] M. Pljakić, D. Jovanović, B. Matović, S. Mićić, Macro-level accident modeling in Novi Sad: A spatial regression approach, Accident Analysis and Prevention 132 (2019) 105259. doi:10.1016/j.aap.2019.105259.

[47] R Core Team, A language and environment for statistical computing, Vienna, 2015.
URL http://www.r-project.org

[48] Minitab.
URL https://www.minitab.com/en-us/

[49] Brazilian Institute of Geography and Statistics (IBGE), Household census by subareas, 2010.
URL http://www.ibge.gov.br

[50] G. Dong, R. Harris, A. Mimis, HSAR: An R package for integrated spatial econometric and multilevel modelling, GIS Research UK 2016 held at the Faculty of Architecture, Computing and Humanities, 2016.

[51] I. Svetunkov, Greybox: Toolbox for model building and forecasting. R package version 0.5.9, 2021.
URL https://cran.r-project.org/package=greybox

[52] Y. Song, D. Thatcher, Q. Li, T. McHugh, P. Wu, Developing sustainable road infrastructure performance indicators using a model-driven fuzzy spatial multi-criteria decision making method, Renewable and Sustainable Energy Reviews 138 (2021) 110538. doi:10.1016/j.rser.2020.110538.

URL https://doi.org/10.1016/j.rser.2020.110538