IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 February 2024, accepted 17 March 2024, date of publication 28 March 2024, date of current version 9 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3382943

RESEARCH ARTICLE

Estimation of Vulnerable Areas to Faults Caused by Tree Vegetation in Power Distribution Systems

ANDRÉIA S. SANTOS[1], LUCAS TELES FARIA[2], LETÍCIA S. BOSCHI[2], MARA LÚCIA M. LOPES[1], AND CARLOS R. MINUSSI[1]

[1]Department of Electrical Engineering, São Paulo State University (UNESP), Ilha Solteira, São Paulo 15385000, Brazil
[2]Department of Energy Engineering, São Paulo State University (UNESP), Rosana, São Paulo 19274000, Brazil

Corresponding author: Carlos R. Minussi (carlos.minussi@unesp.br)

This work was supported in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) under Grant 001.

ABSTRACT Faults in power distribution feeders cause damage to power utilities due to the deterioration of reliability and power quality indexes and the displacement of field maintenance teams to replace or repair power grid equipment. Additionally, consumer units have energy supply interruptions for an undetermined time. Studies in specialized literature usually detect, classify, and locate faults after they occur. In contrast, preventing faults by estimating areas vulnerable to them is crucial to mitigate all inconveniences and additional costs after they occur. Tree vegetation is an essential factor contributing to faults. In this sense, an enhanced method for tree vegetation mapping by areas is developed using multilayer perceptron neural networks trained on high-resolution images from Google Earth. A geographic space is incorporated to estimate the regions vulnerable to failures due to tree vegetation. Geographically weighted spatial analysis is applied from local variables aggregated by areas. Spatial data analysis is used to real faults and tree vegetation data from a medium-sized Brazilian city via QGIS and R programming environments. As a result, thematic maps are produced with the areas whose feeders are vulnerable to faults, where there is a moderate positive correlation by regions between the faults in distribution transformers and tree vegetation in the northeast and southwest areas of the city under study.

INDEX TERMS Exploratory spatial data analysis, power distribution systems, spatial data analysis, steady-state fault.

NOTATION

The notation used throughout this paper is reproduced below for quick reference.

Acronyms:

ANN	Artificial Neural Network;
CTs	Census Tracts;
CUs	Consumer Units;
ESDA	Exploratory Spatial Data Analysis;
ESI	Energy Supply Interruptions;
ETT	Enhanced Tree Trimming
GIS	Geographic Information Systems;
GW	Geographically Weighted;
GWEA	Geographically Weighted Exploratory Analysis;
GWMs	Geographically Weighted Models;
MLP	Multilayer Perceptron;
PDS	Power Distribution Systems;
SDA	Spatial Data Analysis;
SWM	Spatial Weighting Matrix.

Indices:

i,j	Index for census tracts.

Parameters:

$\bar{X}$	Arithmetic mean;
σ	Standard deviation;
b	Bandwidth parameter for the Gaussian kernel function;

The associate editor coordinating the review of this manuscript and approving it for publication was F. R. Islam.

d	Difference between the ordinal position of two variables in the same census tract;
d_{ij}	Distance between the centroids of census tracts i and j;
n	Number of census tracts;
$\boldsymbol{W}_{(n\times n)}$	Spatial weighting matrix;
w_{ij}	Weighting between the census tracts i and j;
y_i, z_i	Attributes associated with a census tract at i.
Set:	
Ω_k	Set of k classes used in the legend of thematic maps.
Statistical Metrics:	
$\rho(z_i, y_i)$	Geographically weighted Pearson's local correlation coefficient;
$c(z_i, y_i)$	Geographically weighted covariance;
$m(z_i)$	Geographically weighted mean;
r_s	Spearman's correlation coefficient;
$s\,(z_i)$	Geographically weighted standard deviation.
Variables:	
NFT	Number of faults in distribution transformers by census tracts;
PFT	Percentage of faults in distribution transformers by census tracts;
PTV	Percentage of tree vegetation by census tracts.

I. INTRODUCTION

Power distribution systems (PDS) are continually subject to events that can result in disturbances and, consequently, failures [1]. These failures cause numerous interruptions in the energy supply and negatively impact PDS's reliability and power quality, seriously threatening its operation [2], [3].

ESIs are associated with multiple factors such as (i) tree vegetation close to the overhead utility grid, (ii) adverse weather conditions (wind gusts, snow, storms, electrical discharges), (iii) damaged equipment (missing maintenance, obsolescence, or manufacturing defect), (iv) animals as birds and insects, (v) vehicle collision, (vi) overload on utility grid; (vii) clandestine connection or vandalism, (viii) human failures, (ix) protective equipment acting, and (x) kites [4].

Tree vegetation represents a relevant cause associated with faults among all the factors above, and its impact is a recurring problem for power utilities that results in significant economic losses [5], [6]. This impact is related to the uncontrolled growth of tree vegetation close to distribution lines, plus adverse weather conditions such as wind gusts and storms [7], [8]. Additionally, the contact of tree branches with energized conductors belonging to the utility grid often causes short circuits, failures, damage to system infrastructure, and risks to human health [9], [10].

Tree vegetation close to overhead distribution lines also can interfere with the operation of the utility grid in each area, making it more vulnerable to failures. The overlapping of tree branches over the cables can cause interference, ESIs, and compromise the safety of consumer units (CUs).

Faults resulting from tree vegetation can be classified into two categories: the first refers to interruptions caused by tree branches that meet utility grid cables as they grow close to the feeder. The second category is directly related to adverse weather conditions, which can result in trees falling onto distribution lines or forced contact of branches with the feeder.

Adopting preventive measures is crucial to mitigate the faults, inconveniences, and additional costs after they occur in PDS. To the best of the authors' knowledge, specialized literature studies aim to detect, classify, and locate faults after they occur – post-fault studies [11], [12]. On the other hand, estimating regions whose distribution feeders are vulnerable to faults can be a valuable tool to power utilities in planning and executing preventive actions to minimize the occurrence of faults and all maintenance costs related to them.

As mentioned above, several factors can make some regions or cities more vulnerable to ESIs; however, some may be more relevant than others. For example, wind gusts may be appropriate in coastal towns; conversely, tree vegetation can be applicable in heavily wooded cities. Thus, the variables associated with faults depend on the region or city under study. In this context, spatial data analysis (SDA) is a tool for preventing faults by estimating the city's regions whose feeders are vulnerable to them. Furthermore, the leading causes of faults can be identified in the city under study.

This work aims to incorporate the geographic space study to estimate areas whose utility grid is vulnerable to failures due to tree vegetation. The variables evaluated in this study are shutdowns in distribution transformers that caused steady-state faults and the percentage of area occupied by tree vegetation per census tracts (CTs), which are directly associated with faults in utility grid feeders. It is observed that all transformers and tree vegetation are georeferenced in a time window; that is, their geographic coordinates are known, enabling SDA application.

A. LITERATURE REVIEW

Faults in PDS related to tree vegetation are addressed in recent works. Cerrai et al. [5] assessed the impact of enhanced tree trimming (ETT) on ESI, in which two different methodologies were implemented. First, the authors applied a statistical analysis to verify the trend of reduction in interruptions as the number of ETTs increases. An interruption prediction model was addressed in the second stage, where the aim is to evaluate changes in the number of interruptions before and after ETT.

An approach for predicting vegetation-related interruptions was presented in [7]. They were categorized into two groups: those caused by tree growth and those caused by vegetation due to adverse weather conditions.

The predictions applied two models: a statistical approach based on time series algorithms and an approach based on non-linear machine learning.

In [11], the interruptions' frequency, extent, and duration were considered. The analysis of tree pruning operations was evaluated based on a set of actual data, and two autoregressive models were introduced: the spatial autoregressive model and the spatial Durbin model. The results demonstrated that tree pruning operations reduced interruptions and the number of affected CUs.

Chen and Kezunovic [13] implemented a predictive method for managing interruptions in transmission and distribution systems, where the impact due to the combination of winds and tree vegetation was considered. A geographic information systems (GIS) framework was introduced to correlate energy system data with multiple climate data layers, including wind and vegetation information. As a result, maps were presented with the areas vulnerable to interruptions.

Local variables associated with faults due to trees during storms were analyzed in [14]. Five fault risk models due to tree vegetation were developed and compared using a random forest algorithm, where the physical structure of trees plus local environmental factors such as utility infrastructure, faults history, tree pruning data, soil, and land cover information were considered.

Taylor et al. [15] evaluated the impact of vegetation management to reduce ESIs during storms. A machine learning model was applied, where data on vegetation management, soil cover, climate data, and electrical infrastructure were inserted. The study quantified the damage reduction in the power grid due to ETT during adverse weather events. It was found that annual reductions in hotspots varied between 25.7% and 42.5% due to ETT.

The authors empirically demonstrate the impact of ETT on reducing faults during storms [16]. ESI reduced significantly in areas where ETT treatment was applied, while more failures occurred in untreated lines.

The aforementioned studies evaluate the impact of ETT on reducing ESI. On the other hand, the relevance of our research consists of assessing the relationship between tree vegetation and ESI, typically caused by factors such as wind, snow/ice accumulation, and falling branches, considering the city's geographic space. Given that power utilities implement tree vegetation management programs such as ETT, this study becomes pertinent when there is a need to enhance the understanding or optimize strategies within these programs. For example, this study can help in the decision-making process to select critical areas that can be the targets of ETT programs.

In [17], a spatial fuzzy influence diagram was presented to identify vulnerable points in the power grid with a high risk of interruptions caused by trees during extreme weather conditions. Data such as forest resources, soil type, power grid, and history of interruptions were considered. As a result, the authors presented a map showing the areas whose power grid is vulnerable to interruptions.

Finally, Onaolapo et al. [18] presented an outage prediction model based on an artificial neural network (ANN). Two models were built, one with ANN and the other with multiple linear regression, where climate variables influencing the power grid were introduced. The ANN-based model improved performance by incorporating ten variables: (1) outage history, (2) number of clouds, (3) minimum and maximum temperatures, and (4) number of frost days. The study highlighted the complex relationship between weather events and failures in the power grid that result in ESI. On the other hand, the city's areas most vulnerable to climate events were not highlighted. For example, strong winds can cause trees and branches to fall onto the power grid. Therefore, the city's heavily wooded regions will be more vulnerable to faults.

B. CONTRIBUTIONS

This study performs exploratory spatial data analysis (ESDA) from tree vegetation data associated with faults and shutdowns in distribution transformers. Its main contributions are described below.

- Geographical space study is incorporated to produce thematic maps showing areas vulnerable to faults due to tree vegetation. These maps are an easy-to-interpret visual tool for power utilities to plan and execute actions to prevent ESIs. Therefore, all inconveniences and additional costs are minimized after the fault occurs in PDS.
- Geographically weighted exploratory analysis (GWEA) by regions uses variables associated with each CT: distribution transformer faults and tree vegetation. Local geographically weighted (GW) measurements, such as standard deviation and local correlations, are applied. GWEA considers the variable values for each CT and the influence of closest or neighboring areas.

C. PAPER STRUCTURE

This paper follows in Section II with SDA, spatial analysis with data aggregated by areas (Subsection II-A), Spearman's correlation coefficient (Subsection II-A.1), spatial weighting matrix (Subsection II-A.2), and GWEA – Subsection II-A.3. Subsection II-B presents tree vegetation mapping procedure. Section III presents the implementation of GWEA in a Brazilian city. Subsections III-A and III-B present tree vegetation mapping and database description, respectively. An ESDA is performed in Subsection III-C, from the percentage of tree vegetation (Subsection III-C.1) and faults in distribution transformers – Subsection III-C.2. GW statistical summary is performed in Subsection III-C.3 considering the weighting among areas. Finally, Section IV shows the conclusions of our work.

II. SPATIAL DATA ANALYSIS

SDA aims to measure the properties and relationships of events considering their spatial location – geographical coordinates. In this way, geographic space is incorporated into the

analysis; therefore, there is a visual perception of the spatial distribution of the event under study [19].

This study uses SDA metrics to perform an ESDA on data aggregated by areas named CTs. In this way, each location is associated with a value for each variable under study: faults percentage in distribution transformers and percentage of area covered by tree vegetation.

ESDA is the crucial step of a study that applies SDA. It consists of metrics to identify patterns based on the spatial distribution of the event under study. Thus, hypotheses can be formulated about the variables influencing its spatial distribution [20], [21].

ESDA application shows essential effects: spatial autocorrelation and spatial heterogeneity [22]. Spatial autocorrelation refers to the analysis of variation in spatial dependence based on comparing observations from a given geographic area and its neighbors. This concept was defined by Tobler, who stated the first Law of Geography: all things are similar; however, closer things look more than farther away [23]. Autocorrelation has negative or positive values, where positive values indicate that close observations tend to exhibit similar attributes; therefore, there is aggregation, similarity, or grouping of the event under analysis. On the other hand, spatial heterogeneity shows the variation of observations in geographic space [19], [22].

A. SPATIAL ANALYSIS WITH DATA AGGREGATED BY AREAS

This study is linked to an SDA branch with several methods for analyzing data aggregated by areas delimited by polygons. They are applied to studying events aggregated by areas where their location is unavailable. There is only one value for each variable associated with each region. Thus, it is possible to analyze patterns, trends, and relationships in data sets with specific information about geographic areas such as CTs, cities, or countries [19], [24].

The presentation of data aggregated by areas is prepared using colored maps (thematic maps) that display the spatial pattern of the event under analysis [19].

A branch of SDA with data aggregated by area allows the use of public information from the demographic census. They are grouped into small areas named CTs, whose area is a function of their population density; CTs with greater-density populations have smaller areas, and vice versa [4].

1) CORRELATION COEFFICIENT OF SPEARMAN

The Spearman's coefficient correlation is a metric that identifies whether two variables are associated, where its value belongs to the continuous interval $[-1, 1]$.

Negative correlation values indicate that the variables are inversely proportional. Positive values indicate that the variables are directly proportional. Correlation strength is evaluated as this metric approaches the two extremes $\{-1, 1\}$. On the other hand, positive or negative correlations close to zero indicate no significant correlation. Consider two hypothetical variables, A and B: if A is associated with B, it does not necessarily mean that A causes B and vice versa.

The r_s coefficient is obtained by ranking the values of both variables in ascending (or descending) order, where original values are replaced by a positive integer value representing the variable's ordinal position in the data series. The r_s coefficient is calculated by:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \tag{1}$$

where n is the total number of CTs and d is the difference between the ordinal position of two variables in the same CT.

Spearman's coefficient is not sensitive to asymmetries or outliers in the data series. Furthermore, the pair of variables do not need to be linearly associated [25].

Finally, Spearman's coefficient is a global metric that does not consider the variables' space distribution. In contrast, the following two subsections present a spatial weighting matrix (SWM) representing the influence of closest or neighboring areas for obtaining GW metrics values.

2) SPATIAL WEIGHTING MATRIX

The neighborhood structure among areas establishes their influence relationship, and it is crucial in studies that apply SDA techniques with data aggregated by areas.

SDA techniques are widely applied in epidemiology, botany, soil studies, mineral resources prospecting, and criminology. A neighborhood structure among areas from Euclidean distance among centroid areas is usually applied in these knowledge areas, where closer areas have more significant influence than distant areas [23]. This same neighborhood structure is used in this study because an abrupt transition of tree vegetation or faults among neighboring areas is not expected.

In this sense, SWM is applied to estimate the spatial variability in data aggregated by areas. SWM $\boldsymbol{W}_{(n\times n)}$ is constructed from a discrete set of n areas $\{A_1, ..,A_n\}$, where each element w_{ij} represents the spatial dependence among the variables observed in A_i and A_j areas.

Fig. 1 shows an example of obtaining SWM. There are six CTs (A, B, C, D, E, and F) of the city in Fig. 1 (a); on the other hand, in Fig. 1 (b), there is the corresponding SWM $\boldsymbol{W}_{(6\times 6)}$, since $n = 6$ CTs and whose obtaining rule is according to (2).

For example, CT B has borders in common with two CTs: A and E. Thus, according to (2), SWM elements w_{21} (weighting between B and A) and w_{25} (weighting between B and E) are equal to $1/2$. The other areas do not have common borders with B; therefore, they have zero weighting – $w_{23} = 0$, $w_{24} = 0$, and $w_{26} = 0$.

$$w_{ij} = \begin{cases} \dfrac{1}{No.\ of\ Borders\ CTs}, & if\ i\ and\ j\ are\ borders\ CTs \\ 0, & otherwise \end{cases} \tag{2}$$

SWM elements can also be obtained via kernel functions. They are decreasing and monotone functions of the distance

(a)

	A	B	C	D	E	F
A	1	0.2	0.2	0.2	0.2	0.2
B	0.5	1	0	0	0.5	0
C	0.5	0	1	0.5	0	0
D	0.33	0	0.33	1	0	0.33
E	0.33	0.33	0	0	1	0.33
F	0.33	0	0	0.33	0.33	1

(b)

FIGURE 1. An illustrative example of obtaining a SWM with six CTs of the city (a) and its corresponding SWM (b).

between A_i and A_j centroids. Therefore, it follows Tobler's first law, where the weighting among nearby CTs is more significant than for distant CTs. Gaussian kernel is applied in our study to obtain SWM elements according to (3):

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right) \tag{3}$$

where d_{ij} represents the distance between the centroids of CTs i, and j and b is a bandwidth parameter for the function's decay rate [26].

3) GEOGRAPHICALLY WEIGHTED EXPLORATORY

Geographically weighted models (GWMs) are tools belonging to a particular branch of non-stationary spatial statistics, which intuitively incorporate local spatial relationships into their structure. These techniques are helpful in contexts where a global model does not adequately describe spatial data, as they enable the estimation of local parameters in geographic space [26], [27].

The GW model outputs are mapped to provide a helpful tool that typically precedes more sophisticated statistical analysis.

The kernel function is crucial in GW modeling, which aims to quantify the spatial dependence relationship among the variables observed in CTs.

GW local summary statistic is obtained from a set of spatial data and a SWM. Consider the attributes z_i e y_i associated with a CT at i. The following metrics can be obtained via (4)–(7), respectively: GW mean, GW standard deviation, GW Pearson's local correlation coefficient, and GW covariance [26].

$$m(z_i) = \frac{\sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} w_{ij}} \tag{4}$$

$$s(z_i) = \sqrt{\frac{\sum_{j=1}^{n} w_{ij}(z_j - m(z_i))^2}{\sum_{j=1}^{n} w_{ij}}} \tag{5}$$

$$\rho(z_i, y_i) = \frac{c(z_i, y_i)}{s(z_i)\, s(y_i)} \tag{6}$$

$$c(z_i, y_i) = \frac{\sum_{j=1}^{n} w_{ij} \left[(z_j - m(z_i))(y_j - m(y_i))\right]}{\sum_{j=1}^{n} w_{ij}} \tag{7}$$

where $\boldsymbol{w_{ij}}$ are SWM elements.

B. METHODOLOGY FOR TREE VEGETATION MAPPING

Tree vegetation mapping is performed via image classification of each CT by pixel identification. This process involves identifying the RGB pattern of each pixel, allowing precise identification of the area percentage occupied by tree vegetation.

Some remote sensing techniques are limited and cannot adequately identify tree vegetation due to satellite images' pixel length of around 30 m [28]. Therefore, an ANN multilayer perceptron (MLP) is applied to classify tree vegetation from the pixels' RGB pattern [29].

Fig. 2 shows an illustrative example of a CT with highlighted tree vegetation and the image pixels in a zoom version. In this sense, Fig. 3 shows a flowchart with the steps to determine the area percentage of each CT occupied by tree vegetation: (i) obtaining high-resolution images from Google Earth of each CT; (ii) extraction of the RGB pattern of each pixel; (3) Application of MLP to check the compatibility of the pixel's RGB pattern with tree vegetation; (4) Determination of CT area occupied by tree vegetation.

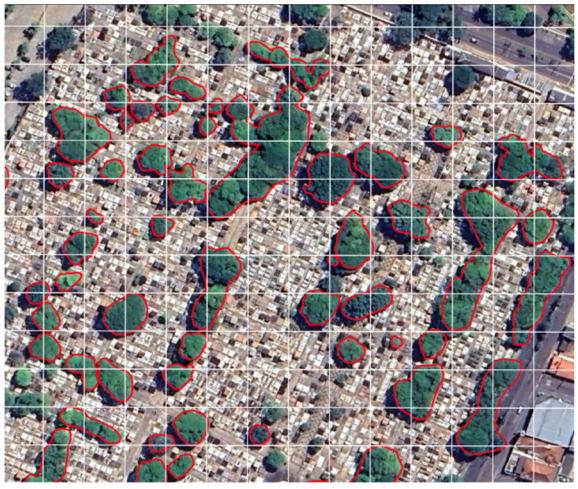


FIGURE 2. Illustrative example of tree vegetation mapping.

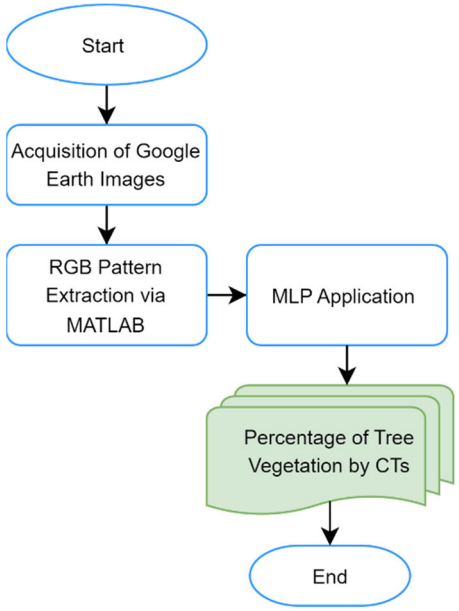


FIGURE 3. Flowchart of tree vegetation mapping.

III. EXPLORATORY SPATIAL DATA ANALYSIS IN A BRAZILIAN CITY

This work applies an ESDA to evaluate the influence of tree vegetation on making some city CTs vulnerable to steady-state faults.

In this sense, Fig. 4 shows some relevant factors that caused 9,266 ESI with more than a 1-minute duration over four years – 2009 to 2012. Tree vegetation occupies the second position in the ranking as the cause of 19% of ESI in the city under study. These data highlight the importance of developing effective strategies for adequately managing tree vegetation to mitigate their impact on interruptions, improving the PDS reliability and the service quality provided to CUs.

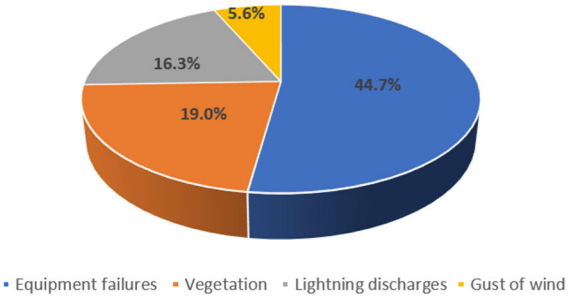


FIGURE 4. Relevant factors that caused steady faults in the city under analysis.

The power grid analyzed has feeders located in a Brazilian city. All simulations are performed via QGIS and R software version 3.30 and 4.1.2, respectively.

QGIS is a GIS software that enables visualization, editing, and SDA [30]. On the other hand, R is a free software for statistical and graphical computing, with several packages, including the GWmodel package applied in this study. Although not a GIS, R can perform similar functions [26], [31].

All simulations in our work are performed on a computer with a Windows operating system, an AMD Ryzen 7 3700X processor, 3.6 GHz, 64-bit, and 16 GB of RAM.

A. TREE VEGETATION MAPPING

An MLP is applied to identify the pixel's RGB pattern with tree vegetation from high-resolution images obtained from Google Earth. Image processing via MATLAB software is used to achieve this objective [32].

MLP has a topology with three inputs (RGB pattern), 20 neurons (intermediate layer), and one neuron in the binary output layer, where the RGB pattern of some pixels is compatible with tree vegetation. MLP training uses 815 samples with different shades, some compatible with tree vegetation and others non-compatible. Therefore, there is a clear distinction between tree vegetation and undergrowth.

Table 1 shows the MLP parameters applied to tree vegetation mapping for all city CTs [29].

TABLE 1. Multilayer perceptron parameters for tree vegetation mapping.

Parameter	Value
Training	Supervised
Architecture	Feedforward
Training Algorithm	Backpropagation
Topology	3–20–1
Constant Training	10^{-1}
Training Accuracy	10^{-6}
Training Set	815 samples (~70%)
Validation Set	349 samples (~30%)
Learning Type	Off-line

B. DATABASE DESCRIPTION

This section evaluates the percentage of tree vegetation (PTV) by CTs, which is responsible for permanent faults, and if it can make some city areas vulnerable to faults.

In this sense, the terminology used in spatial regression is applied where a set of independent variables explains a dependent variable's distribution in geographic space. In this work, the dependent or study variable evaluated is shutdowns in distribution transformers by CTs that caused steady-state faults. The faults are associated with distribution transformers because they are georeferenced, enabling the SDA application. An independent or explanatory variable evaluated, directly related to faults in utility grid feeders, is the percentage of area occupied by tree vegetation by CTs.

Consider a CT_i with $i = 1, \ldots, n$ where n is the number of city's CTs. The independent variable PTV_i is shown in (8). It is obtained from the ratio between the area occupied by tree

vegetation in CT_i and its area:

$$PTV_i = \frac{Tree\ Vegetation\ Area\ in\ CT_i}{CT_i\ Area}\ with\ i = 1, \ldots, n \tag{8}$$

The dependent variable percentage of faults in distribution transformers at CT_i named PFT_i is shown in (9). It is obtained from the ratio between the number of permanent faults at CT_i NFT_i that caused shutdowns in distribution transformers and the number of distribution transformers at CT_i named T_i.

$$PFT_i = \frac{NFT_i}{T_i} \text{with } i = 1, \ldots, n \tag{9}$$

PFT_i is a fault probability estimating in distribution transformers by CTs. CTs with more transformers T_i tend to have a more significant number of faults in transformers NFT_i. Therefore, PFT_i is a more effective variable than just the number of faults NFT_i because it represents a vulnerability measure of CTs to faults.

All variables applied in ESDA are described in Table 2.

TABLE 2. Variables description.

Description	Acronym	Type Variable
Number of Faults in Distribution Transformers by CTs	*NFT*	Dependent Variables
Percentage of Faults in Distribution Transformers by CTs	*PFT*	
Percentage of Tree Vegetation by CTs	*PTV*	Independent Variable

C. EXPLORATORY SPATIAL DATA ANALYSIS

An ESDA is performed in *NFT*, *PTV*, and *PFT* variables in the following two subsections, where the focus is on visualizing the variable distribution in the geographic space of the city via thematic maps.

Thematic map legends are performed via standard deviation intervals, which consider the statistical distribution of the variable to be mapped. This technique is more appropriate for use in data series that follow normal distribution. However, it can show the locational distortions on maps of some CTs concerning the global average [25], [33].

In this sense, the mean $\bar{X}$ and standard deviation σ of the distribution are calculated; then, the data series is segmented into class intervals whose limits are proportions of the variable's standard deviation. The set Ω_k of k classes can be generalized in (10). Alternatively, 0.5σ or 0.25σ can be used instead of σ for data series with small sparsity.

$$\Omega_k = \{\ldots, (\bar{X} - 2\sigma), (\bar{X} - \sigma), \bar{X}, (\bar{X} + \sigma), (\bar{X} + 2\sigma), \ldots\} \tag{10}$$

1) PERCENTAGE OF TREE VEGETATION

An ESDA is performed for the independent variable *PTV* in this section. Figs. 5 (a)–(c) show the *PTV* distribution by CTs from 2010 to 2012. The maps are very similar because tree vegetation usually changes very slowly. The CTs with the highest *PTV* are found on the city outskirts, mainly in the east, northeast, and southeast, and some CTs in the central and extreme west. As expected, the city's central region (red circle) has a low *PTV*.

Table 3 shows descriptive statistics metrics for *PTV* by CTs from 2010 to 2012. The metrics values are very similar for all years evaluated.

TABLE 3. Statistical summary for PTV by census tracts.

Parameter	Year Evaluated		
	2010	2011	2012
Max.	64.43	70.00	71.00
Min.	0.00	0.00	0.00
Mean	8.02	8.28	8.45
Standard Deviation	6.05	6.23	6.55

The numbers in parentheses in the Figs. 5 legend indicates the number of CTs whose *PTV* falls within the range. Figs. 5 (a)–(c) most CTs have $PTV \in [5\%, 8\%[$ for all years.

2) FAULTS IN DISTRIBUTION TRANSFORMERS

This section performs an ESDA for the dependent variables *NFT* and *PFT*. This study analyzed 2,848 interruptions caused by the failure of distribution transformers.

Figs. 6 (a)–(c) show the *NFT* distribution by CTs from 2010 to 2012. Fig. 6 (a) refers to *NFT* in 2010; 135 CTs have $NFT \in]0, 3]$. Most CTs have *NFT* in this range for all years.

The CTs with the highest *NFT* were found mainly in the south region in 2010 and the south plus northeast regions in 2011. *NFT* fell drastically in 2012. Power utilities periodically organize task forces to replace damaged equipment (such as transformers) in critical CTs. It cannot be conclusively stated that *NFT* fell in 2012 due to this, but it is a relevant hypothesis to consider. The city's central region has low *NFT* for all years.

Table 4 shows some descriptive statistics metrics for *NFT* by CTs from 2010 to 2012. On average, each CT has between two and four annual faults. However, in 2011, a single CT was the target of 27 shutdowns in distribution transformers.

TABLE 4. Statistical summary for NFT by census tracts.

Parameter	Year Evaluated		
	2010	2011	2012
Max.	21	27	13
Min.	0	0	0
Mean	3.55	3.83	2.08
Standard Deviation	3.63	4.28	2.35
Total	1069	1153	626

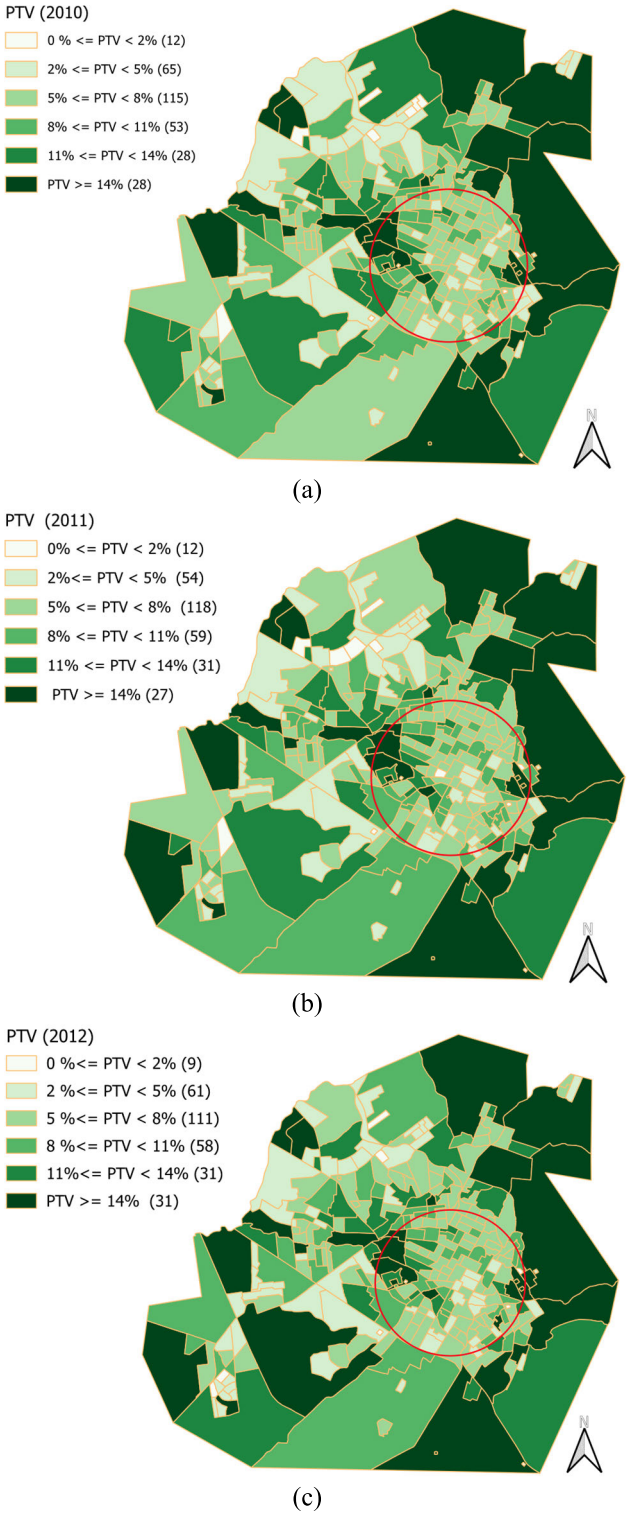


FIGURE 5. PTV by CTs for the years 2010 (a), 2011 (b), and 2012 (c).

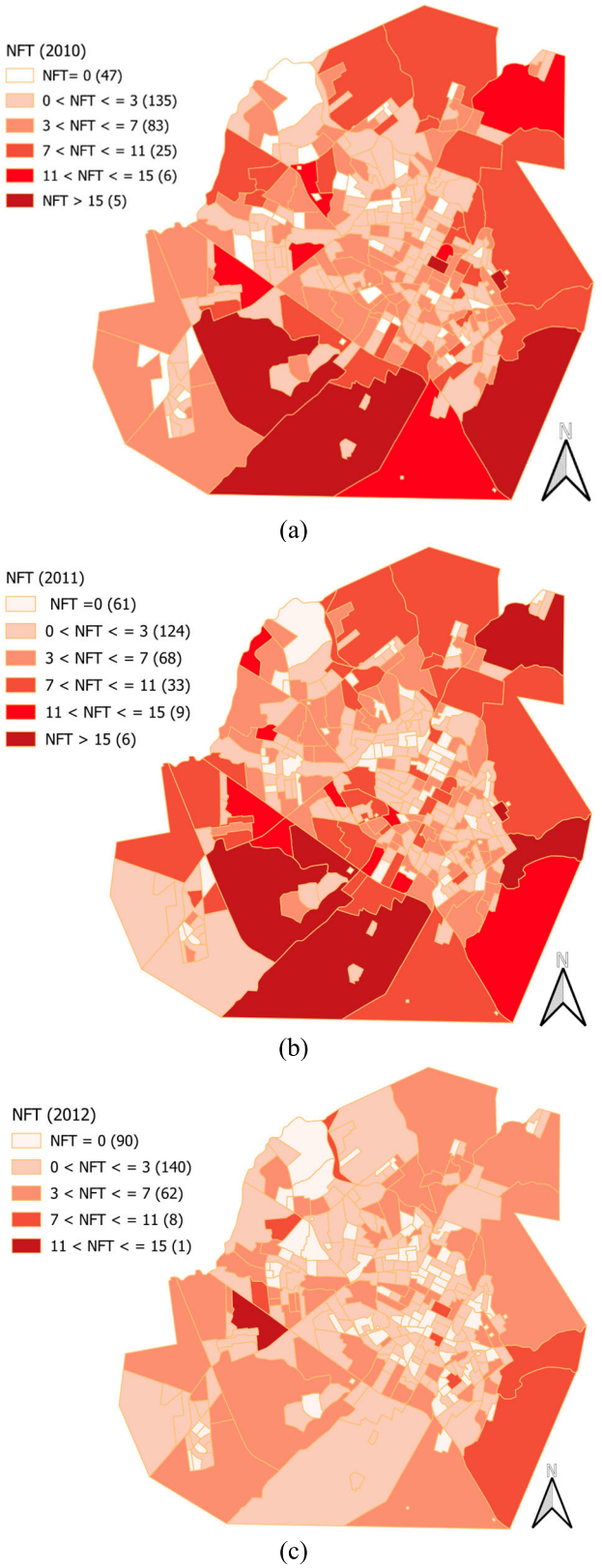


FIGURE 6. NFT by CTs for the years 2010 (a), 2011 (b), and 2012 (c).

On the other hand, thematic maps and statistical summaries are shown in Fig. 7 and Table 5, respectively, for the *PFT* variable.

Figs. 7 (a)–(c) shows *PFT* distribution by CTs from 2010 to 2012. CTs with the highest *PFT* were found in the southeast, west, and northwest regions in 2010 and in the southeast, west, northwest, and northeast areas in 2011.

PFT fell drastically in 2012, too, due to the *NFT* drop in 2012 – Fig. 6 (c).

Table 5 shows some descriptive statistics metrics for *PFT* by CTs from 2010 to 2012. It is worth pointing out that *PFT* has maximum values exceeding 100% for all years. This is possible because several shutdowns can occur in the same transformer over the year in a CT_i, that is, if $NFT_i > T_i$, then $PFT_i > 100\%$ according to (9).

TABLE 5. Statistical summary for PFT by census tracts.

Parameters	Year Evaluated		
	2010	2011	2012
Max.	400.00	366.67	300.00
Min.	0.00	0.00	0.00
Mean	48.39	49.30	30.83
Standard Deviation	50.11	53.22	45.34

Finally, there is a distinction between *NFT* and *PFT*. *NFT* only represents the number of transformer faults per CT. On the other hand, *PFT* is a measure of CTs' vulnerability to faults, as it considers the percentage of damaged transformers over the year for each CT.

3) GEOGRAPHICALLY WEIGHTED SUMMARY STATISTICS

Spearman's correlation global coefficient is shown in Table 6 between $PTV - PFT$ and $PTV - NFT$ variables for the years 2010 to 2012, according to (1). *PTV* and *NFT* have a moderate positive global correlation for all years. Thus, a cause-effect relationship can exist between them; therefore, an increase in *PTV* causes an increase in *NFT*. Conversely, there is no significant global correlation between *PTV* and *PFT* variables. Additional studies must be performed to confirm or refute this hypothesis.

TABLE 6. Spearman's global correlation coefficient.

Pair of Variables	Year Evaluated		
	2010	2011	2012
PTV and *PFT*	0.1261	0.1573	0.2405
PTV and *NFT*	0.3535	0.3390	0.4482

Noteworthy, global metrics do not consider neighborhood structure represented by SWM. Additionally, correlation at the global level represents all cities' CTs by a single numerical value. Therefore, global and local correlations can present different results, and no correlation at a global level may be confirmed at the local level or vice versa [33].

For a more detailed study, GWEA is performed considering neighborhood structure via SWM, obtained from the Gaussian kernel in (3). Thematic maps from GW metrics (4)–(7) are done in this section.

In this sense, Figs. 8 (a)–(c) show the standard deviation GW for *PTV* by CTs from 2010 to 2012. For all years, there has been a high local variability in peripheral regions in the extreme south, southeast, and northeast. Noteworthy, these regions are close to the city's rural area under study. High local variability indicates nearby CTs with very different *PTV* values.

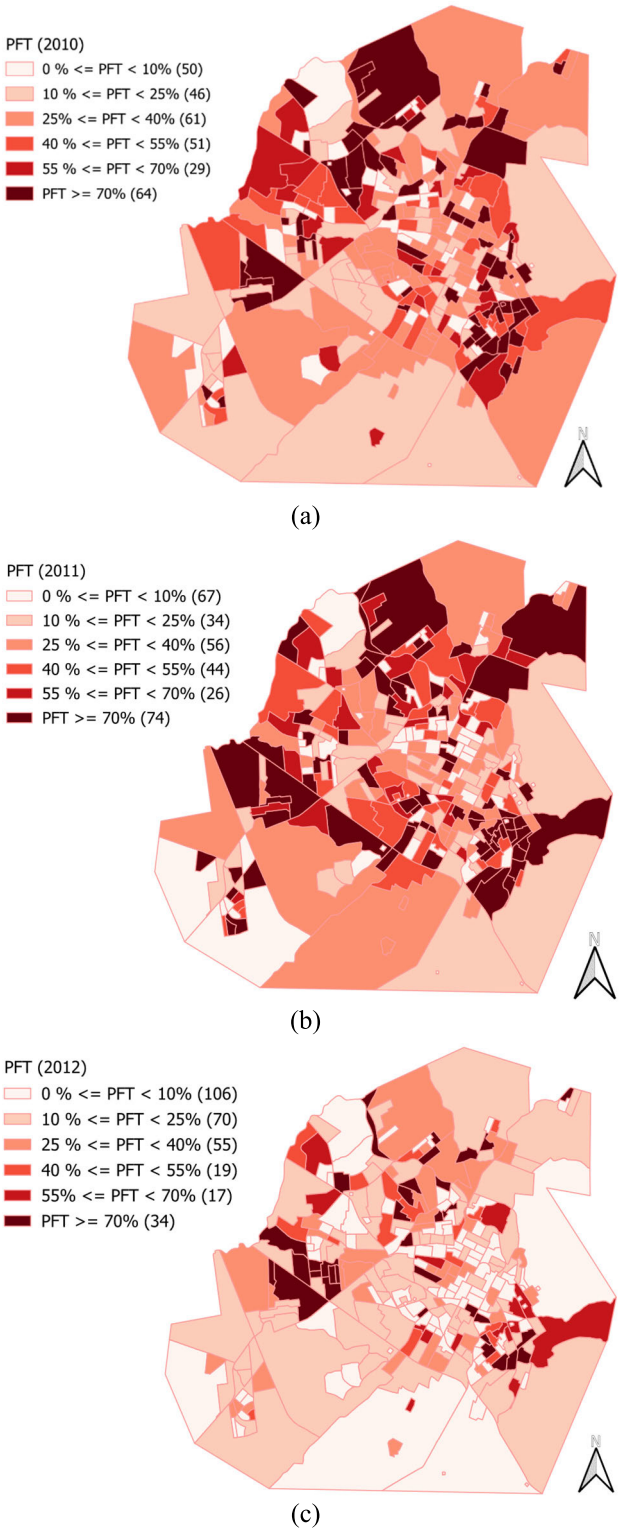


FIGURE 7. PFT by CTs for the years 2010 (a), 2011 (b), and 2012 (c).

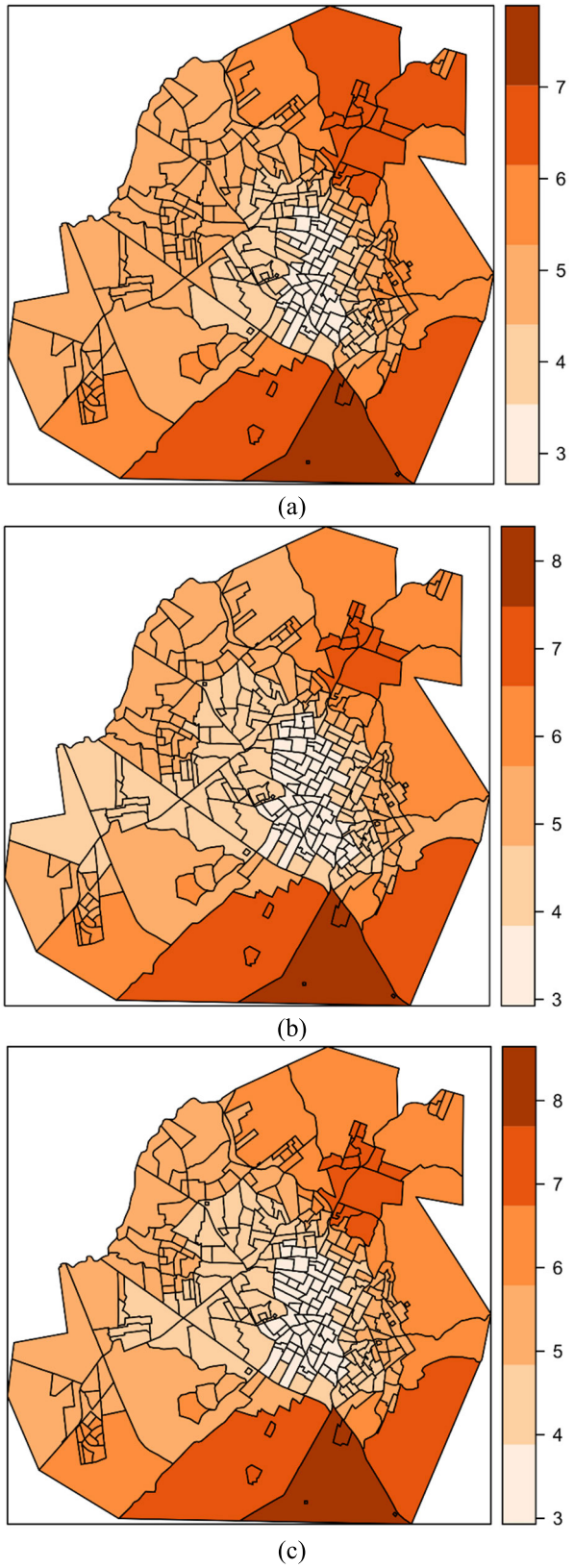


FIGURE 8. GW standard deviation for PTV by CTs for the years 2010 (a), 2011 (b), and 2012 (c).

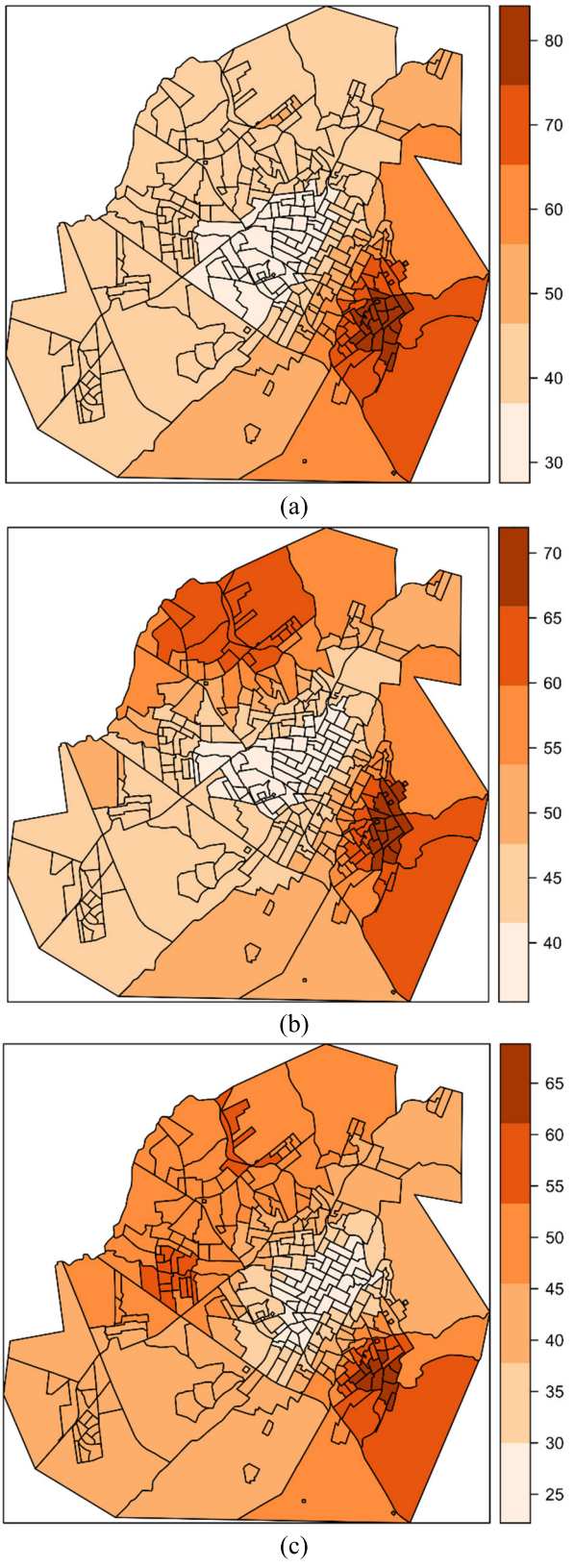


FIGURE 9. GW standard deviation for PFT by CTs for the years 2010 (a), 2011 (b), and 2012 (c).

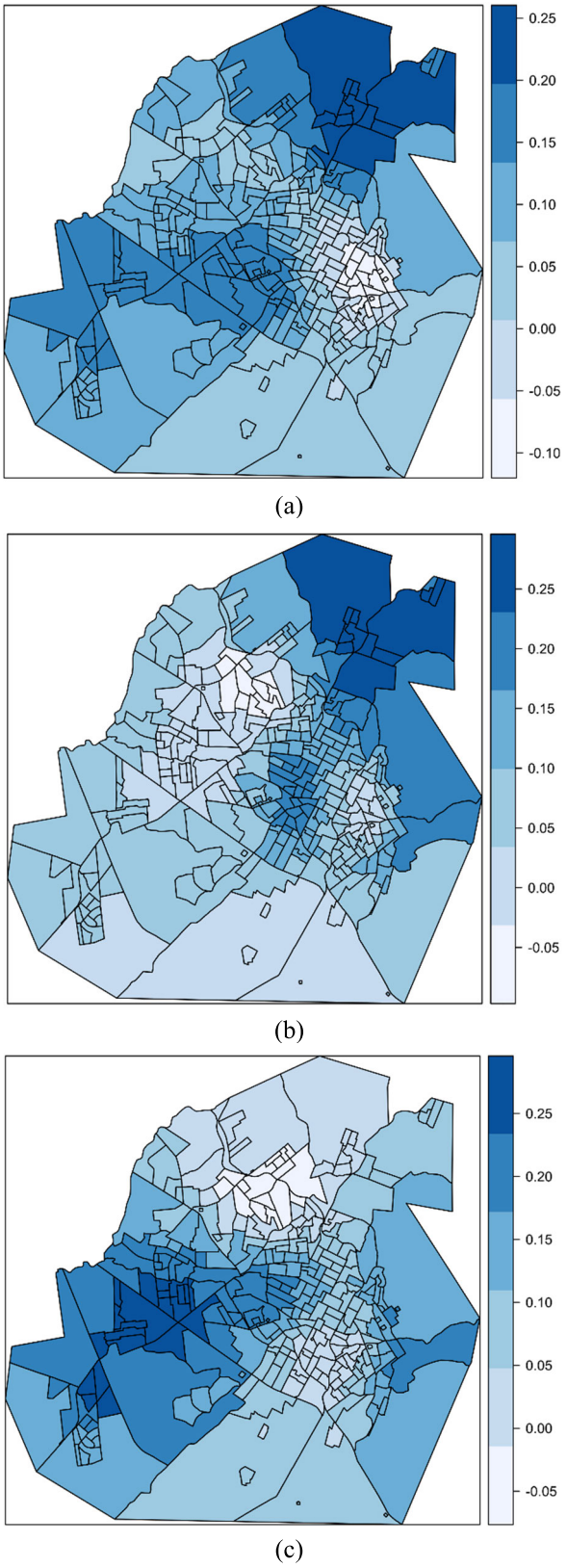


FIGURE 10. GW local Pearson's correlation between NFT and PTV for the years 2010 (a), 2011 (b), and 2012 (c).

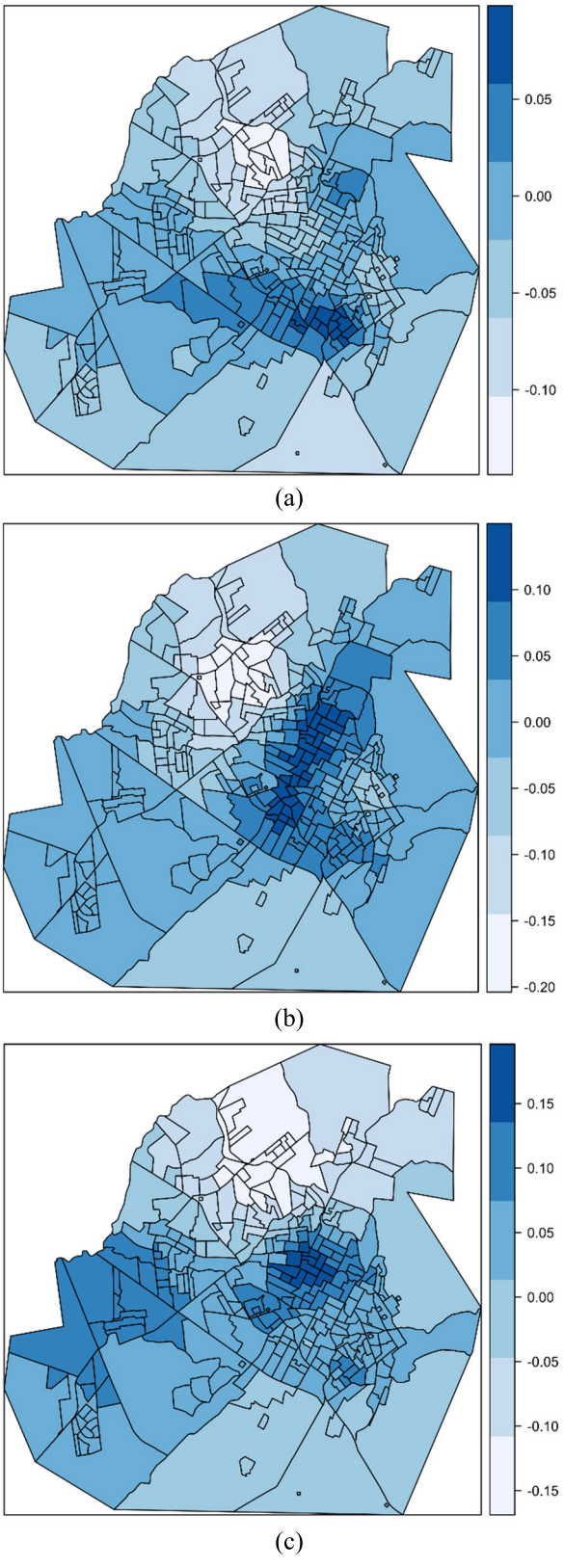


FIGURE 11. GW local Pearson's correlation between PFT and PTV for the years 2010 (a), 2011 (b), and 2012 (c).

Figs. 9 (a)–(c) show the standard deviation GW for *PFT* by CTs from 2010 to 2012. There is a high local variability in peripheral regions in the extreme southeast in 2010 and northwest and southeast regions in 2011 and 2012. Noteworthy is the attenuation of regional variability in 2012.

GW correlation is shown in Figs. 10 (a)–(c) between *NFT* and *PTV* from 2010 to 2012. Both variables have a non-stationary relationship with a moderate positive GW correlation in the northeast (in 2010 and 2011) and southwest (in 2012) regions. A positive correlation means that, in these regions, a rise in *NFT* is accompanied by an increase in *PTV* and vice versa.

Conversely, Figs. 11 (a)–(c) show the GW correlation between *PFT* and *PTV* from 2010 to 2012. Both variables have a non-stationary relationship; however, they have no significant correlation in all CTs and years evaluated.

Therefore, from Fig. 10, there is numerical evidence that tree vegetation (*PTV*) is a relevant factor associated with interruptions in the utility grid (*NFT*) in the northeast (in 2010 and 2011) and southwest (in 2012) regions.

On the other hand, the GW correlation between *PTV* and *PFT* is not significant from Fig. 11. Therefore, only the tree vegetation variable does not make regions vulnerable to faults in the city under study. According to Fig. 4, several other variables are associated with faults. Combining several variables could make some city regions vulnerable to faults.

ESDA applied in this study supports the execution of more sophisticated SDA models to add other variables that, together with *PTV*, can make some city areas vulnerable to faults.

Noteworthy, the non-stationary relationship between *PTV* and *NFT* and between *PTV* and *PFT* indicates that a global spatial regression model would not be appropriate for modeling faults; conversely, a local regression model would be more relevant to represent the non-stationarity at the local level [33].

Finally, an adaptive bandwidth with the influence of 30 closest CTs is considered in GWEA. The closest CT value corresponds to 10% of CTs in the city under study where there are $n = 301$ CTs [26].

IV. CONCLUSION

In this study, a fundamental step of spatial data analysis (SDA) called exploratory spatial data analysis (ESDA) was applied to produce thematic maps that showed the city's regions or census tracts (CTs) whose feeders are vulnerable to faults due to tree vegetation.

Actual data georeferenced by CTs were vital for this study: the dependent or study variables number of faults in transformers (*NFT*), percentage of faults in transformers (*PFT*), and independent or explanatory variable percentage of tree vegetation (*PTV*).

PTV variable was obtained from an enhanced method for tree vegetation mapping by areas using multilayer perceptron (MLP) artificial neural network (ANN) trained on high-resolution images from Google Earth.

ESDA was performed by applying descriptive statistics metrics and visualizing the spatial distribution of the variables in the city's geographic space. Geographically weighted (GW) summary statistics showed local spatial variability for *PFT* and *PTV* variables.

GW Pearson's local correlation showed a moderate positive correlation between *NFT* and *PTV* in the northeast (in 2010 and 2011) and southwest (in 2012) regions. Conversely, the GW correlation between *PTV* and *PFT* is insignificant for all evaluated years.

Therefore, tree vegetation (*PTV*) is a relevant factor associated with interruptions (*NFT*) in the utility grid for some regions. Conversely, only the tree vegetation variable does not become regions vulnerable to faults in the city. There are other variables traditionally associated with faults: electrical discharges, fires, equipment failures, and adverse weather conditions. The set of several variables can make some city regions vulnerable to faults.

ESDA performed in this study will support the implementation of more advanced models for estimating areas whose feeders are vulnerable to faults. Furthermore, incorporating other variables, such as those mentioned in the previous paragraph, will provide more robustness in future works.

REFERENCES

[1] S. S. Gururajapathy, H. Mokhlis, and H. A. Illias, "Fault location and detection techniques in power distribution systems with distributed generation: A review," *Renew. Sustain. Energy Rev.*, vol. 74, pp. 949–958, Jul. 2017, doi: 10.1016/j.rser.2017.03.021.

[2] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electr. Power Syst. Res.*, vol. 152, pp. 140–147, Nov. 2017, doi: 10.1016/j.epsr.2017.07.005.

[3] Q. Zhou, X. Li, J. Liao, and T. Xiong, "Power failure risk assessment and management based on stochastic line failures in distribution network including distributed generation," *IEEJ Trans. Electr. Electron. Eng.*, vol. 13, no. 9, pp. 1303–1312, Sep. 2018, doi: 10.1002/TEE.22696.

[4] A. S. Santos, L. T. Faria, M. L. M. Lopes, and C. R. Minussi, "Power distribution systems' vulnerability by regions caused by electrical discharges," *Energies*, vol. 16, no. 23, p. 7790, Nov. 2023, doi: 10.3390/en16237790.

[5] D. Cerrai, P. Watson, and E. N. Anagnostou, "Assessing the effects of a vegetation management standard on distribution grid outage rates," *Electr. Power Syst. Res.*, vol. 175, Oct. 2019, Art. no. 105909, doi: 10.1016/j.epsr.2019.105909.

[6] W. O. Taylor, P. L. Watson, D. Cerrai, and E. Anagnostou, "A statistical framework for evaluating the effectiveness of vegetation management in reducing power outages caused during storms in distribution networks," *Sustainability*, vol. 14, no. 2, p. 904, Jan. 2022, doi: 10.3390/su14020904.

[7] M. Doostan, R. Sohrabi, and B. Chowdhury, "A data-driven approach for predicting vegetation-related outages in power distribution systems," *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 1, pp. 1–21, Jan. 2020, doi: 10.1002/2050-7038.12154.

[8] T. Dokic and M. Kezunovic, "Predictive risk management for dynamic tree trimming scheduling for distribution networks," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 4776–4785, Sep. 2019, doi: 10.1109/TSG.2018.2868457.

[9] A. U. Melagoda, T. D. L. P. Karunarathna, G. Nisaharan, P. A. G. M. Amarasinghe, and S. K. Abeygunawardane, "Application of machine learning algorithms for predicting vegetation related outages in power distribution systems," in *Proc. 3rd Int. Conf. Electr. Eng. (EECon)*, Sep. 2021, pp. 25–30, doi: 10.1109/EECon52960.2021.9580947.

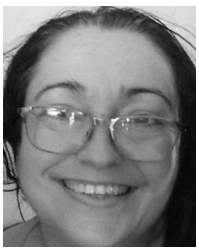
[10] J. B. Leite, J. R. S. Mantovani, T. Dokic, Q. Y. P.-C. Chen, and M. Kezunovic, "Failure probability metric by machine learning for online risk assessment in distribution networks," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf. Latin Amer. (ISGT Latin America)*, Quito, Ecuador, Sep. 2017, pp. 1–6.

[11] A. Gallaher, M. Graziano, and M. Fiaschetti, "Legacy and shockwaves: A spatial analysis of strengthening resilience of the power grid in connecticut," *Energy Policy*, vol. 159, Dec. 2021, Art. no. 112582.
[12] A. D. S. Santos, L. T. Faria, M. L. M. Lopes, A. D. P. Lotufo, and C. R. Minussi, "Efficient methodology for detection and classification of short-circuit faults in distribution systems with distributed generation," *Sensors*, vol. 22, no. 23, p. 9418, Dec. 2022, doi: 10.3390/s22239418.
[13] P.-C. Chen and M. Kezunovic, "Fuzzy logic approach to predictive risk analysis in distribution outage management," *IEEE Trans. Smart Grid*, vol. 7, no. 6, pp. 2827–2836, Nov. 2016.
[14] H. Wedagedara, C. Witharana, R. Fahey, D. Cerrai, D. Joshi, and J. Parent, "Modeling the impact of local environmental variables on tree-related power outages along distribution powerlines," *Electr. Power Syst. Res.*, vol. 221, Aug. 2023, Art. no. 109486, doi: 10.1016/j.epsr.2023.109486.
[15] W. O. Taylor, P. L. Watson, D. Cerrai, and E. N. Anagnostou, "Dynamic modeling of the effects of vegetation management on weather-related power outages," *Electr. Power Syst. Res.*, vol. 207, Jun. 2022, Art. no. 107840, doi: 10.1016/j.epsr.2022.107840.
[16] J. R. Parent, T. H. Meyer, J. C. Volin, R. T. Fahey, and C. Witharana, "An analysis of enhanced tree trimming effectiveness on reducing power outages," *J. Environ. Manage.*, vol. 241, pp. 397–406, Jul. 2019, doi: 10.1016/j.jenvman.2019.04.027.
[17] Z. Zhang, U. Demšar, S. Wang, and K. Virrantaus, "A spatial fuzzy influence diagram for modelling spatial objects' dependencies: A case study on tree-related electric outages," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 2, pp. 349–366, Oct. 2017, doi: 10.1080/13658816.2017.1385789.
[18] A. K. Onaolapo, R. P. Carpanen, D. G. Dorrell, and E. E. Ojo, "A comparative assessment of conventional and artificial neural networks methods for electricity outage forecasting," *Energies*, vol. 15, no. 2, p. 511, Jan. 2022, doi: 10.3390/en15020511.
[19] S. Druck, M. S. Carvalho, G. Câmara, and A. M. V. Monteiro, "Spatial analysis of geographic data," EMBRAPA Editor, Brasília, Brazil, 2004.
[20] J. Le Gallo and C. Ertur, "Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995," *Papers Regional Sci.*, vol. 82, no. 2, pp. 175–202, Apr. 1980.
[21] R. Haining, W. Stephen, and J. Ma, "Exploratory spatial data analysis in a geographic information system environment," *J. Roy. Stat. Soc.*, vol. 47, no. 3, pp. 457–469, 1998.
[22] S. Dall'erba and Z. Chen, "Exploratory spatial data analysis," in *International Encyclopedia of Human Geography*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2020, pp. 357–365, doi: 10.1016/B978-0-08-102295-5.10541-4.
[23] W. Tobler, "On the first law of geography: A reply," *Ann. Assoc. Amer. Geographers*, vol. 94, no. 2, pp. 304–310, Jun. 2004, doi: 10.1111/j.1467-8306.2004.09402009.x.
[24] L. Ventura, G. E. Felix, R. Vargas, L. T. Faria, and J. D. Melo, "Estimation of non-technical loss rates by regions," *Electr. Power Syst. Res.*, vol. 223, Oct. 2023, Art. no. 109685.
[25] M. C. Ferreira, "Introduction to geospatial analysis: Theory, techniques, and examples for geoprocessing," UNESP Editor, São Paulo, Brazil, 2014.
[26] I. Gollini, B. Lu, M. Charlton, C. Brunsdon, and P. Harris, "GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models," *J. Stat. Softw.*, vol. 63, pp. 1–52, Jun. 2015. [Online]. Available: http://www.jstatsoft.org/
[27] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ, USA: Wiley, 2002.
[28] W. Li, R. Dong, H. Fu, J. Wang, L. Yu, and P. Gong, "Integrating Google Earth imagery with Landsat data to improve 30-m resolution land cover mapping," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111563.
[29] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.
[30] QGIS Development Team. (2009). *QGIS Geographic Information System*. [Online]. Available: http://qgis.osgeo.org
[31] R Core Team, Vienna, Austria. (2022). *R: A Language and Environment for Statistical Computing*. [Online]. Available: https://www.R-project.org/
[32] *MATLAB, Version 9.13.0 (R2022b)*, MathWorks, Natick, MA, USA, 2022.
[33] R. S. Bivand, E. Pebesma, and V. Gómez-Rubio, *Applied Spatial Data Analysis With R*, 2nd ed. New York, NY, USA: Springer, 2013.

ANDRÉIA S. SANTOS received the B.S. degree in electrical engineering from AEMS, Mato Grosso do Sul, Brazil, in 2018, and the M.S. degree in electrical engineering from São Paulo State University (UNESP), Ilha Solteira, Brazil, in 2020, where she is currently pursuing the Ph.D. degree in electrical engineering. Her research interests include detecting, classifying, and localizing short-circuit faults in power distribution systems.

LUCAS TELES FARIA received the B.S. degree in electrical engineering from the Federal University of Goiás (UFG), Goiás, Brazil, in 2009, and the M.S. and Ph.D. degrees in electrical engineering from São Paulo State University (UNESP), Ilha Solteira, São Paulo, Brazil, in 2012 and 2016, respectively. He is currently an Assistant Professor with UNESP, Rosana, São Paulo. His research interests include spatial data analysis, soft computing techniques, fraud detection, analysis, and control of electrical power systems.

LETÍCIA S. BOSCHI received the B.S., M.S., and Ph.D. degrees in cartographic engineering from São Paulo State University (UNESP), Presidente Prudente, São Paulo, Brazil, in 2002, 2005, and 2011, respectively. She is currently an Assistant Professor with UNESP, Rosana, São Paulo, Brazil. Her research interests include remote sensing and spatial analysis of geographic data.



MARA LÚCIA M. LOPES received the B.S. degree in mathematics from the Federal University of Mato Grosso do Sul (UFMS), Mato Grosso do Sul, Brazil, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from São Paulo State University (UNESP), Ilha Solteira, São Paulo, Brazil, in 2000 and 2005, respectively. She is currently an Assistant Professor with UNESP. Her research interests include power transmission and distribution systems, artificial neural networks, artificial immune systems, fuzzy logic, load forecasting, and dynamic systems.

CARLOS R. MINUSSI received the B.S. degree in electrical engineering from the Federal University of Santa Maria (UFSM), Rio Grande do Sul, Brazil, in 1978, and the M.S. and Ph.D. degrees in electrical engineering from the Federal University of Santa Catarina (UFSC), Santa Catarina, Brazil, in 1981 and 1990, respectively. He is currently a Full Professor with São Paulo State University (UNESP), Ilha Solteira, São Paulo, Brazil. His research interests include smart grids, transient stability, dynamic preventive control, anticipatory systems, load forecasting, machine learning, fuzzy logic, and artificial immune systems.