Spatial-Temporal Estimation for Nontechnical Losses

Lucas Teles Faria, *Student Member, IEEE*, Joel David Melo, *Member, IEEE*, and Antonio Padilha-Feltrin, *Senior Member, IEEE*

***Abstract*—This paper presents a novel method for estimating the spatial distribution in geographical space of the nontechnical losses over time. The method progresses in two stages: in the first stage, a generalized additive model is used to generate a map of current loss probabilities. The second stage employs the Markov chain to generate a map that indicates possible future changes in loss probabilities. The method yields an assessment of the location of the nontechnical losses now and in the future at the city subarea level, even indicating the variables that have greater statistical correlation with the nontechnical losses. We apply the method to a city with approximately 81 000 consumers, and the results are compared with those obtained through inspections carried out by a Brazilian power utility. The detection rate surpasses 78% in inspected subareas. The method we propose offers improved estimation of distribution of the nontechnical losses in urban regions.**

***Index Terms*—Electricity theft, generalized additive models, nontechnical losses, spatial-point pattern analysis.**

I. Introduction

NONTECHNICAL losses (NTLs) are present in almost all electric power distribution systems and are the source of considerable expenses for many power utilities [1]. The introduction of smart grids and smart meters may contribute to a significant reduction in such costs by eliminating some types of losses [2]. However, the development of such technological advances often progresses more slowly in developing countries (particularly those with high rates of NTLs). For these reasons, a real need exists for further research on more efficient NTLs evaluation techniques.

This paper seeks to answer the following questions: where, why, and when do NTLs occur? The methodology we present here differs from that proposed in other NTLs studies in that ours considers the spatial distribution of socioeconomic characteristics and electrical infrastructure in the city subareas where the losses occur. From these spatial distributions, two loss probability maps are produced: one representing the present, which is devised through a generalized additive model (GAM), and another representing the future via the Markov chain. These loss probability maps indicate the city subareas vulnerable to NTLs.

The task of detecting NTLs is one of the most challenging in the field of power systems and, as such, NTLs detection remains a primary focus in a lot of current research. Recently in [1], a state estimation-based approach to the load estimation of distribution transformers was exploited to detect meter tampering and provide quantitative evidence of NTLs. In the literature, several techniques in the area of intelligent systems, or soft computing, have been employed, some of which are: multiple classifiers and wavelet coefficients [3]; fuzzy logic [4], [5]; text mining [6]; Bayesian networks [7]; the pattern-recognition technique via optimum path forest [8]; data mining [9]; data mining using support vector machines [10]; using extreme leaning machines [11]; and using generalized rule induction [12] and fuzzy inference systems [13].

The aforementioned references detect consumer units (CUs) with NTLs without considering the characteristics of the geographic areas in which they occur. However, such characterization can improve the detection of regions with loss and may clarify why they are more concentrated in certain areas of the city. The detection of the geographic regions where there is greater loss probabilities is useful because it enables a series of combat and prevention actions for NTLs. The loss probability maps for the present and future, using spatial temporal estimation, enable the visualization of geographical regions with the highest loss probabilities. These maps can be used as a general guide to the power utilities for all actions to treatment and reduction of the NTLs in low-voltage (LV) CUs with meters that belong.

There are other problems in power systems that can be benefit from spatial-temporal estimations. For example, in [14], a spatial-point pattern analysis to determine the input data for a spatial-temporal simulation of the load growth is explained and in [15], a spatial-temporal approach for estimating the load demand of battery electric vehicles charging in small residential areas was proposed.

In [16], a qualitative approach is presented where socioeconomic aspects that provide a favorable environment for the emergence of the NTLs are analyzed. One-hundred-two countries were evaluated from 1980 to 2000. NTLs are associated with governmental and social weaknesses, such as: political instability, government ineffectiveness, high levels of corruption, poverty, high birth rate, low human development index, etc. It appeared that the socioeconomic characteristics of the subareas where NTLs occurred were relevant, and they will be considered in this paper.

Manuscript received March 13, 2015; revised July 20, 2015; accepted August 09, 2015. Date of publication September 28, 2015; date of current version January 21, 2016. This work was supported in part by Coordination for the Improvement of Higher Education Personnel (CAPES), in part by National Council of Scientific and Technological Development (CNPq), and in part by Grant 2014/06629-0, São Paulo Research Foundation (FAPESP). Paper no. TPWRD-00323-2015.

The authors are with the Department of Electrical Enginnering at São Paulo State University (UNESP), Ilha Solteira Campus, Ilha Solteira, São Paulo 15385 000, Brazil (e-mail: lucas.teles.faria@gmail.com; joeldavidm@gmail.com; padilha@dee.feis.unesp.br).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRD.2015.2469135

In this study, a method for identifying the regions vulnerable to NTLs is presented, and the corresponding variables that contribute most significantly to the emergence of these losses are established. The method results are probability maps of the emergence of losses in the present and future. We apply the method in a Brazilian city with approximately 81 000 CUs. The maps are compared with inspections conducted by the power utility, showing a loss identification rate of more than 78%. The main contribution of the proposed method is the estimation of the spatial variation of the NTLs by city subarea over time.

This paper is organized as follows. In Section II-A, it has been discussed how to determine the loss probabilities in the present via spatial-point pattern analysis. Section II-B includes the Case-Control study, and how to determine the current loss state of each subarea through GAM is presented in Sections II-C and D. In Section III, the future loss state of each subarea via a Markov chain is determined (Sections III-A and B). In Section IV, the proposed methodology by comparing the loss states provided with real data from inspections in a Brazilian city is applied and validated (Sections IV-A and IV-B). In Section IV-C, the variables that have the greater statistical correlation with the NTLs are identified. The concluding remarks are presented in Section V.

II. Vulnerability to NTLs in the Present

This section contains techniques used to estimate, at present, the NTLs vulnerability in each city subarea.

A. Spatial-Point Pattern Analysis

Spatial analysis enables incorporating space and revealing the characteristics of the subareas where the NTLs occur. It comprises a set of tools to explore and model processes that are expressed through a spatial distribution known as geographic phenomena. Spatial analysis measures properties and relationships in order to explicitly consider the spatial location of the phenomenon under study [17].

A point process is a statistical process in which some events of interest within a limited region A are observed. The location of events generated by a point process in the area of study is called a point pattern [18].

The main interest of the spatial analysis of point events is to analyze the point patterns and determine whether they show any systematic pattern. If a pattern of point events shifts significantly compared to a stochastic distribution (usually a Poisson distribution), this is an indication of a spatial distribution other than complete randomness that should be the subject of further analysis [19].

In this paper, the CUs are represented by points in space (point events). The term "event" refers to any type of phenomenon localizable space by geographic coordinates.

In order to determine the spatial variation of the NTLs, we performed a Case-Control study.

B. Case-Control Study

One of the epidemiology applications in others areas is to associate contagious diseases with crimes, because the occurrence of diseases and crimes follows a similar pattern [20].

In this context, we consider a case-control study [21], [22]. In public health, for example, sick people (cases) are compared to healthy people (controls). It is assumed that all cases and controls have been exposed to risk factors for the disease.

According to [18], the distribution of cases is influenced by the heterogeneous distribution of the risk population. Therefore, it is necessary to estimate their spatial distribution and compare it to existing cases. The control set represents the spatial variation of the risk population. In this paper, the entire risk population is regular CUs, because the NTLs can appear in any regular CU of the city.

The location of CUs with NTLs (cases) and the CUs with no loss (controls) are the input data for the GAM. Then, the GAM is used to obtain the probabilities of finding cases (CUs with NTLs) in city subareas.

C. Generalized Additive Model

The GAM is a semiparametric model that enables inclusion of distribution network variables, socioeconomic variables, and others as a function of the city subareas [23].

The inclusion of the effects of variables in the model is affected according to [22] and is shown in (1). In (1), x is the vector of variables, β are coefficients of the variables, and $g(s)$ is a smooth function of the spatial coordinates s to model other information that is not available

$$\text{logit}\{P(s,x)\} = \log\left\{\frac{P(s,x)}{1-P(s,x)}\right\} = \beta x + g(s). \quad (1)$$

In (1), a probability surface $P(s,x)$ is estimated, where a linear effect of the x variables and a residual spatial variation represented by $g(s)$ are considered. Moreover, (1) has a logistic regression model extended by an additive component $g(s)$, which is smoothly varying in space and is independent of the x variables.

The procedure for estimating β and $g(s)$ is based on the usual iterative methods for the GAM [23]. These parameters are calculated simultaneously during the iterative process for the GAM. After several experiments, we observe that the influence of these parameters in the method efficiency depends on the spatial distribution of the input data.

Applying (1) to each location or point belonging to the base GAM input data obtains a probability $P(s,x)$ of the emergence of NTLs at the spatial location s.

D. Determining the Present Loss State

The vulnerability for the NTLs in subareas is represented by three loss states: regular, attention, and critical. These states indicate that vulnerability to losses are low, intermediate, and high, respectively.

After execution of the GAM, every point of the case-controls database is associated with a probability $P(s,x) \in [0,1]$ of the emergence of NTLs. The current loss state of each city subarea is determined by means of grouping and sorting of the probabilities contained in each subarea.

Let us consider m points from the case-control database contained in a particular subarea. The objective is to determine the percentage of these m points that are in each of the loss

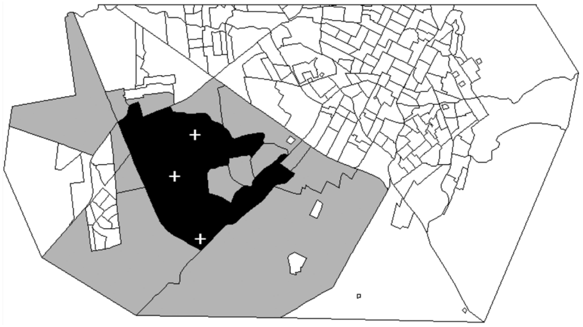


Fig. 1. Map of the southern part of a city highlighting one of its subareas (in black) that contains three points from a case-control database and 14 neighboring subareas (in gray).

states: regular, attention, and critical. These percentages represent a probability for each loss state and are placed in a vector $[\,\%P_R \quad \%P_A \quad \%P_C\,]_{1\times 3}^{A}$. This vector is associated with each subarea A. The terms $\%\mathrm{P}_{\mathrm{R}}$, $\%\mathrm{P}_{\mathrm{A}}$, and $\%\mathrm{P}_{\mathrm{C}}$ represent the percentage of the m probabilities contained in the subarea A, which are in states of regular, attention, and critical losses, respectively. Modal class determines the current loss state in the subarea, that is, the loss state in which the probabilities are most frequent.

There is a lower bound (LB) and an upper bound (UB) to each of the three loss states such that $\mathrm{P}(s,x) \in [LB, UB]$. The LB and UB are specific to each under analysis region. They are calibrated for each loss state after an exploratory analysis on the GAM probabilities. Techniques as standard deviation, natural boundaries, quantiles, equal intervals, among others, can be used for determining the class limits of the loss state, as explained in [19].

For illustrative purposes, the southern area of a city partitioned by the respective subarea, as shown in Fig. 1, is considered. The subarea in black contains three locations of CUs, represented by points from the case-control database $(m = 3)$. It is assumed that the bounds for each loss state such that $P(s,x) \in [LB, UB]$ are $P(s,x) \in [0, 1/3]$ in the regular state; $P(s,x) \in (1/3, 2/3]$ in the attention state, and $P(s,x) \in (2/3, 1]$ in the critical state. It is assumed also that the probabilities associated with each point, resulting from the implementation of the GAM, are 0.02 (regular state), 0.09 (regular state), and 0.38 (attention state). Therefore, probabilities of 2/3 and 1/3 are in the regular and attention states, respectively, and 0/3 is in the critical state. The current loss probabilities for each loss state for the black subarea of Fig. 1 are $[2/3 \quad 1/3 \quad 0]$. It can be concluded that the subarea is in the state of regular loss, because the GAM probabilities are more frequent in this loss state. The subarea in black has common boundaries with 14 other subareas (highlighted in gray).

III. Vulnerability to NTLs in the Future

In this section, the Markov chain is used to estimate the subareas vulnerable to NTLs in the future.

The Markov chain is a mathematical model used to describe stochastic processes [24]. The outputs are the probabilities of occurrence of each one of the discrete states as discrete-time functions. It is an empirical method that focuses on the relationships among its variables, assuming the relationships of the past will remain in the near future, per

$$\Pi_{(t+1)} = \mathbf{P}\Pi_{(t)} \tag{2}$$

where $\Pi_{(t)}$ is the system state at time t, $\Pi_{(t+1)}$ is the system state after an interval of $(t+1)$, and $\mathbf{P}$ is the transition matrix among states. The transition matrix represents the probability of a state i to stay the same or change to a state j during the time interval t. The conditional probabilities $P\{X(t_{k+1}) = x_{k+1} | X(t_k) = x_k\}$, called transition probabilities, represent the state probability $X(t_{k+1})$ to be x_{k+1} at time t_{k+1} if the state $X(t_k)$ is x_k at time t_k. The transition probabilities are obtained from samples at a time interval [24].

A stochastic process is considered a Markov process of the first order if the future state depends only on the present state and not on past states. This methodology does not ignore the past; however, it assumes that all past information is concentrated in the present state. The transition probabilities do not change with time, which characterizes a stationary process.

According to [19], there is no single solution to model dynamic spatial phenomena. In this paper, some considerations are made to enable the use of the Markov chain to model the spatial variation of the NTLs. For a short-term horizon, the NTLs are approximately stationary. The losses are influenced by the socioeconomic conditions of the city, which change slowly. Thus, the process is approximately stationary, and states in the recent past resemble the present states.

A. Transition Matrix

The probability is usually defined as (3). The probability $P(E)$ of an event E is the relative frequency at which this event occurs in a series of attempts under constant conditions

$$P(E) = \frac{N_E}{n} \tag{3}$$

where N_E is the number of times that event E occurs in n trials. The occurrence of an event E in a particular observation is entirely uncertain; however, the relative frequency with which it occurs in repeated observations has stable properties. If not, probability theory may not apply [25]. N_E represents the number of CUs with NTLs found in a total of n CUs inspected by teams in city subareas.

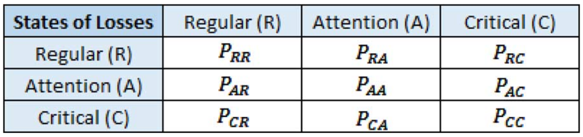
For a fixed time interval, the loss probability is obtained from (4) after adjusting (3). Given that $\hat{P}_A^y$ is the annual estimated loss probability related to the subarea A in the year y

$$\hat{P}_A^y(\text{Losses}) = \left(\frac{\text{Number of CUs with NTLs}}{\text{Number of CUs Inspected}}\right). \tag{4}$$

Having determined the current loss state of each city subarea, the transition matrices are used to estimate the future loss state.

Fig. 2 presents the general structure of the transition matrix with all of its elements. P_{ij} is the probability for a subarea to remain in the same loss state (for $i = j$) or change state (for $i \neq j$) after a fixed transitional period of one year.

The transition matrix of each subarea is obtained from the number of annual changes in loss state (or lack thereof) for each



States of Losses	Regular (R)	Attention (A)	Critical (C)
Regular (R)	P_{RR}	P_{RA}	P_{RC}
Attention (A)	P_{AR}	P_{AA}	P_{AC}
Critical (C)	P_{CR}	P_{CA}	P_{CC}

Fig. 2. Loss state transition matrix by city subarea.

subarea and its neighbors. The loss state is defined from the loss probabilities in (4).

The goal is to obtain the elements of the transition matrices for all subareas. In this example, the LB and UB boundaries of each of the three loss states are the same as in the previous example, that is, $P(s,x) \in [0, 1/3]$ in the regular state; $P(s,x) \in (1/3, 2/3]$ in the attention state; and $P(s,x) \in (2/3, 1]$ in the critical state. The building of the transition matrix related to the subarea highlighted in black in Fig. 1 is considered in detail. At the beginning, all of the elements of the matrix are null. This subarea has loss probabilities that are defined in (4), with 0.05 (regular state), 0.08 (regular state), and 0.37 (attention state) in 2009, 2010, and 2011, respectively. So, in two possible annual transitions (from 2009 to 2010 and from 2010 to 2011), one regular state remained regular (2009 to 2010); in another possible transition (2010 to 2011), the regular state changes to attention state. Therefore, a unit in the elements of the transition matrix P_{RR} (probability of the regular state remain in this same state the next year) and P_{RA} (probability of the regular state changes to the attention state in the next year) is increased. This process is repeated for the other 14 neighboring subareas of the subarea, highlighted in black in Fig. 1, to obtain the full transition matrix.

In the example of Fig. 1, 15 subareas (the subarea highlighted in black and 14 neighboring subareas it) each contain two transitions (from 2009 to 2010 and from 2010 to 2011), resulting in 30 transitions. The number of transitions in which loss state i remained the same or changed to loss state j is calculated after each fixed transition period of one year.

In (5), we present the transition matrix $\boldsymbol{P}$ for the subarea highlighted in black in Fig. 1 after normalization of the amount of changes in the loss state (or lack thereof). It is worth mentioning that the sum of each transition matrix line is unitary [24]

$$\boldsymbol{P} = \begin{bmatrix} \frac{3}{5} & \frac{1}{10} & \frac{3}{10} \\ 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}_{3\times 3}. \tag{5}$$

B. Determining the Future Loss State

The forecast of the future loss state is obtained by multiplying the current loss state vector matrix by the transition matrix for the subarea. The result is a 1 × 3 vector containing the future loss states' probabilities. The most probable resulting vector determines the loss state predicted for the evaluated subarea.

In (6), the future loss state of the subarea highlighted in black in Fig. 1 is estimated considering the loss probability vector and the transition matrix shown in Sections II-D and III-A, respectively

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{5} & \frac{1}{10} & \frac{3}{10} \\ 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{6}{15} & \frac{1}{15} & \frac{8}{15} \end{bmatrix}. \tag{6}$$

Note that the probability for the subarea to be in the states of regular, attention, and critical loss in 2012 is 6/15, 1/15, and 8/15, respectively. Therefore, it is concluded that the loss state for this subarea will change from regular (current loss state) to critical (future loss state).

The method proposed here can be summarized as follows.

Step 1) *Build the Case-Control database.*
A database is constructed from georeferenced inspections and the power utility customer base. This base is associated with socioeconomic variables from the census and is the input data for the GAM.

Step 2) *Execute the GAM in (1) and construct the probability map.*

Step 3) *Determine the current loss state of each subarea using the probabilities calculated in the previous step.*
The current loss state of each subarea is determined based on the percentage of the probability map that is contained in each area and the loss state it is in: regular, attention, or critical. The current loss state of the subarea is determined by the loss state in which the GAM probabilities are most frequent.

Step 4) *Estimate the transition matrix $\boldsymbol{P}$ for each subarea.*
Matrix $\boldsymbol{P}$ is constructed from (4) by evaluating the annual transitions in the loss state for subareas and their neighbors.

Step 5) *Forecast the future loss state.*
The future loss state forecast is obtained through matrix multiplication between the vector of current loss states (Step 3) and the transition matrix (Step 4). The result is a vector containing the probabilities of each subarea being in each of the three possible future loss states. The loss state with the greatest probability in the resulting vector determines the forecasted loss state for the evaluated subarea.

IV. Case Study: Application in a Brazilian City

In this section, the application of the methodology proposed here is examined in detail. All simulations were performed on R software version 2.15.3. The libraries used are as follows: mgcv, spatial, spatstat, and splancs. The mgcv library is employed to execute the GAM [26]. This library implements the nonparametric estimator, as in [22].

A. Input Data

In the present study, actual data from georeferenced inspections provided by a Brazilian power utility and data from the latest Brazilian census are used [27], [28].

The inspected CUs belong to an LV network of a medium-sized city in the State of São Paulo, Brazil, with approximately 200 000 inhabitants. Inspections conducted in 301 subareas of the urban area are used.

The power utility inspected 9278 CUs as 2463, 1103, 3777, and 1935 in 2009, 2010, 2011, and 2012, respectively. Inspection teams found 1133 CUs with NTLs as 165, 88, 454, and 426 from 2009 to 2012, respectively.

The inspections are commonly performed in the form of campaigns; that is, officials inspect all CUs in a city subarea in which there may be a high number of NTLs.

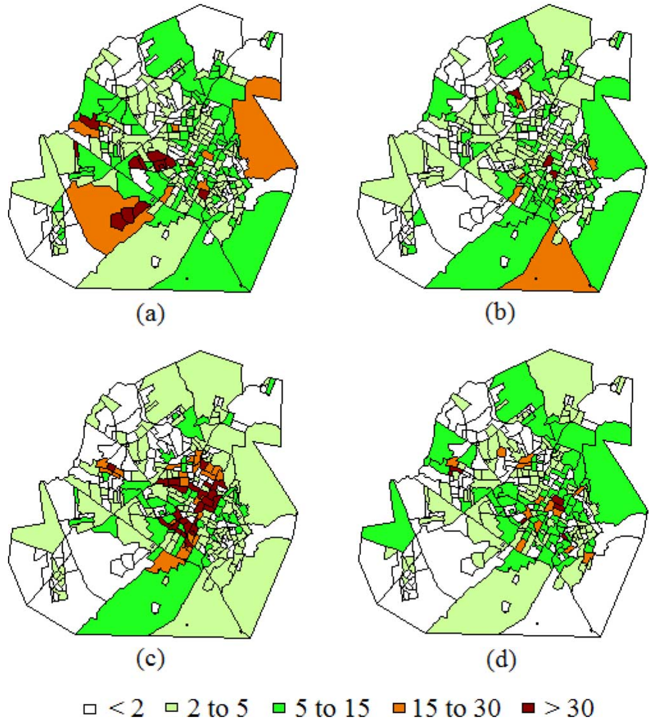


Fig. 3. Thematic maps of the total number of CUs inspected annually by subareas for the years 2009, 2010, 2011, and 2012, which are (a), (b), (c), and (d), respectively.

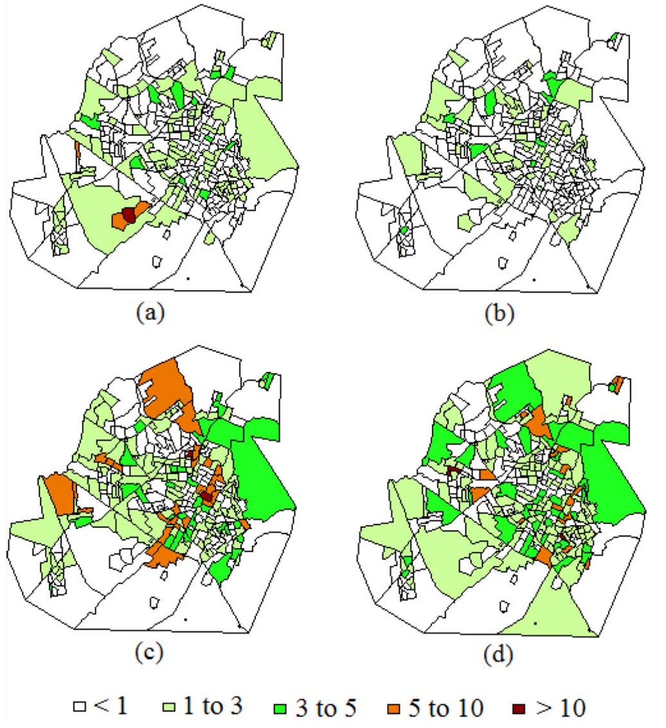


Fig. 4. Thematic maps of the total number of CUs with NTLs annually by subareas for the years 2009, 2010, 2011, and 2012, which are (a), (b), (c), and (d), respectively.

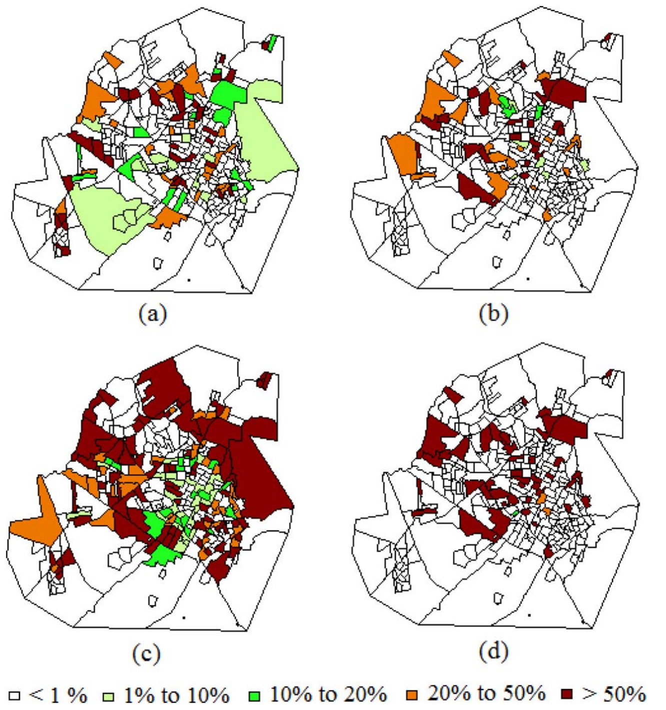


Fig. 5. Thematic maps related to annual loss probabilities by subareas for the years 2009, 2010, 2011, and 2012, which are (a), (b), (c), and (d), respectively.

In Figs. 3–5, the information from inspections has been represented graphically by city subareas.

Fig. 3 shows the number of annual inspections by subareas. Those in dark red were targeted inspection campaigns.

Fig. 4 shows the number of CUs with NTLs by subareas. Equation (4) is employed to calculate the annual loss probabilities for the subareas in Fig. 5.

The CUs of the loss database are not georeferenced. Therefore, in order to apply the method proposed here, they are grouped by the transformers to which they are connected; the distribution network transformers are georeferenced, unlike the CUs.

1) Case-Control Database Construction: The input data are sampled according to the case-control study. The set of cases consists of all irregular CUs from 2009 to 2011 (the year 2012 is reserved to validate the proposed methodology). The set of controls is a random sampling of regular CUs for which loss has not occurred.

The database is built with a (9:1) sampling scheme; that is, for each irregular CU that belongs to the set of cases, there are nine regular CUs belonging to the set of controls. The cases contain all 707 CUs that are found to be guilty of irregularities, and the controls contain 6363 regular CUs, randomly selected from all of the CUs inspected; both from 2009 to 2011.

Fig. 6 displays the distribution of CUs in the case-control database on the map of the urban area. The central region (inscribed in the circle) has the largest number of inspections and CUs with NTLs.

It is noteworthy that the sampling scheme (9:1) represents the real situation of field inspections of Brazilian utilities; on average, every ten inspections reveals one irregular CU. In this paper, for example, on average, there is an irregular CU for every 8.2 inspections.

2) Description of the GAM Variables: In the GAM, 11 x variables are used as in (1). These variables are described in Table I and characterize the socioeconomic aspects, the power

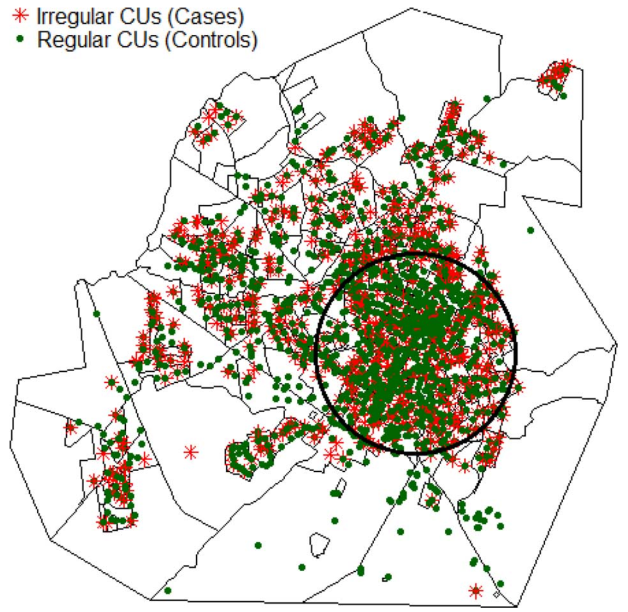


Fig. 6. Spatial distribution of CUs in the case-control database in the urban area of the city from 2009 to 2011. The CUs are classified into irregular (cases) and regular (controls).

TABLE I
DESCRIPTION OF THE VARIABLES USED IN THE GAM

	Variables	Description
Socioeconomic Variables	*Average Number of Residents*	Average number of residents in permanent households
	Average Income	Nominal average monthly income of persons 10 years of age or older (with and no income)
	%Rented Residency[a]	Percentage of leased private households
	%Residency with Water[a]	Percentage of private households with water supply
	%Residency with Garbage Collection[a]	Percentage of private households with garbage collection service
	%Residency with Pavement[a]	Percentage of private households with pavement in surrounding areas
	%Literates[b]	Percentage of literate people five years of age or older
Technical Variables	*LOSSTRAFO*	Percentage of irregular CUs by transformer
	LOSSAREA	Percentage of irregular CUs in the subarea
	NTRAFO	Number of transformers in the subarea
	CAMPAIGN	Binary variable indicating whether there were campaign actions in the subarea

[a] Percentage with respect to total residency in the subarea.
[b] Percentage with respect to the total number of individuals in the subarea.

grid, and the concentration of the NTLs by city subarea. These variables are grouped into two groups: 1) socioeconomic variables (those from the census) and 2) technical variables (those from the power utility). They were chosen after an exploratory analysis in accordance with the recommendations of the power utility experts.

Among the socioeconomic variables available in census data by subarea, seven were selected: *Average Number of Residents, Average Income, % Rented Residency, % Residency with Water, % Residency with Garbage Collection, % Residency with Pavement*, and *% Literate*. The other variables were taken from inspections; *LOSSTRAFO* and *LOSSAREA* are related to the concentration of irregular CUs by transformer and by subarea, respectively; *NTRAFO* is linked to the extension of the distribution network, and *CAMPAIGN* indicates whether there were a high number of CUs inspected (more than 30); all variables are related to the subarea.

B. Validation of the Methodology

The method was validated by comparing the forecast to inspections for the year 2012.

For the application of the proposed method, the bounds for each loss state such that $\mathrm{P(s,x)} \in \lceil \mathrm{LB}, \mathrm{UB} \rceil$ are $\mathrm{P(s,x)} \in [0, 0.15]$ in the regular state; $\mathrm{P(s,x)} \in (0.15, 0.35]$ in the attention state; and $\mathrm{P(s,x)} \in (0.35, 1]$ in the critical state. Now we can evaluate the loss probabilities in the present and future and define the loss states in each city subarea.

In order to validate the estimated loss state, the loss state for each subarea was determined again from the number of NTLs in 2012. A subarea is considered to be in regular or critical state if the number of NTLs is less than three or greater than eight, respectively. If these conditions are not met, the subarea is in the attention state.

Fig. 7 shows the result of applying the proposed methodology. Fig. 7(a) and (b) presents the current loss state of each subarea (via GAM) and the forecasted loss states (via Markov chain), respectively. As noted before, in order to validate the method, the subareas are again categorized into one of three loss states according to the number of NTLs in 2012 and compared to the forecast obtained in Fig. 7(b). In a total of 301 subareas in the urban area, 280 subareas were evaluated. Among these, there were 219 hits (78.2%) and 61 errors (21.8%). The states of 21 subareas were not forecasted due to a lack of data.

We observed that although the proposed method has not properly recognized subareas in the critical state, the proposal detects subareas which do not change the state [see Fig. 7(c)], considering a conservative approach, that is, the largest possible margin of error for each loss state. In addition, the proposal identifies the locations that should not be visited.

C. Variables With Statistical Significance

In addition to being essential for the construction of the probability map, the GAM indicates that the variables have statistical relevance and, therefore, influence the NTLs. Table II presents the estimates, standard deviations, and statistical significance for the GAM x variables in (1).

The p-value corresponds to a significance for which the hypothesis H_0 (null hypothesis) cannot be rejected. H_0 is the hypothesis in which the variables are not relevant. H_0 is rejected if the p-value is less than or equal to the predefined level of significance. In this paper, the significance level is fixed to $\alpha = 0.1$.

In this context, the significant variables (boldface in Table II), that is, those that influence the NTLs in the analyzed city, are *LOSSTRAFO* (p-value $< 2.10^{-16}$), *LOSSAREA* (p-value $5.92.10^{-6}$), *CAMPAIGN* (p-value $< \quad 2.10^{-16}$), *NTRAFO* (p-value $< 6.07.10^{-4}$), and *Average Income* (p-value $4.81.10^{-4}$). The significance of *LOSSTRAFO* and *LOSSAREA* indicates a higher concentration of the NTLs in transformers and in subareas with a high percentage of irregular CUs, respectively. The significance of *NTRAFO* indicates a higher concentration of NTLs in subareas with more transformers,

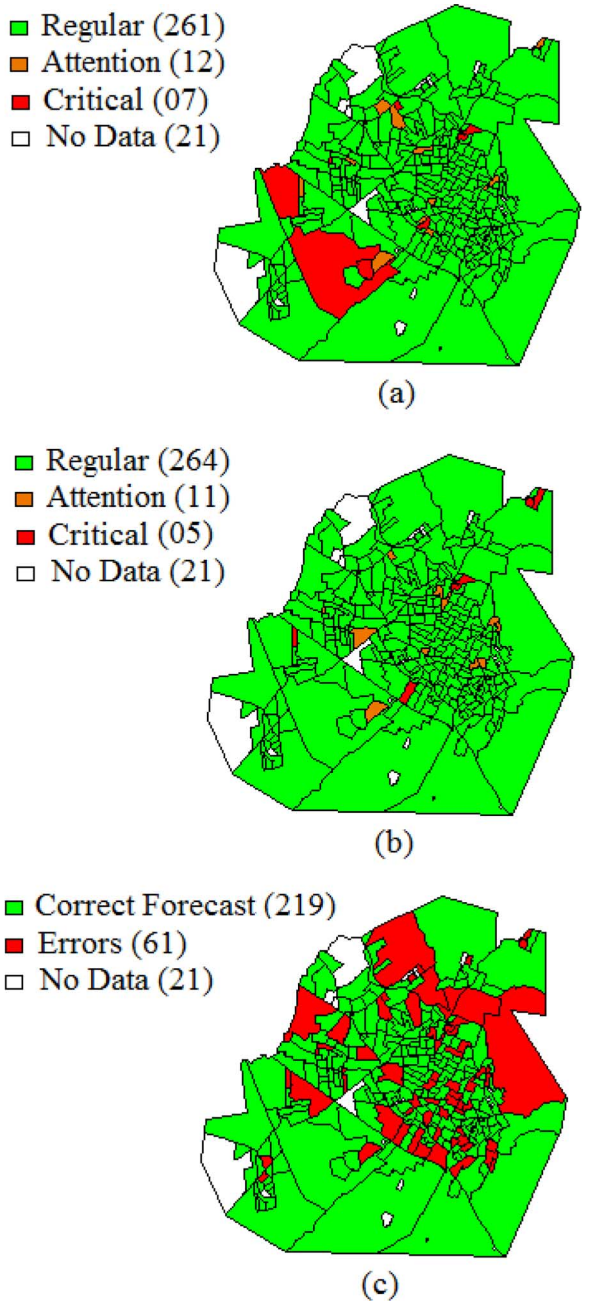


Fig. 7. (a) and (b) Current loss state (obtained via GAM) and forecasted loss state (obtained via Markov chain), respectively, by the subarea of the urban area. (c) Validation methodology by comparison between the forecasted loss state in (b) and the number of NTLs, by subarea, in 2012.

TABLE II
ESTIMATES, STANDARD DEVIATION AND STATISTICAL SIGNIFICANCE OF THE GAM VARIABLES

Variables	Estimates	Standard Deviation	p-value
Average Number of Residents	-0.10250	0.25620	0.68922
Average Income	**-0.00024**	**0.00007**	**0.00048**
%Rented Residency	-1.29400	1.07800	0.22998
%Residency with Water	-0.79650	1.16100	0.49257
%Residency with Garbage Collection	3.14900	2.14800	0.14264
%Residency with Pavement	-0.33760	0.86280	0.69558
%Literates	1.37600	1.87800	0.46367
LOSSTRAFO	***13.20000***	***0.89600***	***0.00000***
LOSSAREA	**6.43700**	**1.42100**	**0.00000**
NTRAFO	***-0.02196***	***0.00640***	***0.00061***
CAMPAIGN	***-1.13700***	***1.17600***	***0.00000***

while the significance of *Average Income* indicates the concentration of NTLs in subareas with average income that is higher than the average income of the city. Finally, the significance of *CAMPAIGN* demonstrates that there are more NTLs in targeted subareas of the inspection campaigns.

From Fig. 3, in each year, there are subareas of the central region that are targets for many inspections. This fact explains the increased amount of NTLs found in this area, as can be seen in Fig. 4. Moreover, the central region, in accordance with the result of the GAM, has a high population density and an extensive distribution network.

We must emphasize that it is not possible to state conclusively that variables without statistical significance have no influence on the presence of the NTLs. Statistical significance must be interpreted with discretion upon exploratory analysis of the problem [19]. Moreover, the significance of each variable is modified to the extent that the variables of the analysis are included (or excluded). Significance is also influenced by a sample of the case-control database.

The residual term of splines' smoothing function is significant (p-value 0.073). This means that there may be some residual spatial variation not explained by the variables associated with the GAM.

The proposed methodology is easily implementable in statistical analysis tools. It was applied in our database, and the required computational time was 5 s on a personal computer with an Intel Core TM i7 processor at 2.8-GHz frequency and 4 GB of RAM.

V. Conclusions

In this paper, a method to estimate the vulnerability of NTLs in city subareas presently (via GAM) and in the future (via Markov chain) was shown.

The regional spatial-temporal behavior of the NTLs is represented by the loss state of each area, which is determined from a probability map. This map is obtained via GAM and considers socioeconomic variables from census data and the power grid. The Markov chain is used to model the dynamics of the losses; that is, to check for changes in the loss state by subarea. The method was validated with existing inspection data, and it has been proved feasible to forecast the loss of each subarea with an accuracy rate of 78.2%.

The main contributions of the methodology we propose here are to regard the place where the NTLs occur and estimating the spatial variation of the NTLs by city subarea over time. Through it, three fundamental questions were answered: 1) where the losses are located, that is, which are the critical subareas; 2) what the causes of loss are, that is, which variables create a favorable environment for losses in certain city areas; and 3) finally, where the losses will be located; in other words, what subareas will be critical in the future.

REFERENCES

[1] S. C. Huang, Y. L. Lo, and C. N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2959–2966, Aug. 2013.

[2] W. Han and Y. Xiao, "NFD: A practical scheme to detect non-technical loss fraud in smart grid," in *Proc. IEEE/ICC Int. Conf. Commun. Inf. Syst. Security Symp.*, 2014, pp. 605–609.

[3] R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey, "Wavellet based feature extraction and multiple classifiers for electricity fraud detection," in *Proc. IEEE/Power Energy Soc. Transm. Distrib. Conf. Exhibit.*, 2002, vol. 3, pp. 2251–2256.

[4] J. E. Cabral, E. M. Gontijo, J. O. P. Pinto, and J. R. Filho, "Fraud detection in electrical energy consumers using rough sets," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2004, pp. 3625–3629.

[5] E. W. S. dos. Angelos, O. R. Saavedra, O. A. C. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Trans. Power Del.*, vol. 26, no. 4, pp. 2436–2442, Oct. 2011.

[6] J. I. Guerreiro, C. León, and F. biscarri, "Increasing the efficiency in non-technical losses detection in utility companies," in *Proc. 15th IEEE Mediterranean Electrotech. Conf.*, 2010, pp. 136–141.

[7] P. R. F. M. Bastos, B. A. Souza, and N. Ferreira, "Diagnosis of non-technical energy losses using Bayesian networks," in *Proc. 20th Int. Conf. Elect. Distrib.*, 2009, pp. 1–4.

[8] C. C. O. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcão, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.

[9] A. H. Nizar, Z. Y. Dong, and P. Zhang, "Detection rules for non-technical losses analysis in power utilities," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2008, pp. 1–8.

[10] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Non-technical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.

[11] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[12] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millán, "Variability and trend-based generalized rule induction model to NTL detection in power companies," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 1798–1807, Nov. 2011.

[13] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1284–1285, Apr. 2011.

[14] J. D. Melo, E. M. Carreno, and A. Padilha-Feltrin, "Estimation of a preference map of new consumers for spatial load forecasting simulation methods using a spatial analysis of points," *Int. J. Elect. Power Energy Syst.*, vol. 67, pp. 299–305, 2015.

[15] J. D. Melo, E. M. Carreno, and A. Padilha-Feltrin, "Spatial-temporal simulation to estimate the load demand of battery electric vehicles charging in small residential areas," *J. Control, Autom. Elect. Syst.*, vol. 25, pp. 470–480, 2014.

[16] T. B. Smith, "Electricity theft: A comparative analysis," *Energy Policy*, vol. 32, pp. 2067–2076, 2004.

[17] P. J. Diggle, *Statistical Analysis of Spatial Point Patterns*. London, U.K.: Academic Press, 1983.

[18] R. S. Bivand, J. Pebesma, and V. Gómez-Rubio, *Applied Spatial Data Analysis With R*. New York, USA: Springer, 2008.

[19] C. Bailey and A. C. Gatrell, *Interactive Spatial Data Analysis*. Essex, U.K.: Longman Scientific & Technical, 1995.

[20] T. A. Akers, R. H. Potter, and C. V. Hill, *Epidemiological Criminology: A Public Health Approach to Crime and Violence*. Hoboken, NJ, USA: Wiley, 2013.

[21] C. R. V. Kiffer, C. G. Camargo, S. E. Shikamura, P. J. Ribeiro, T. C. Bailey, A. C. C. Pignatari, and A. M. V. Monteiro, "A spatial approach of the epidemiology of antibiotic use and resistance in community-based studies: The emergence of urban clusters of Escherichia coli quinolone resistance in São Paulo, Brasil," *Int. J. Health Geographics*, pp. 1–10, 2011.

[22] J. E. Kelsall and P. J. Diggle, "Spatial variation in risk of disease: A nonparametric binary regression approach," *Appl. Stat.*, vol. 47, pp. 559–573, 1998.

[23] S. N. Wood, *Generalized Additive Models: An Introduction With R*. Boca Raton, FL, USA: CRC, 2006.

[24] N. A. J. Hastings, *Dynamic Programming: With Management Applications*. London, U.K.: Butterworths, 1973.

[25] R. Larson and B. Faber, *Elementary Statistics: Picturing the World*, 5th ed. Essex, U.K.: Pearson, 2011.

[26] W. N. Venables and D. M. Smith, An Introduction to R: A Programming Environment for Data Analysis and Graphics, Jul. 2014. [Online]. Available: http://cran.r-project.org/doc/manuals/R-intro.pdf

[27] IBGE, (in Portuguese), Household census by subareas. Jan. 2010. [Online]. Available: http://www.ibge.gov.br

[28] SEADE, (in Portuguese), Information of Paulistas municipalities. Jan. 2010. [Online]. Available: http://www.seade.sp.gov.br

Lucas Teles Faria (S'15) received the B.Sc. degree in electrical engineering from the Federal University of Goiás, Goiás, Brazil, and the B.Sc. degree in communication networks from the Federal Institute of Goiás, Goiás, both in 2010, and the M.Sc. degree in electrical engineering from São Paulo State University, São Paulo, Brazil, in 2012, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests include soft computing, fraud detection in electrical systems, and spatial analysis.

Joel David Melo (M'14) received the B.S. degree in electrical engineering from UNMSM, Lima, Peru, in 2006, and the M.S and Ph.D. degrees in electrical engineering from UNESP, Ilha Solteira, Brazil, in 2010 and 2014, respectively.

Currently he is a Postdoctoral Researcher with UNESP, Ilha Solteira. His main interests are power network planning and spatial analysis.

Antonio Padilha-Feltrin (M'89–SM'06) received the B.Sc. degree in electrical engineering from the EFEI and the M.Sc and Ph.D. degrees in electrical engineering from Campinas State University, Campinas, São Paulo, Brazil.

Currently, he is a tenured Professor with the Electrical Engineering Department, Ilha Solteira, São Paulo. From 1995 to 1997, he was Visiting Faculty in the Electrical and Computer Engineering Department of the University of Wisconsin, Madison, WI, USA. His main fields of interest are analysis and control of electrical power systems.