

Pós-Graduação em Engenharia Elétrica

LUCAS TELES DE FARIA

**ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS
NO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA**

Ilha Solteira

2016



UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Ilha Solteira

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LUCAS TELES DE FARIA

**ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS
NO SISTEMA DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA**

Tese apresentada à Faculdade de Engenharia
– UNESP – Câmpus de Ilha Solteira, para
obtenção do título de Doutor em Engenharia
Elétrica.

Área de Conhecimento: Automação.

Orientador: Dr. Antonio Padilha Feltrin

Ilha Solteira

2016

FICHA CATALOGRÁFICA
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

- F224e Faria, Lucas Teles de.
Estimação espaço-temporal das perdas não técnicas no sistema de distribuição de energia elétrica / Lucas Teles de Faria. -- Ilha Solteira: [s.n.], 2016
158 f. : il.
- Tese (doutorado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2016
- Orientador: Antonio Padilha Feltrin
Inclui bibliografia
1. Perdas de energia elétrica. 2. Perdas não técnicas. 3. Perdas comerciais. 4. Análise espacial de dados geográficos. 5. Regressão espacial. 6. Cadeia de Markov.

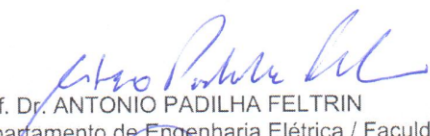
CERTIFICADO DE APROVAÇÃO

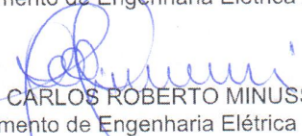
TÍTULO DA TESE: Estimção espaço-temporal das perdas não técnicas no sistema de distribuição de energia elétrica.

AUTOR: LUCAS TELES DE FARIA

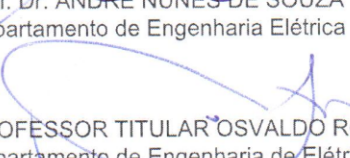
ORIENTADOR: ANTONIO PADILHA FELTRIN

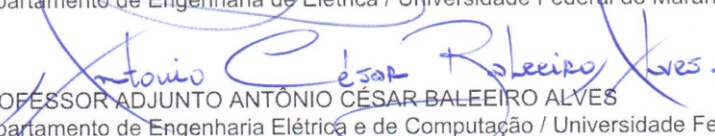
Aprovado como parte das exigências para obtenção do Título de Doutor em ENGENHARIA ELÉTRICA, área: AUTOMAÇÃO, pela Comissão Examinadora:


Prof. Dr. ANTONIO PADILHA FELTRIN
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira


Prof. Dr. CARLOS ROBERTO MINUSSI
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Ilha Solteira


Prof. Dr. ANDRE NUNES DE SOUZA
Departamento de Engenharia Elétrica / Faculdade de Engenharia de Bauru


PROFESSOR TITULAR OSVALDO RONALD SAAVEDRA MENDEZ
Departamento de Engenharia de Elétrica / Universidade Federal do Maranhão


PROFESSOR ADJUNTO ANTÔNIO CÉSAR BALEEIRO ALVES
Departamento de Engenharia Elétrica e de Computação / Universidade Federal de Goiás

Ilha Solteira, 26 de fevereiro de 2016

Dedico este trabalho a Deus, aos meus pais, José Maria (In memoriam) e Cinelândia, aos meus irmãos Luciana e Leandro e a minha querida namorada Andréia.

AGRADECIMENTOS

A Deus pela oportunidade de trabalhar em uma instituição pública de excelência.

A minha família por estar sempre ao meu lado nesta etapa desafiadora da minha vida.

A minha namorada Andréia pela paciência e compreensão devido ao tempo despendido para produção desta tese. Obrigado por tornar meus dias mais felizes!

Ao professor Antonio Padilha Feltrin, por ter aceitado orientar-me no mestrado e no doutorado, pela paciência e dedicação a este trabalho.

Aos parceiros de pesquisa Joel David Trujillo Melo e Tiago Torres.

Aos amigos e colegas do Laboratório de Planejamento de Sistemas de Energia Elétrica (LaP-SEE) pelo companheirismo. Em especial aos amigos e colegas: Marlon Borges, Renzo Vargas, Mahdi Pourakbari, Juan Camilo, Fernando Cerna (“amiguinho”), Leonardo Henrique, Maria Nataly, Ozy Daniel, Roberto Pessoa (“café”), Ruben Ernesto, Silvia Mara, Carlos Sabillon, Darwin Alexis, Diogo Rupolo, Jeferson Back (“alemão”), Jônatas Leite, Katiani Pereira, Marcel Chuma (“braddock”), Patricia Fernanda (“paty girl”), Raiani Piacente, Thays Abreu, Cibele Trinca (“migucha”), Stefani Carolline, Eliane de Souza, Victor Garcia, Alejandra Tabares, Marcelo Semensato, Luciano de Souza (“carçu”), Esau Figueroa, Joel Villavicencio (“joelzinho”), Nelson Taquire (“*el dragón*”), Augusto Medina, John Fredy e tantos outros.

As minhas companheiras de “república” Adriana Vieira (“Adrianinha”) e Fabiana Oliveira (“Fabi”).

A UNESP, ao Departamento de Engenharia Elétrica da FEIS pela excelente estrutura oferecida para o pleno desenvolvimento deste trabalho.

A CAPES pelo indispensável aporte financeiro.

“Uma probabilidade razoável é a única certeza”.

Samuel Gridley Howe.

RESUMO

Neste trabalho o espaço geográfico é incorporado ao estudo das perdas não técnicas. Os trabalhos avaliados em perdas comumente não consideram a localização espacial das mesmas de forma explícita. No entanto, o estudo das características do lugar onde elas ocorrem pode trazer informações imprescindíveis para melhor compreensão do problema. O espaço é incorporado via técnicas de análise espacial de dados geográficos. A saber: análise espacial de padrões de pontos e análise espacial de dados agregados por áreas. A localização das perdas é obtida através de dados de inspeções reais georreferenciados obtidos a partir de uma concessionária de energia elétrica. Os atributos socioeconômicos do censo demográfico e da rede de distribuição de energia do lugar onde ocorrem as perdas são considerados via técnicas de regressões espaciais. São elas: modelo aditivo generalizado (GAM) e regressão geograficamente ponderada (GWR). Esses atributos são as variáveis independentes das regressões espaciais e auxiliam na explicação da disposição das perdas no espaço geográfico do município em estudo. Essas regressões são combinadas com as cadeias de Markov para produção de mapas de probabilidades de perdas. Esses mapas indicam as subáreas do município que são mais vulneráveis às perdas em termos probabilísticos. Por meio deles, estima-se a evolução das perdas não técnicas no espaço geográfico do município ao longo do tempo. Os mapas de probabilidade de perdas são uma ferramenta gráfica, de fácil interpretação e que auxiliam no planejamento de uma série de ações de prevenção e combate às perdas. Este estudo foi realizado em um município de porte médio do interior paulista com aproximadamente 81 mil unidades consumidoras, sendo que os resultados das simulações foram comparados com dados reais de inspeções em campo. A taxa de acerto para estimação das áreas vulneráveis às perdas via modelo aditivo generalizado (GAM) e cadeias e Markov foi superior a 80%.

Palavras-chave – Perdas não técnicas. Perdas comerciais. Análise espacial de dados geográficos. Regressões Espaciais. Sistema de distribuição de energia elétrica.

ABSTRACT

In this work the geographic space is incorporated into the study of non-technical losses. Studies on non-technical losses do not often consider the spatial location of them explicitly. However, the study of the characteristics of the place where they occur can provide essential information to better understanding of the problem. The space is incorporated via spatial analysis techniques of geographical data; to know: spatial analysis of point patterns and spatial analysis of data aggregated by areas. The location of the losses is determined via georeferenced inspections data obtained from an electrical power utility. Socioeconomic attributes of the census and the distribution network of energy of the place where the losses occur are considered using the spatial regressions techniques; namely: generalized additive model (GAM) and geographically weighted regression (GWR). These attributes are the independent variables of spatial regressions and assist in the provision of the explanation of the losses in the geographical space of the city under study. These regressions are combined with Markov chains to produce the loss probability maps. These maps show the city subareas that are more vulnerable to losses in probabilistic terms. Through them, the evolution of non-technical losses in the geographical area of the city over the time is estimated. The loss probability maps are a graphical tool, easy to interpret and to assist in planning a series of actions to prevent and combat to losses. This study was conducted in a medium-sized city of São Paulo with about 81,000 consumer units, and the simulation results were compared with real data obtained in field inspections. The hit rate for the estimation of areas vulnerable to losses via generalized additive model (GAM) and Markov chains surpasses 80%.

Keywords – Non-technical losses. Commercial losses. Spatial analysis of geographical data. Spatial regression models. Electric power distribution system.

LISTA DE FIGURAS

Figura 1 – Distribuição das PNTs por classes consumidoras. _____	24
Figura 2 – Mapa da cidade de Londres com óbitos por cólera e poços de água representados por pontos e cruzeiros, respectivamente. _____	45
Figura 3 – Estimador de intensidade de distribuição de pontos. _____	55
Figura 4 – Fluxograma do modelo para geração do mapa de probabilidade de perdas. _____	62
Figura 5 – Matriz de ponderação espacial de primeira ordem normalizada. _____	66
Figura 6 – Fluxo de produção de conhecimento em análise espacial. _____	75
Figura 7 – Mapa neural bidimensional de Kohonen topologia hexagonal 6×6 – 36 neurônios. _____	80
Figura 8 – Classificação dos modelos para representar a dinâmica de fenômenos espaciais. _____	86
Figura 9 – Mapa da zona sul de uma cidade qualquer com destaque para uma de suas subáreas (em vermelho) que contém três pontos da base de dados casos-controles e quatorze subáreas vizinhas (em verde). _____	89
Figura 10 – Matriz de transição dos estados de perdas por subáreas do município. _____	90
Figura 11 – Histogramas das frequências relativas das variáveis socioeconômicas, técnicas e derivadas, por subáreas (a) Inspeções; (b) Densidade de Carga (kVA/km ²); (c) UCs Irregulares; (d) Renda Média (em R\$); (e) Taxa de Sucesso percentual e (f) %UCs Alugadas. _____	105
Figura 12 – Histogramas de frequências relativas das variáveis socioeconômicas, técnicas e derivadas por subáreas. (a) %Alfabetizados; (b) %UCs com Água Tratada; (c) Média de Moradores; (d) %UCs com Coleta de Lixo; (e) %UCs com Medidor Comum e (f) %UCs com pavimentação. _____	106
Figura 13 – Mapas temáticos com a distribuição espacial das variáveis por subáreas de 2009 a 2012. (a) Inspeções realizadas; (b) UCs Irregulares e (c) Taxa de Sucesso percentual das inspeções. _____	107
Figura 14 – Mapas temáticos com a distribuição espacial por subárea da (a) Densidade de Carga em kVA/km ² ; (b) Renda Média em reais e (c) %UCs Alugadas. _____	108
Figura 15 – Mapas temáticos com a distribuição espacial por subáreas. (a) %Alfabetizadas; (b) Média de Moradores e (c) %UCs com Medidor Comum. _____	109
Figura 16 – Mapas temáticos com a distribuição espacial por subáreas. (a) %UCs com Água Tratada; (b) %UCs com Coleta de Lixo e (c) %UCs com Pavimentação Asfáltica. _____	110
Figura 17 – Mapas temáticos referentes à Taxa de Sucesso anualmente por subárea. (a), (b), (c), (d) Mapas referentes aos anos de 2009, 2010, 2011 e 2012, respectivamente. _____	115
Figura 18 – Mapas temáticos referentes ao número total de Inspeções anualmente por subáreas. (a), (b), (c), (d) Mapas de UCs inspecionadas em 2009, 2010, 2011 e 2012, respectivamente. _____	116
Figura 19 – Mapas temáticos referentes ao número total de UCs Irregulares anualmente por subáreas. (a), (b), (c), (d) Mapas de UCs irregulares em 2009, 2010, 2011 e 2012, respectivamente. _____	117
Figura 20 – Função $G(r)$ ou função de distribuição cumulativa de UCs com PNTs. _____	119
Figura 21 (a), (b), (c) e (d) – Mapas de densidade espacial ponderada e normalizada de UCs irregulares em 2009, 2010, 2011 e 2012, respectivamente. As cruzeiros representam as localizações das inspeções realizadas. _____	120
Figura 22 – Distribuição espacial dos eventos Casos-Controles na região urbana do município sob análise. _____	123

Figura 23 – Mapa de probabilidades para a distribuição espacial das perdas não técnicas.	125
Figura 24 – (a), (b) Estados de perdas atuais (via GAM) e estados de perdas previstos (via cadeias de Markov), respectivamente, por subárea. (c) Validação da metodologia via comparação entre o estado de perda previsto em (b) e o número de PNTs, por subárea, em 2012.	128
Figura 25 – Subáreas vizinhas ou semelhantes após execução do SOM.	132
Figura 26 – Correlação GW entre a variável dependente Taxa de Sucesso de Perdas e as variáveis explicativas Densidade de Carga (a); Renda Média (b) e %UCs Alugadas (c). Em (d) a correlação GW entre as variáveis explicativas Densidade de Carga e Renda Média.	133
Figura 27 – Coeficientes GWR para a variável independentes Densidade de Carga por subáreas.	136
Figura 28 – Coeficientes GWR para a variável independente Renda Média por subáreas.	136
Figura 29 – Coeficientes GWR para a variável independente %UCs Alugadas por subáreas.	137
Figura 30 – Mapas de probabilidade de perdas no futuro via GWR e via cadeias de Markov. (a), (b) Distribuição espacial e histograma de frequência relativa; ambos, com respeito às UCs flagradas em irregularidades em 2012, respectivamente. (c), (d) Distribuição espacial e histograma de frequência relativa; ambos, com respeito às UCs irregulares estimadas ou previstas em 2012, respectivamente. (e), (f) Distribuição espacial e histograma de frequência relativa; ambos, com respeito ao erro absoluto entre as UCs irregulares reais e as UCs irregulares estimadas em 2012, respectivamente.	138
Figura 31 – Mapeamento das amostras de entrada do espaço n -dimensional para o espaço bidimensional da grade neural – $\mathbb{R}^n \Rightarrow \mathbb{R}^2$.	154
Figura 32 – Disposição geométrica dos vetores de pesos dos neurônios do SOM no espaço bidimensional após treinamento da rede neural.	155
Figura 33 – Mapa de contexto 4×4 para uma topologia bidimensional após treinamento do SOM.	157

LISTA DE TABELAS

Tabela 1 – Principais irregularidades em UCs pertencentes à base de clientes da distribuidora de energia. ____	31
Tabela 2 – Síntese das causas mais recorrentes e medidas de prevenção e combate às PNTs no SDEE. _____	33
Tabela 3 – Resumo da eficiência das principais estratégias para localização de UCs com PNTs. _____	36
Tabela 4 – Exemplos de perguntas que podem dar início ao procedimentos de análise espacial em GIS e categorias da análise espacial nas quais as mesmas estão associadas. _____	44
Tabela 5 – Tipos de representações geométricas em análise espacial. _____	47
Tabela 6 – Intervalos do coeficiente de correlação e magnitude da correlação. _____	69
Tabela 7 – Requisitos para modelar a dinâmica ou a variação espaço-temporal de eventos espaciais. _____	84
Tabela 8 – Classificação das principais estruturas temporais para modelagem dinâmica. _____	86
Tabela 9 – Distribuição das inspeções e irregularidades em UCs das principais classes consumidoras. _____	100
Tabela 10 – Diagnóstico da concessionária de energia com respeito às inspeções realizadas ao longo de quatro anos. _____	101
Tabela 11 – Descrições das variáveis analisadas por subáreas. _____	101
Tabela 12 – Perfil socioeconômico do município sob análise e comparativo com o Estado de São Paulo (SP). _____	103
Tabela 13 – Valores das medidas de tendência central e de dispersão para as variáveis da Tabela 11 por subáreas: Média $\bar{X}$; Moda (Mo); Mediana (Md); Desvio Padrão σ ; Coeficiente de Variação (CV); Coeficiente de Assimetria (S) e Curtose (kt). _____	111
Tabela 14 – Matriz simétrica de coeficientes globais de correlação de Spearman (r_s) entre pares de variáveis. _____	112
Tabela 15 – Estimativas, desvios padrão e significâncias estatísticas das variáveis independentes do GAM. _	124
Tabela 16 – Estimativas, desvios padrão e significâncias estatísticas da variáveis do GAM. _____	129
Tabela 17 – Parâmetros para treinamento do mapa auto-organizável de Kohonem – SOM. _____	132
Tabela 18 – Estimativas, desvios padrão e significâncias estatísticas dos coeficientes das variáveis explicativas da regressão linear global. _____	134
Tabela 19 – Resumo dos coeficientes estimados das variáveis independentes do GWR. _____	135
Tabela 20 – Comparativo entre os modelos de regressões espaciais GAM e GWR. _____	140
Tabela 21 – Algumas das funções R^* utilizadas neste estudo. _____	141

LISTA DE SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
CAR	<i>Conditional Autoregressive Models</i>
CSR	<i>Complete Spatial Randomness</i>
GAM	<i>Generalized Additive Models</i>
GIS	<i>Geographic Information System</i>
GW	<i>Geographically Weighted</i>
GWMs	<i>Geographically Weighted Models</i>
GWR	<i>Geographically Weighted Regression</i>
HPP	<i>Homogeneous Poisson Process</i>
IPP	<i>Inhomogeneous Poisson Processes</i>
MGP	Modelos Geograficamente Ponderados
NTL	<i>Non-Technical Losses</i>
PNTs	Perdas Não Técnicas
PTs	Perdas Técnicas
RGP	Regressão Geograficamente Ponderada
SAR	<i>Simultaneous Autoregressive Models</i>
SDEE	Sistema de Distribuição de Energia Elétrica
SIG	Sistema de Informação Geográfica
SOM	<i>Self Organization Maps</i>
UCs	Unidades Consumidoras

LISTA DE SÍMBOLOS

Γ	Autocorrelação espacial.
w_{ij}	Elementos da matriz de ponderação espacial entre as subáreas i e j .
z_i	Variáveis espaciais aleatórias referentes à subárea i .
ξ_{ij}	Medida de correção entre as variáveis que representam as subáreas i e j .
d_{ij}	Distância Euclidiana entre os pontos i e j .
n	Número de pontos ou número de subáreas (ou setores censitários).
S_i	Subárea i .
Z	Processo estocástico.
A	Área genérica do espaço geográfico ou conjunto de subáreas S .
s	Coordenadas geográficas (x,y) .
i, j	Eventos arbitrários localizados no plano bidimensional do espaço geográfico.
r	Distância a partir de um ponto arbitrário.
$E\{. \}$	Operador esperança.
C	Covariância espacial.
λ	Intensidade do processo de pontos.
h	Distância Euclidiana entre dois pontos.
b	Largura de banda ou <i>bandwidth</i> .
$\kappa(.)$	Função de <i>kernel</i> .
$G(.)$	Função de distribuição cumulativa
$K(.)$	Função K de Ripley.
n_1	Número de eventos no conjunto de <i>casos</i> no estudo de <i>casos-controles</i> .
n_0	Número de eventos no conjunto de <i>controles</i> no estudo de <i>casos-controles</i> .
λ_1	Intensidade dos <i>casos</i> no estudo de <i>casos-controles</i> .
λ_0	Intensidade dos <i>controles</i> no estudo de <i>casos-controles</i> .
$\rho(s)$	Risco relativo em s .
$\hat{T}$	Teste estatístico global do risco.
x	Variáveis explicativas ou independentes das regressões espaciais.
y	Variável dependente da regressão espacial.
β	Coefficientes das variáveis independentes x em uma regressão espacial.
$P(s, x)$	Probabilidade em função das coordenadas s e das variáveis explicativas x .
$g(s)$	Função de suavização ou <i>spline</i> das coordenadas espaciais s .

W	Matriz de ponderação espacial, matriz de vizinhança ou matriz de proximidade espacial.
r_s	Coefficiente de correlação de <i>Spearman</i> .
ϵ	Erros aleatórios com distribuição normal, média nula e variância constante.
ρ	Coefficiente espacial autorregressivo utilizado na regressão SAR.
x_k	Vetor das k amostras de entrada para treinamento do SOM.
$\mu(z_i)$	Média GW na subárea i cujo atributo é z_i .
$s(z_i)$	Desvio padrão GW na subárea i cujo atributo é z_i .
$\rho(z_i, y_i)$	Coefficiente de correlação de Pearson GW na subárea i entre os atributos z_i e y_i .
$c(z_i, y_i)$	Covariância GW na subárea i entre os atributos z_i e y_i .
$\Pi_{(t)}$	Estado de transição discreto no tempo t .
P	Matriz de transição de processos de Markov.
$\bar{X}$	Média aritmética simples – estatística descritiva.
Mo	Moda – estatística descritiva.
Md	Mediana – estatística descritiva.
σ	Desvio padrão – estatística descritiva.
CV	Coefficiente de variação – estatística descritiva.
S	Coefficiente de assimetria – estatística descritiva
kt	Curtose – estatística descritiva.

SUMÁRIO

1	ESCOPO DO TRABALHO	20
1.1	CONTEXTO E JUSTIFICATIVA	20
1.1.1	Objetivos	23
1.1.2	Definição do Problema	23
1.1.3	Hipótese da Pesquisa	25
1.1.4	Contribuição Científica Pretendida	26
1.2	ESTRUTURA DA TESE	27
2	PERDAS NÃO TÉCNICAS EM SISTEMAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA	28
2.1	INTRODUÇÃO ÀS PERDAS NÃO TÉCNICAS	28
2.1.1	Perdas Técnicas	30
2.2	MEDIDAS DE COMBATE E PREVENÇÃO ÀS PERDAS NÃO TÉCNICAS	31
2.2.1	Varredura	34
2.2.2	Denúncias	34
2.2.3	Análise dos Dados dos Consumidores	34
2.2.4	Programas computacionais dedicados	35
2.3	REVISÃO BIBLIOGRÁFICA	36
2.4	CONSIDERAÇÕES FINAIS	41
3	ANÁLISE ESPACIAL DE DADOS GEOGRÁFICOS	42
3.1	INTRODUÇÃO À ANÁLISE ESPACIAL DE DADOS GEOGRÁFICOS	43
3.1.1	Tipos de Dados em Análise Espacial	46
3.1.2	Representação Computacional de Dados Geográficos	46
3.2	CONCEITOS BÁSICOS EM ANÁLISE ESPACIAL	47
3.2.1	Dependência Espacial	47
3.2.2	Inferência Estatística em Dados Espaciais	48
3.2.3	Estacionaridade e Isotropia	49
3.2.4	Modelos inferenciais	50
3.3	ANÁLISE ESPACIAL DE PADRÕES DE PONTOS	50
3.3.1	Análise Exploratória de Padrões de Pontos: Caracterização de Distribuições de Pontos	51
3.3.1.1	Aleatoriedade Espacial Completa	52
3.3.2	Análise Estatística de um Processo Espacial de Pontos	53
3.3.2.1	Processos de Poisson	53
3.3.2.2	Estimação da Intensidade: Kernel Não Paramétrico	54

3.3.3	Propriedades de Segunda Ordem	56
3.3.3.1	Função G: Distância para o Evento (ou Vizinho) mais Próximo	57
3.3.3.2	Função K	57
3.3.4	Processo Pontual Marcado: Estudo <i>Casos-Controles</i>	58
3.3.4.1	Estudo de Casos-Controles	58
3.3.4.2	Variação Espacial do Risco Relativo	59
3.3.4.3	Estimador de Regressão Binária	59
3.3.4.4	Teste Global do Risco	59
3.3.4.5	Regressão Binária via Modelo Aditivo Generalizado	61
3.3.4.6	Mapa de Probabilidade de Perdas	61
3.4	ANÁLISE ESPACIAL DE DADOS AGREGADOS POR ÁREAS	63
3.4.1	Distribuição dos Dados em Áreas	65
3.4.2	O Problema da Unidade de Área Modificável e o Problema de Escala	65
3.4.3	Matriz de Ponderação, de Vizinhança ou Matriz de Proximidade Espacial	66
3.4.4	Análise Exploratória	67
3.4.4.1	Coefficiente de Correlação de Spearman r_s	68
3.4.4.2	Técnicas de Corte da Variável para Classificação e Construção de Legendas em Mapas Temáticos	69
3.4.5	Modelos de Regressão	70
3.4.5.1	Modelos de Regressões Espaciais Globais	72
3.4.5.2	Modelos de Regressões Espaciais Locais	73
3.5	CONSIDERAÇÕES FINAIS	74
4	MODELOS GEOGRAFICAMENTE PONDERADOS E MODELAGEM DINÂMICA	76
4.1	INTRODUÇÃO AOS MODELOS GEOGRAFICAMENTE PONDERADOS	76
4.1.1	Definição da Matriz de Ponderação Convencional	77
4.1.2	Redefinição da Matriz de Ponderação para o problema das PNTs: a <i>Vizinhança Semântica</i>	78
4.1.3	Análise Exploratória: Resumo Estatístico Geograficamente Ponderado	81
4.1.4	Regressão Geograficamente Ponderada	82
4.2	INTRODUÇÃO À MODELAGEM DINÂMICA	83
4.3	COMPONENTES PARA MODELAR A DINÂMICA DE FENÔMENOS ESPACIAIS	84
4.3.1	Representação do Espaço	85
4.3.2	Representação do Tempo	85
4.3.3	Modelos de representação	86
4.4	MODELOS EMPÍRICOS	87
4.4.1	Cadeias de Markov de Tempo Discreto	87

4.4.2	Metodologia para Estimaco Espaco-Temporal das Perdas No Tcnicas via GAM e Markov	88
4.4.2.1	Vulnerabilidade s Perdas No Tcnicas no Presente	88
4.4.2.2	Vulnerabilidade s Perdas No Tcnicas no Futuro	90
4.4.2.3	Matriz de Transico	90
4.4.2.4	Determinaco do Estado de Perda Futuro	91
4.4.3	Modelos de Regresso	92
4.4.4	Modelos Logsticos de Difuso	93
4.5	MODELOS DE SISTEMAS	94
4.5.1	Modelos de Simulaco de Ecossistemas	94
4.5.2	Modelos de Simulaco da Dinmica Espacial	95
4.5.2.1	Autmatos Celulares	95
4.5.2.2	Sistemas Multiagentes	95
4.6	CONSIDERAES FINAIS	96
5	ESTIMACO ESPACO-TEMPORAL DAS PERDAS NO TCNICAS	98
5.1	CONSIDERAES INICIAIS	98
5.2	DADOS DE ENTRADA	99
5.3	O PERFIL MUNICIPAL	102
5.4	ANLISE EXPLORATRIA	104
5.4.1	Mapeamento Anual das Inspees, Irregularidades e Taxa de Sucesso de Perdas	114
5.4.2	Propriedades da Informaco Espacial	117
5.4.3	Funo de Distribuico Cumulativa	118
5.4.4	Mapas de Densidade Anual de Perdas No Tcnicas	119
5.5	ESTUDO DE CASO 1: ESTIMACO DA DISTRIBUIO ESPACIAL DAS PERDAS NO TCNICAS via GAM	121
5.5.1	Estudo de <i>Casos-Control</i>es	122
5.5.2	Modelo Aditivo Generalizado (GAM) para Anlise Espacial das PNTs	122
5.5.3	Mapa de Probabilidades de Perdas No Tcnicas	123
5.6	ESTUDO DE CASO 2: ESTIMACO ESPACO-TEMPORAL DAS PERDAS NO TCNICAS VIA GAM E CADEIAS DE MARKOV	126
5.6.1	Estimaco da Variao Espaco-Temporal das Perdas No Tcnicas em um Municpio do Interior Paulista	126
5.6.1.1	Descrio das Variveis utilizadas no Modelo Aditivo Generalizado	126
5.6.1.2	Validaco da Metodologia	127
5.6.1.3	Variveis com Significncia Estatstica	127

5.7	ESTUDO DE CASO 3: ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS VIA GWR E CADEIAS DE MARKOV	130
5.7.1	Construção da Vizinhança Semântica	130
5.7.2	Resumo Estatístico Geograficamente Ponderado	131
5.7.3	Regressão Geograficamente Ponderada	134
5.8	CONSIDERAÇÕES FINAIS	139
6	CONCLUSÕES	142
6.1	PRINCIPAIS CONTRIBUIÇÕES DO TRABALHO	143
6.2	TRABALHOS FUTUROS	145
	APÊNDICE A – MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN	153

1 ESCOPO DO TRABALHO

Este capítulo de apresentação visa fornecer uma introdução abrangente à proposta deste trabalho que consiste na incorporação inédita do espaço geográfico ao estudo do problema das perdas não técnicas (PNTs) ou *non-technical losses* (NTLs); também denominadas perdas comerciais no sistema de distribuição de energia elétrica (SDEE). Ao final deste capítulo, é apresentada a estrutura do presente trabalho.

1.1 CONTEXTO E JUSTIFICATIVA

As PNTs estão presentes na maioria dos SDEE e representam valores elevados para muitas distribuidoras brasileiras. A disseminação das redes inteligentes ou *smart grids* e dos medidores inteligentes contribuirão para uma redução significativa desses valores por eliminação de alguns tipos de perdas. No entanto, tais avanços tecnológicos deverão ser mais lentos e paulatinos nos países em desenvolvimento, justamente naqueles que possuem as maiores perdas. Portanto, a busca por técnicas mais eficientes para avaliação das PNTs faz-se necessária.

O desvio de energia elétrica ou fraude é um problema internacional que prejudica a sociedade e acarreta aumento na tarifa de energia e injustiça social. As ligações irregulares na rede de distribuição de energia representam grande risco para a segurança pública, uma vez que modificam as características da rede de distribuição e podem causar sérios acidentes.

Em Dantas (2006) apresenta-se um resumo dos principais danos à sociedade, ocasionados pela existência das PNTs. A saber:

- ❖ *Insegurança*: em geral, as ligações clandestinas são realizadas sem rigor técnico e sem um estudo prévio da rede elétrica local. As consequências são: acidentes graves, redução do nível de tensão local e aumento das interrupções no fornecimento de energia para clientes regulares que compartilham a mesma rede.
- ❖ *Concorrência desleal*: o furto de energia permite reduzir ilicitamente os custos de atividades comerciais ou industriais, gerando uma concorrência desleal em relação às empresas honestas. Tais empresas são, dessa forma, estimuladas a também aderir a essa prática fraudulenta por uma questão de sobrevivência no mercado.
- ❖ *Aumento tarifário*: as empresas concessionárias de energia elétrica são concessões de serviço público e são regidas por uma política tarifária. Para determinar o percentual de reajuste, a Agência Nacional de Energia Elétrica (ANEEL) considera a

variação de custos das distribuidoras nos últimos doze meses. Na conta de consumo de energia elétrica de cada cliente, há uma parcela referente às PNTs que é medida em valores monetários e não em número de clientes irregulares. Nessa perspectiva, o consumidor regular irá pagar pela energia perdida por meio da elevação da tarifa, o que configura uma injustiça social.

- ❖ *Desperdício de energia*: consumidores fraudadores não pagam a energia elétrica que consomem e, por isso, a tendência é de que eles não tenham hábitos de racionalização, o que ocasiona grande desperdício de energia. São comuns, nesses casos, lâmpadas acesas durante todo o dia ou aparelhos de ar-condicionado ligados ininterruptamente.
- ❖ *Proliferação do furto de energia elétrica*: a impunidade leva à proliferação de criminosos que oferecem uma forma ilícita de economia através da redução ou mesmo da anulação do valor da conta de energia elétrica.
- ❖ *Não arrecadação de impostos*: a arrecadação de vários impostos é reduzida devido às fraudes e às ligações clandestinas. Tais recursos não arrecadados pelo Estado deixam de ser aplicados em benefício da sociedade.
- ❖ *Degradação ambiental*: uma parcela considerável da energia elétrica consumida a nível mundial é gerada em usinas termoeletricas. Tais usinas utilizam combustíveis fósseis cuja queima libera gases que ocasionam inúmeros impactos ambientais tais como: poluição do ar, chuva ácida, danos à camada de ozônio, aquecimento por intensificação do efeito estufa, etc. Logo, o aumento da eficiência energética e redução das perdas elétricas contribuem para tornar o sistema de geração de energia mais eficiente e, conseqüente, redução dos impactos ambientais supracitados.

Ao contrário do estudo das perdas técnicas (PTs), apesar dos inúmeros prejuízos à sociedade supracitados, poucos trabalhos se dedicaram ao estudo das PNTs antes dos anos 2000. Acreditava-se que essas perdas eram ocasionadas majoritariamente devido à desorganização dos processos internos da distribuidora de energia (erros de leituristas ou no cadastro de clientes, por exemplo) (COMETTI; VAREJÃO, 2005).

A partir do início do século XXI, surgiram inúmeros estudos em PNTs. Citam-se Jiang (2002) e mais recentemente Rodrigues, Ramos e Papa (2015). Os estudos avaliados em perdas, em sua maioria, são focados na detecção das mesmas por meio de metodologias que empregam técnicas da área de sistemas inteligentes tais como: redes neurais, lógica *fuzzy*, algoritmos de aprendizado de máquina, etc.

Neste trabalho, pretende-se complementar os trabalhos existentes em perdas. De posse dos dados de inspeções georreferenciados e dos dados cadastrais das unidades consumidoras (UCs) de um município do interior paulista, desenvolve-se uma metodologia que permite a inclusão e o estudo das características socioeconômicas e da rede de distribuição de energia elétrica do lugar onde as perdas ocorrem. Parte-se do pressuposto de que as características do lugar onde as perdas estão localizadas influem nas mesmas e, por isso, devem ser estudadas (SMITH, 2004). O resultado da metodologia será uma estimativa estatístico-probabilística da evolução das perdas no espaço geográfico de um município do interior paulista via *mapas de probabilidades de perdas*. Esses mapas são uma ferramenta gráfica, de fácil interpretação e que indicam as áreas do município com maior (e menor) vulnerabilidade às PNTs. Tais mapas podem ser utilizados como uma ferramenta para auxílio no planejamento de inúmeras ações de combate e prevenção às perdas como, por exemplo, a seleção otimizada das áreas que receberão mais (e menos) inspeções, a realização de campanhas educativas em locais específicos, etc. Dessa forma, os inúmeros prejuízos ocasionados pelas perdas, citados em Dantas (2006), podem ser minorados via um combate mais eficaz às mesmas.

Observa-se que a aplicação da metodologia proposta restringiu-se a um município do interior paulista devido à disponibilidade de dados para essa localidade. No entanto, ela pode ser estendida para outros municípios a partir dos dados de inspeções disponíveis e do estudo do perfil socioeconômico e da rede elétrica dos mesmos.

Com relação à regulamentação, o atual regime de regulação para o setor de distribuição de energia elétrica é o da regulação pelo preço-teto ou *price-cap*, que se caracteriza por ser uma regulação por incentivos. Nesse, o mecanismo que incentiva as empresas a buscarem uma maior eficiência na gestão do seu negócio está na fixação, pelo regulador, de parâmetros gerenciáveis durante um período de tempo (AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL, 2010). No caso das PNTs, o incentivo está na fixação do nível de perdas durante um período específico, e posterior avaliação dos resultados para assegurar que os ganhos de eficiência sejam repassados aos consumidores.

A fixação do nível de perdas deve sinalizar por uma busca permanente de maior eficiência no combate às PNTs pelas empresas de distribuição de energia elétrica. A correta fixação deve considerar fatores não gerenciáveis, específicos de cada área de concessão, muitas vezes relacionados a questões socioeconômicas e, por outro lado, impedir que sejam repassados às tarifas fatores relacionados à ineficiência das distribuidoras no combate às PNTs (ANEEL, 2011; ANEEL, 2013).

1.1.1 Objetivos

Os objetivos gerais deste trabalho consistem em incorporar o espaço geográfico ao estudo das PNTs e em estimar a evolução das mesmas ao longo do espaço geográfico de um município do interior paulista.

Em termos específicos, pretendem-se localizar, espacialmente, as PNTs, explicar o porquê elas estão mais concentradas em determinadas subáreas do município e prever espacialmente, em termos probabilísticos, a localização das mesmas no futuro. Deseja-se responder a três questões fundamentais:

- ❖ Onde estão as PNTs? Quais são as subáreas críticas do município onde se localizam um elevado número de UCs com PNTs ou de UCs irregulares?
- ❖ Por que as PNTs ocorrem? As características socioeconômicas e da rede elétrica das subáreas onde ocorrem as PNTs contribuem ou explicam o aparecimento e o crescimento das mesmas?
- ❖ Onde estarão as PNTs? Predizer quais regiões do município terão mais casos de perdas de maneira que ações de prevenção e combate as mesmas possam ser planejadas com antecedência.

Na Figura 1 apresenta-se a distribuição das PNTs por classes consumidoras. Essas perdas ocorrem tanto em clientes (UCs ativas) quanto em não clientes (UCs desligadas¹ e ligações clandestinas) da concessionária de energia elétrica. Ocorrem perdas também na iluminação pública devido, por exemplo, a sua expansão e a não atualização do cadastro da mesma. Entre os clientes do grupo B, existem as classes: residencial, comercial, industrial, rural, serviço público, poder público e consumo próprio da concessionária de energia elétrica. Neste trabalho, no estudo das PNTs, ênfase maior é dada às UCs ativas pertencentes às classes residencial, comercial e industrial; classes nas quais ocorre a parcela majoritária das PNTs.

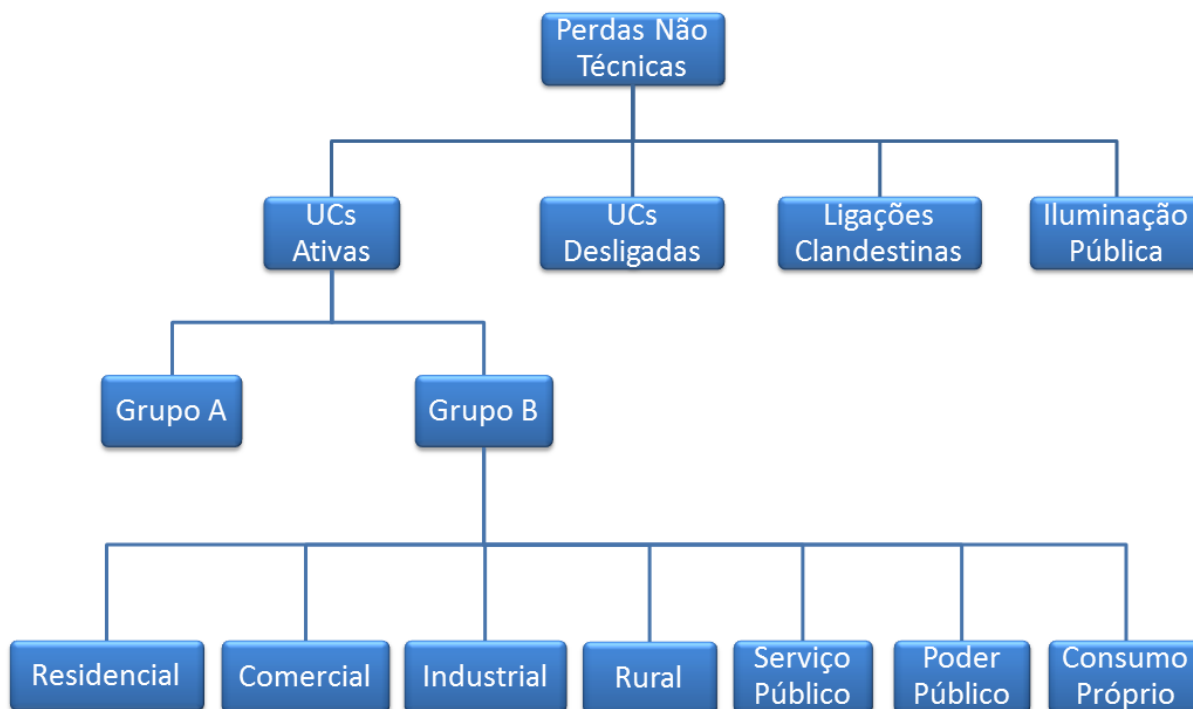
1.1.2 Definição do Problema

As pesquisas avaliadas na área de PNTs comumente direcionam o estudo para a localização ou detecção pontual de UCs irregulares. Para tal finalidade, empregam inúmeras técnicas da área de sistemas inteligentes ou *soft computing* tais como: redes neurais, lógica *fuzzy*, algoritmos de aprendizado de máquina, mineração de dados e de texto, redes Bayesianas, etc. Os clientes suspeitos de fraudes são visitados pelas equipes de inspeção da concessionária de energia que procuram irregularidades em clientes da rede de distribuição. Essas téc-

¹ O termo “UCs desligadas” refere-se àquelas UCs que foram desligadas da rede elétrica pela distribuidora de energia (por inadimplência, por exemplo) e que se autorreligam ilicitamente.

nicas de sistemas inteligentes têm como dados de entrada a base de clientes da concessionária que incluem: dados cadastrais, dados de consumo mensal em kWh, histórico de inspeções, etc.

Figura 1 – Distribuição das PNTs por classes consumidoras.



Fonte: Elaboração do próprio autor.

Os trabalhos avaliados em PNTs seguem a linha de pesquisa supracitada e, por isso, possuem algumas limitações que podem ser complementadas com este trabalho. Em primeiro lugar, em sua maioria, não consideram o lugar onde ocorrem as perdas. No entanto, em Smith (2004), relacionam-se as PNTs com as características socioeconômicas do lugar onde elas ocorrem. Portanto, as características das subáreas onde ocorrem as perdas podem auxiliar no entendimento do problema.

Outra limitação refere-se à grande dificuldade em se detectar com precisão, pontualmente, quais são os clientes suspeitos. Isso é explicado em parte por deficiências da base de dados da concessionária. Há casos em que os dados cadastrais, por exemplo, encontram-se desatualizados e/ou incompletos. Ademais, os dados de consumo mensal em kWh são influenciados por inúmeros fatores de natureza estocástica (climáticos, sazonais, etc.) gerando, em muitos casos, falso-positivos. Essa situação ocorre quando clientes regulares possuem perfis de consumo de energia análogos a clientes fraudadores e, por isso, são rotulados erroneamente como tal por metodologias orientadas à detecção. Falso-positivos implicam em perda finan-

ceira para a concessionária quando investem em equipes de inspeções que perdem tempo ao visitarem UCs regulares.

Para o estudo das PNTs, é preciso extrair informações úteis a partir de uma grande base de dados. No entanto, essa é uma tarefa árdua que requer muito processamento computacional; frequentemente, os dados estão repetidos, incompletos ou inconsistentes. Como não há uma única metodologia para resolver esse problema, é necessário trabalhar em conjunto com a concessionária para se detalhar melhor os objetivos e entender todas as restrições do problema no município em estudo.

Em resumo, o problema das PNTs é altamente complexo e de natureza estocástica. Envolve múltiplas variáveis de natureza diversa; variáveis socioeconômicas e variáveis elétricas como, por exemplo, o número de casas alugadas e a densidade de carga (em kVA/km²) em uma subárea, respectivamente. Nesse contexto, neste trabalho, utiliza-se uma abordagem espacial estatístico-probabilística. Por meio dela, investigam-se as perdas considerando o espaço geográfico no qual as mesmas estão inseridas.

Segundo Ferreira (2014, p. 65), a estatística é, sobretudo, uma ciência da incerteza. Ao utilizá-la talvez seja por que não se está totalmente seguro com relação ao conhecimento do todo. A estatística parte do pressuposto de que não é possível se conhecer, com segurança e exatidão, o todo que se estuda; de que, pelo contrário, somente se sabe de maneira provável; e de que em toda certeza há um resíduo de inexatidão.

Nessa perspectiva, com respeito ao problema das PNTs, dada a escassez de dados, a frequente não integridade total da base de dados da concessionária (histórico de consumo mensal, histórico de inspeções, dados cadastrais, etc.) e a multiplicidade das variáveis envolvidas; são indicativos de que uma abordagem espacial estatístico-probabilística seja adequada para capturar a incerteza relacionada aos dados e a variação espaço-temporal do problema em estudo.

1.1.3 Hipótese da Pesquisa

A hipótese principal da pesquisa é de que as características socioeconômicas e da rede de distribuição de energia do lugar onde ocorrem as PNTs influem diretamente nas mesmas e; por isso, são imprescindíveis à completa compreensão do problema. Por meio dessas características, determina-se a variação espaço-temporal das perdas. Dessa maneira, será possível explicar o porquê elas estão mais concentradas em determinadas regiões do município e detectar subáreas vulneráveis às perdas no presente e no futuro via mapas de probabilidades de perdas.

Ressalta-se que neste trabalho não há associação direta entre fraudes e pobreza. As fraudes ocorrem em todas as classes sociais e em todas as classes consumidoras – residencial, comercial, etc. O uso de variáveis socioeconômicas do censo demográfico do Instituto Brasileiro de Geografia e Estatística, IBGE (2010) objetiva unicamente aproveitar informações úteis, confiáveis e que estão disponíveis publicamente. Observa-se que há poucas informações a respeito de cada UC; logo, o uso da base de dados do IBGE é uma forma de contornar essa limitação de dados e de melhor embasar as análises realizadas.

1.1.4 Contribuição Científica Pretendida

A tese de doutorado de Oliveira (2009) foi pioneira na área de perdas na distribuição do Laboratório de Planejamento de Sistemas de Energia Elétrica (LaPSEE) na qual foi proposto um método simplificado para cálculo das perdas técnicas (PTs).

Na sequência, com a dissertação de mestrado de Faria (2012), apresenta-se um trabalho orientado à detecção das PNTs nas UCs do SDEE. Nesse estudo, construiu-se um sistema inteligente híbrido intercomunicativo que combina um sistema de inferência *fuzzy* (SIF) com as redes neurais *perceptron* multicamadas (PMC) e os mapas auto-organizáveis de Kohonen ou *self organization maps* – SOM (KOHONEN, 2001).

Neste trabalho pretende-se ampliar a discussão. Os trabalhos na área de PNTs comumente implementam metodologias da área de sistemas inteligentes visando a detecção de UCs irregulares ou com PNTs. Além de localizar as subáreas críticas da cidade com altas perdas no presente e no futuro, pretende-se também explicar o porquê das perdas através dos atributos socioeconômicos e da rede elétrica do lugar onde as mesmas ocorrem. Tais atributos podem auxiliar na compreensão do porquê as PNTs estão mais concentradas em determinadas subáreas da cidade.

As PNTs são abordadas, neste estudo, no espaço geográfico da cidade do ponto de vista estatístico-probabilístico. Neste contexto, citam-se alguns trabalhos do LaPSEE que serviram como inspiração tais como: Carreño (2008), Melo (2010), Melo, Carreño e Padilha-Feltrin (2012). Tais trabalhos abordam o problema da previsão espacial de demanda considerando a dinâmica urbana da cidade.

Em suma, considera-se que as características das subáreas onde ocorrem as perdas são importantes e serão abordadas neste trabalho. A partir delas, estima-se a dinâmica ou a variação espaço-temporal das PNTs de maneira a localizá-las, por subáreas do município, no tempo presente e prever quais serão as subáreas críticas, com altas perdas no futuro.

1.2 ESTRUTURA DA TESE

Neste estudo apresenta-se a proposta e o desenvolvimento de uma nova metodologia que incorpora o espaço à análise do problema das PNTs no SDEE. O texto está dividido em seis capítulos. A saber:

- ❖ no capítulo 2 abordam-se os conceitos teóricos que fundamentam o estudo e cuja revisão visa auxiliar na compreensão do leitor acerca do problema tratado e da solução proposta para o mesmo. Apresenta-se também uma revisão acerca da literatura consultada sobre as PNTs;
- ❖ no capítulo 3 abordam-se os aspectos fundamentais da análise espacial, com ênfase à análise espacial de padrões de pontos e análise espacial de dados agregados por áreas. Abordam-se também um modelo aditivo generalizado ou *generalized additive models* (GAM) aplicado em uma base de dados de *casos-controles* que permite a inclusão de variáveis que caracterizam o lugar onde ocorrem as PNTs;
- ❖ no capítulo 4 abordam-se uma família de modelos geograficamente ponderados (MGPs) com ênfase à regressão geograficamente ponderada (RGP) ou *geographically weighted regression* (GWR); modelo utilizado para representar as perdas agregadas por áreas. Na segunda parte do capítulo, abordam-se alguns modelos para representar a variação espaço-temporal de fenômenos espaciais com ênfase às cadeias de Markov que serão utilizadas para representar a variação espaço-temporal das PNTs;
- ❖ no capítulo 5 são apresentadas as simulações realizadas a partir de dados reais de inspeções e de dados públicos do censo demográfico (IBGE, 2010). Esse capítulo é dividido em três partes: na primeira, realiza-se uma análise exploratória nos dados de entrada (variáveis socioeconômicas e da rede elétrica) e uma análise do perfil do município sob estudo; na segunda, produz-se um mapa de probabilidade de perdas via GAM que leva em consideração os atributos de cada subárea do município; em último, realiza-se uma análise da variação espaço-temporal das PNTs via GAM e GWR; ambas regressões espaciais são combinadas individualmente com as cadeias de Markov para produção de mapas de probabilidade de perdas no presente e no futuro;
- ❖ no capítulo 6 apresentam-se as conclusões, principais contribuições da pesquisa e sugestões para trabalhos futuros.

2 PERDAS NÃO TÉCNICAS EM SISTEMAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

As perdas elétricas também denominadas perdas globais de energia na rede de distribuição são definidas como sendo a diferença entre a quantidade de energia comprada e distribuída pela concessionária e a quantidade de energia paga pelos consumidores. As perdas são agrupadas conforme sua origem em duas categorias: perdas técnicas e perdas não técnicas. As primeiras são as perdas de energia por efeito Joule, por exemplo, durante o processo de transporte de energia, ocasionadas pelas resistências internas dos condutores e equipamentos elétricos de transmissão e de distribuição de energia. Tais perdas podem ser reduzidas através de investimentos na construção de novas redes, da correta manutenção e melhoria dos equipamentos elétricos e da melhoria dos processos de distribuição de energia elétrica (COMETTI; VAREJÃO, 2005; OLIVEIRA, 2009).

As PNTs englobam a energia comprada pela concessionária e não faturada a seus consumidores, descontadas as perdas técnicas. Ocorrem por variadas causas e envolvem todos os tipos de perdas com exceção das perdas técnicas. As causas mais comuns são: ligações irregulares, erros de leituristas, defeitos em equipamentos de medição de energia, diversas modalidades de fraudes, etc. Existe uma vasta literatura a respeito das perdas técnicas. Em contrapartida, o estudo das PNTs é mais recente e passou a ser estudado mais fortemente somente nos últimos anos do século XX.

Neste capítulo, introduzem-se as PNTs no SDEE e uma revisão acerca das metodologias avaliadas para tratar o problema.

2.1 INTRODUÇÃO ÀS PERDAS NÃO TÉCNICAS

O problema das PNTs é de cunho estocástico, não estacionário e é composto por múltiplos componentes individuais; ademais, seu estudo varia em função da região sob estudo. O comportamento das PNTs é influenciado por múltiplos fatores temporais, sazonais, aleatórios, socioeconômicos, criminalidade, corrupção, etc.

Em geral, relacionam-se as PNTs, popularmente conhecidas como “gatos”, a áreas carentes, cuja população é mal assistida por serviços públicos essenciais como água tratada, esgoto, etc. Conforme Heppenstall et al. (2012, p. 406), a relação entre o risco de crime e o ambiente social não é trivial. Comunidades carentes, por exemplo, conforme supracitadas, normalmente possuem altas taxa de criminalidade; no entanto, o contrário também tem sido

encontrado. Em outras palavras, o crime de furto de energia ocorre tanto em regiões mais pobres bem como em bairros nobres.

O estudo das PNTs é influenciado pelas características da região na qual as perdas estão inseridas. Em cada cidade, há motivações diferentes para a realização de fraudes. Conforme Oliveira (2013), na Inglaterra, por exemplo, um terço de toda energia elétrica furtada é destinada ao cultivo de maconha. Os produtores utilizam a eletricidade desviada para aquecer estufas que aceleram o crescimento das plantas. O prejuízo pode chegar a R\$ 673 milhões anuais. Adicionalmente, tal prejuízo financeiro pode estar subestimado visto que os dados oficiais consideram somente os furtos de eletricidade comprovados, sendo de 25 mil casos em 2010. Os furtos são executados predominantemente via adulteração do medidor de energia elétrica.

Com respeito aos medidores de energia, existem basicamente dois tipos: os analógicos e os digitais. A maior parte das UCs brasileiras ainda possuem medidores do tipo analógico. Sabe-se que engendrar fraudes em um medidor analógico requer menos conhecimento técnico do que em um medidor digital. Em alguns casos são realizadas fraudes grosseiras como, por exemplo, a colocação de obstáculos físicos no disco do medidor analógico para impedir a rotação do mesmo. Devido a isso e também por razões financeiras (o medidor digital possui menor custo do que o medidor analógico), os medidores analógicos avariados estão sendo paulatinamente substituídos pelas distribuidoras de energia por medidores digitais (COMETTI; VAREJÃO, 2005).

Nesse contexto, para fins de exemplificação, cita-se a distribuidora CPFL energia que iniciou, em 2015, um projeto para modernização do sistema de medição do consumo de clientes da baixa tensão (BT), incluindo os clientes residenciais. Segundo Magnabosco (2015), o projeto, avaliado em R\$ 700 milhões, prevê a instalação de 2 milhões de medidores digitais “inteligentes” em um prazo de cinco a sete anos. Tal iniciativa já foi adotada em todos os 25 mil clientes pertencentes ao grupo A; indústrias e comércios atendidos em alta tensão (AT) e em média tensão (MT). Tais medidores “inteligentes” possibilitam a prestação de novos serviços. A saber: eventuais desligamentos são identificados imediatamente, o restabelecimento do sistema pode ser executado em tempo real, o consumo de energia pode ser aferido remotamente (prescindindo leituristas), a concessionária pode oferecer novos serviços aos consumidores como o acionamento automatizado de aparelhos elétricos, por exemplo, etc.

Na seção subsequente, destacam-se aspectos relevantes das perdas técnicas a fim de melhor elucidação das perdas elétricas no SDEE.

2.1.1 Perdas Técnicas

As perdas técnicas são ocasionadas pelas propriedades físicas dos equipamentos do sistema elétrico de potência (SEP). Referem-se a uma parcela da energia dissipada durante o transporte da energia. A perda nos condutores do sistema elétrico ou perda Joule é um dos exemplos. Ela decorre devido à resistência elétrica do condutor e é proporcional ao quadrado da corrente que o percorre. Todos os equipamentos presentes no sistema elétrico têm perdas técnicas que podem ser medidas e/ou estimadas (PENIN, 2008, p. 09).

Mensurar precisamente as perdas técnicas é um problema complexo devido principalmente à grande quantidade de elementos que constituem o sistema, à grande quantidade de dados necessários, ao caráter aleatório das cargas elétricas e ao contínuo processo de expansão das mesmas.

Os cálculos das perdas técnicas são comumente executados por segmento do sistema elétrico de modo a conferir maior exatidão aos resultados. As perdas técnicas em sistemas elétricos são calculadas comumente em três segmentos: alta, média e baixa tensão. As perdas no segmento de alta tensão são encontradas pela diferença de medição nas subestações. No sistema de distribuição de média tensão, as perdas são calculadas por meio de fluxo de carga. Já a metodologia para o cálculo das perdas técnicas em sistemas de baixa tensão varia entre as concessionárias. Em sua maioria, agregam os componentes da rede de distribuição por tipo (transformador, rede secundária, ramais de serviço e medidores) e realizam os cálculos baseados em curvas de cargas típicas de consumidores, totalmente independente dos dados de faturamento (DANTAS, 2006; OLIVEIRA, 2009).

Além dos segmentos observados, há outras perdas técnicas que são classificadas como perdas diversas. Elas são estimadas como sendo um percentual das perdas técnicas totais calculadas. Tais perdas englobam as perdas nas conexões, efeito corona, fugas nos isolamentos dos equipamentos e nas linhas de transmissão e distribuição, perdas nos bancos de capacitores e de regulares de tensão, etc. Devido à dificuldade de calculá-las de modo preciso, elas recebem um valor estimado entre 5% a 10% do valor total das perdas técnicas nos outros segmentos.

Observa-se o impacto nas perdas técnicas devidos às PNTs. Embora sejam consideradas perdas técnicas tais perdas podem ser consideradas como uma parcela das PNTs devido à própria natureza dos cálculos elétricos das perdas técnicas que são realizados sobre cargas conhecidas. Como as cargas geradoras das PNTs não são conhecidas, desconsiderando-as há uma subdimensionamento das perdas técnicas.

Conhecidas as perdas globais e as perdas técnicas, as PNTs são comumente obtidas pela diferença entre ambas. Logo, erros advindos desses cálculos e estimativas são incorporados no valor estimado para a PNT. Essa é a prática usual do setor elétrico brasileiro (ANEEL, 2013).

2.2 MEDIDAS DE COMBATE E PREVENÇÃO ÀS PERDAS NÃO TÉCNICAS

A detecção de todas as UCs irregulares em uma cidade de médio (ou grande) porte é totalmente inviável. Isso se deve principalmente à grande e crescente quantidade de consumidores, ao custo elevado das inspeções, à informação insuficiente que se tem disponível principalmente com respeito aos clientes do grupo B, à grande diferença na quantidade de clientes regulares e de clientes com algum tipo de irregularidade – segundo Cometti e Varejão (2005), em média, para cada dez clientes regulares há um cliente com alguma irregularidade – e à diversidade de fraudes existentes. Tais fraudes podem compreender desde alterações simples como desvios na instalação elétrica até mecanismos mais elaborados como violação de equipamentos de medição digitais, por exemplo.

Segundo Dantas (2006), para a maioria das concessionárias, fraudes e defeitos na medição são as principais causas de PNTs. Na Tabela 1 são apresentadas as irregularidades mais recorrentes ocasionadas por fraudes e por defeitos na medição e que, por serem de origem externa à concessionária, são de difícil localização.

Tabela 1 – Principais irregularidades em UCs pertencentes à base de clientes da distribuidora de energia.

<i>Irregularidades por Fraudes</i>	<i>Irregularidades por Defeitos na Medição</i>
Ponte no bloco de terminais	Medidor com disco parado
Ligação direta ou autorreligação	Constante de medição errada
Ligação invertida	Consumidor não implantado
Circuito de potencial interrompido	Ligação executada com erro
Desvio aparente antes do medidor	Transformador de corrente (TC) danificado
Desvio embutido na parede	
Medidor avariado	
Medidor com lacre violado	

Fonte: Adaptado de Bastos (2011).

Em Oliveira (2009), relacionam-se diversas ações que visam à redução das PNTs.

A saber:

- ❖ identificação das localidades com maior risco de fraudes e desenvolvimento de campanhas educativas com o apoio dos líderes locais;
- ❖ promoção de estratégias para participação social promovendo ações organizadas nas comunidades com cunho educativo alertando, por exemplo, para o prejuízo que as PNTs causam à sociedade e a divulgação do disque denúncia de fraudes;
- ❖ regularização dos clientes clandestinos;
- ❖ adaptação da rede de distribuição de energia elétrica com a instalação de cabos antifurto, redes compactas ou multiplexadas de média tensão; novos padrões de rede (rede secundária elevada e com cabos concêntricos ou pré-reunidos);
- ❖ implementação de políticas de facilitação de quitação de débitos, cobrança prévia e políticas de cortes;
- ❖ ações no sentido de redução do consumo de energia elétrica. Destacam-se: instalação gratuita de medidores de energia, instalação de aquecedor solar em substituição aos chuveiros elétricos, etc.

Na Tabela 2 apresenta-se uma compilação das causas mais recorrentes que influem diretamente no acréscimo das PNTs nas concessionárias brasileiras. Para cada causa associam-se medidas preventivas que poderiam ser adotadas pelas concessionárias como forma de minimização das PNTs. Observa-se que as causas estão agrupadas em internas e externas às concessionárias. As causas internas tais como erros do leiturista e erros no cadastramento de consumidores são mais facilmente combatidas e sua existência está relacionada ao mau gerenciamento dos procedimentos internos da empresa concessionária. Logo, as PNTs de origem interna são inaceitáveis e devem ser prontamente eliminadas através, por exemplo, das medidas de prevenção e combate da Tabela 2. Em contrapartida, todas as PNTs de origem externa às concessionárias tais como os inúmeros tipos de fraudes da Tabela 1 são mais difíceis de serem localizadas e requerem o desenvolvimento de programas dedicados.

Para combate às perdas de origem externa realizam-se inspeções em UCs previamente selecionadas. As informações colhidas durante as inspeções são fundamentais para quantificação e qualificação das PNTs em uma dada região. Em geral, as inspeções são motivadas por campanhas de combate às PNTs, suspeitas de fraude, denúncias, etc. Dentre as principais estratégias para localização de UCs com PNTs estão: a varredura, denúncias, análise dos consumidores mais suspeitos e a utilização de programas computacionais que implementam ferramentas sofisticadas para detecção de UCs suspeitas. O sucesso de cada uma das estratégias supracitadas é mensurado através da taxa de sucesso ou taxa de acerto das inspeções apresentada em (1):

$$\text{Taxa de Sucesso} = 100 \times \frac{N^{\circ} \text{ UCs Irregulares Encontradas}}{N^{\circ} \text{ UCs Inspeccionadas}} \quad (1)$$

Tabela 2 – Síntese das causas mais recorrentes e medidas de prevenção e combate às PNTs no SDEE.

<i>Distribuidora de Energia</i>	<i>Origem</i>	<i>Medidas de Prevenção e Combate às Perdas</i>
Clientes	<i>Interna:</i> erro do leiturista; defeito ou obsolescência do medidor; erro na ligação; especificação inadequada da medição; engano no cadastro de clientes; funcionários desonestos, etc.	Acompanhamento do processo de ligação e faturamento; atualização cadastral; treinamento do pessoal que especifica o sistema de medição; inspeção nos medidores, etc.
	<i>Externa:</i> fraudes como desvio embutido antes da medição; circuito de potencial interrompido, medidor avariado, ponte nos terminais, etc.	Inspeções; acompanhamento do histórico de consumo; instalação de medição fiscal; desenvolvimento de programas e ferramentas que auxiliem na identificação de desvios de consumo, melhorando a taxa de sucesso das inspeções; lacres no sistema de medição; uso de redes antifurto, etc.
Não Clientes	Ex-clientes desligados que se autorrelegam.	Adequação tarifária; políticas de subsídios; facilidades para negociação de débitos; inspeção das UCs desligadas; lacres no sistema de medição; etc.
	Conexões ilegais: invasões, loteamentos irregulares, etc.	Mapeamento de zonas de risco; realização de políticas sociais em conjunto com a prefeitura, polícia e justiça; opção pela rede antifurto, etc.

Fonte: Adaptado de Bastos (2011).

Por meio da taxa de sucesso, avalia-se a eficácia dos critérios internos das distribuidoras de energia para seleção das UCs que devem ser inspeccionadas.

O resultado da metodologia proposta neste estudo será a construção de mapas de probabilidade de perdas. Por meio desses mapas, localizam-se as subáreas do município vulneráveis às perdas. Esses mapas são uma ferramenta gráfica, de fácil interpretação e úteis em uma série de ações de prevenção e combate às perdas. Destacam-se, por exemplo, o mapeamento das zonas de risco de fraudes (causas externas) melhorando a realização de políticas sociais e o direcionamento das inspeções para as regiões mais vulneráveis às perdas.

Seguem nas seções subsequentes uma síntese a respeito dos principais critérios para localização de UCs irregulares.

2.2.1 Varredura

A estratégia de varredura é uma das primeiras e menos eficiente para localização de clientes com PNTs. Nessa, identificam-se quais são os alimentadores (ou regiões) da rede de distribuição com as maiores PNTs através do cálculo das perdas globais descontadas as perdas técnicas. Em seguida, inspecionam-se todos os consumidores que são supridos por tal alimentador – ou que estão localizados em determinada área suspeita.

Por ter de inspecionar todos os consumidores de uma região pertencentes a um alimentador (ou pertencentes a uma área) com altas perdas; essa estratégia tem alto custo financeiro e consome muito tempo das equipes de inspeções.

As características supracitadas limitam a execução da operação de varredura a regiões nas quais a quantidade de fraudes é suficientemente grande para compensar os altos custos com a operação. Ademais, a operação de varredura é facilmente identificada pelos consumidores que podem mascarar suas fraudes momentos antes da realização da fiscalização.

A taxa de sucesso da operação de varredura é baixa e oscila entre 10% a 12% em média (CABRAL et al., 2004; COMETTI; VAREJÃO, 2005; PERIM; DIAS; COMETTI, 2007).

2.2.2 Denúncias

Em geral, realizam-se inspeções sempre que ocorrem denúncias. Embora a estratégia de denúncias tenha alto índice de acerto, o número de denúncias é muito pequeno frente à quantidade de inspeções que poderiam ser realizadas ao longo do tempo. A taxa de sucesso das denúncias é de 22%, aproximadamente (COMETTI; VAREJÃO, 2005).

2.2.3 Análise dos Dados dos Consumidores

A estratégia de análise dos dados consiste em aplicar regras heurísticas simples para varrer a base de dados e selecionar clientes com histórico de consumo atípico e, devido a isso, tornam-se alvos das inspeções. A análise de dados é uma tarefa árdua que requer muito tempo e esforço por parte dos especialistas em perdas, os quais precisam varrer extensas bases de dados para identificar características suspeitas de cada consumidor.

Uma regra comum é a *regra do consumo nulo* que seleciona clientes com consumo inferior a um valor limite durante um determinado período de tempo com exceção de resi-

dências desocupadas ou de áreas de veraneio que possuem índice de consumo abaixo do padrão (PERIM; DIAS; COMETTI, 2007).

A análise de dados utiliza regras heurísticas baseadas na experiência do especialista. No entanto, uma regra, após ser aplicada sucessivamente ao longo do tempo, tem sua eficácia reduzida. É preciso adaptá-las. Além disso, tais regras cobrem um pequeno número de consumidores. A taxa de sucesso dessa estratégia é moderada. A distribuidora de energia AES Eletropaulo obteve um índice de acerto de 19,8%, em média, nos anos de 2005 a 2007 (FERREIRA, 2008).

A eficácia dessa estratégia depende exclusivamente do conhecimento e experiência do especialista. No entanto, abordagens dependentes exclusivamente do conhecimento do especialista podem impedir que padrões ocultos nos dados sejam encontrados, uma vez que o especialista não tem condições de imaginar todas as possíveis relações e associações existentes em um grande volume de dados. Por isso, recomenda-se a utilização de técnicas de análise dirigidas por computador as quais possibilitam a extração automática (ou semiautomática) de novos conhecimentos a partir de um grande repositório de dados.

2.2.4 Programas computacionais dedicados

A seleção dos consumidores a serem inspecionados é realizada comumente através da busca exaustiva (ou varredura), denúncias ou por meio da análise da base de dados em planilhas eletrônicas por especialistas em perdas. No entanto, se a seleção dos consumidores ou da área a ser inspecionada for feita de maneira automática, através de um programa computacional dedicado, o custo com as inspeções pode ser reduzido, pois elas estariam restritas a pequenas regiões ou pontos bem determinados, aumentando a possibilidade de flagrantes de irregularidades (COMETTI; VAREJÃO, 2005).

Nesse contexto, os programas computacionais para análise das PNTs possuem comumente sistemas classificadores baseados em técnicas de extração automática de conhecimento como a mineração de dados, por exemplo. A partir das fontes de dados disponíveis, tais sistemas computacionais classificadores são capazes de decidir com base nas características atuais de um determinado cliente se o mesmo deve ou não ser inspecionado. Fundamental para essa abordagem é a base de dados com o histórico de inspeções que permite gerar exemplos de clientes regulares e irregulares (fraudadores, por exemplo) para serem utilizados para treinamento e teste do sistema classificador. Aliado a isso, tem-se a aquisição e a incorporação do conhecimento dos especialistas ao sistema desenvolvido. Uma condição fundamental

para o desenvolvimento desses sistemas é torná-lo flexível para que os usuários possam facilmente adaptá-lo a fim de aplicar novas regras heurísticas.

Na Tabela 3 há um resumo da relação custo-benefício, tempo de planejamento, área de abrangência e taxa de sucesso das principais estratégias adotadas pelas empresas concessionárias de energia para combate às PNTs.

Tabela 3 – Resumo da eficiência das principais estratégias para localização de UCs com PNTs.

<i>Estratégia</i>	<i>Retorno Financeiro Custo de Execução</i>	<i>Tempo de Planejamento</i>	<i>Área de Abrangência</i>	<i>Taxa de Sucesso</i>
Varredura	Baixo	Moderado	Por Área	10% a 12%
Denúncias	Alto	Baixo	Pontual	≈ 22%
Análise dos Dados	Moderado	Alto	Pontual	> 20%
Software Dedicado	Alto	Alto	Pontual/Área	> 20%

Fonte: Adaptado de Faria (2012).

2.3 REVISÃO BIBLIOGRÁFICA

As PNTs eram relacionadas recentemente exclusivamente a uma provável desorganização dos processos internos administrativos e de treinamento dos funcionários das distribuidoras de energia. Apenas no fim do século XX, tal problema começou a ser pesquisado mais fortemente, externamente às distribuidoras. Segue uma retrospectiva histórica acerca dos trabalhos avaliados.

Uma parcela importante das PNTs deve-se às fraudes no SDEE. Tais fraudes são um crime tipificado no código penal brasileiro (PENIN, 2008). Nesse contexto, inúmeros trabalhos da área de criminologia levam em consideração o lugar onde ocorrem os crimes. Citam-se, por exemplo, Heppenstall et al. (2012) e Ransford, Kane e Slutkin (2012). Esses são trabalhos na área de criminologia, que estudam os atributos do lugar onde ocorrem crimes comuns tais como: furtos a residência, a pessoa física, a veículos, etc. Tais estudos destacam a importância do lugar onde ocorrem os crimes cujas características podem estimulá-los ou inibi-los.

No entanto, em contraposição aos trabalhos supracitados, os estudos avaliados em PNTs comumente não analisam explicitamente o lugar onde as mesmas ocorrem. Os poucos trabalhos encontrados que analisam alguns atributos locais, do lugar onde estão as perdas, tais como Francisco et al. (2010), fazem-no de forma preliminar, sem o emprego das técnicas de análise espacial de dados geográficos, por exemplo.

Os estudos em perdas são focados, em sua maioria, na detecção ou na localização de UCs irregulares ou com PNTs. Os trabalhos que abordam a detecção das PNTs utilizam comumente técnicas da área de sistemas inteligentes ou *soft computing*. Destaque para as técnicas de mineração de dados e de texto, técnicas estatísticas, lógica *fuzzy*, algoritmos de aprendizado de máquina ou *support vector machine* (SVM) e redes neurais como sendo os métodos com maior êxito nesse campo de pesquisa (GUERRERO; LEÓN; BISCARRI, 2010).

Em Jiang (2002) é apresentada uma abordagem baseada em múltiplos classificadores e coeficientes *wavelets* para identificação de fraudes em clientes cujo consumo de energia foi mensurado em intervalos de quinze minutos. Os resultados das simulações provam que o método é efetivo na identificação de fraudes na distribuição, com 78% de taxa de acerto para os dados de treinamento e 70% para os dados de teste.

Em Cabral et al. (2004) utilizam-se conjuntos rústicos ou *rough sets* como técnica para redução do número de atributos utilizados na indução de um sistema de regras de decisão para detecção de fraudes em consumidores de energia elétrica.

Smith (2004) faz uma abordagem qualitativa de grande relevância na qual se abordam aspectos sociais a nível mundial. Essa referência estimou o furto de eletricidade em 102 países de 1980 até o ano 2000. Ela reconhece a característica complexa e multifacetária do problema das PNTs e o relaciona a aspectos governamentais tais como: instabilidade política, baixa efetividade governamental, altos níveis de corrupção, alta taxa de natalidade, baixo índice de desenvolvimento humano (IDH), etc.

Cometti e Varejão (2005) desenvolveram um sistema para identificação de clientes com PNTs e efetuaram uma análise comparativa entre diversas técnicas da área de sistemas inteligentes. Tais técnicas foram comparadas em termos da especificidade e da confiabilidade negativa, dando maior ênfase à primeira. Seguem as técnicas empregadas em ordem decrescente de desempenho para detecção das PNTs: *Naïve Bayes*, redes Bayesianas, *k-Nearest Neighbor* (KNN), algoritmo C4.5 e redes neurais.

Rauber et al. (2005) mostraram que a extração de novas características, a partir da série temporal de consumo mensal, melhora o desempenho do classificador que determinará a presença ou a ausência de PNTs. A partir dessa constatação, os autores realizaram a análise comparativa entre três metodologias para extração de atributos: coeficientes de Fourier, coeficientes *wavelets* e regressão polinomial. O primeiro método de extração gerou as características mais relevantes, seguido pelos coeficientes *wavelets* e pela regressão polinomial.

Na dissertação de Dantas (2006) foi desenvolvido um *software* para quantificação e estratificação das PNTs por classe e por faixa de consumo no município de Salvador, Bahia.

As UCs foram agrupadas em quatro categorias: irregular com fraude, irregular com defeito, irregular sem perda e normal. Esse trabalho possui embasamento estatístico e baseia-se na amostragem aleatória de dados levantados em campo.

Perim, Dias e Cometti (2007) desenvolveram um sistema baseado em conhecimento (SBC) composto por regras independentes escritas manualmente na linguagem XML baseadas no conhecimento do especialista. Os dados de entrada avaliados são alguns atributos estatísticos como a média e o consumo nulo que são extraídos a partir do histórico de consumo mensal de cada consumidor em kWh.

Na dissertação de Ferreira (2008) foi avaliada uma combinação entre diferentes ferramentas de classificação e bases de dados para melhorar o índice de acerto das inspeções. As bases de dados são compostas pelo histórico de consumo mensal e por dados cadastrais dos clientes. Para tal, utilizam-se ferramentas de aprendizado de máquina tais como: redes neurais, o algoritmo indutor de árvore de decisão C4.5, aprendizado de máquina e *Naïve Bayes*. O trabalho efetua uma análise comparativa entre esses quatro classificadores a fim de detectar aquele que apresenta melhor poder preditivo em termos dos critérios de confiabilidade negativa e de especificidade. O algoritmo C4.5 obteve os melhores resultados.

Nizar, Dong e Zhang (2008) buscam detectar anormalidades em UCs que podem ser devidas às PNTs através da extração de regras em dados cadastrais. Utilizam técnicas de mineração de dados, detecção de *clusters* e técnicas de classificação para construção de um *framework* para detecção das perdas.

Em Nizar, Dong e Wang (2008) foi implementada uma técnica de mineração de dados via *extreme learning machine* (ELM) para melhorar o processo de detecção de UCs com PNTs. Tal técnica utiliza a curva de carga dos consumidores para detectar comportamentos de consumo anormais ou suspeitos. Demonstrou-se, via inúmeros resultados, a superioridade do ELM frente a outras técnicas de classificação como SVM.

Na tese de Bastos (2011) foi feito um diagnóstico ou prospecção das PNTs permitindo identificar suas diversas origens e quantificar as parcelas relacionando-as as causas da mesma entre os consumidores e classes de consumo em várias regiões. Utiliza-se para tal as redes Bayesianas.

No trabalho de Guerrero, León e Biscarri (2010), objetiva-se explicar o comportamento muitas vezes anômalo e suspeito da curva de consumo mensal como sendo consequência das características de consumo do cliente como, por exemplo, de sua atividade econômica. Os dados de entrada da metodologia são as informações textuais e em linguagem natural contidas na base de dados da concessionária de energia. São informações sobre a do-

cumentação, comentários dos inspetores e informações adicionais sobre a instalação elétrica dos consumidores. A metodologia é uma combinação das técnicas de mineração de texto, redes neurais, técnicas estatísticas e experiência de especialistas a fim de extrair conhecimento a partir de informação não estruturada. Dessa forma, pretende-se aumentar a eficiência das campanhas contra perdas através da redução de falso-positivos, isto é, redução de clientes normais ou regulares que são erroneamente rotulados como irregulares pela metodologia empregada.

Em Nagi et al. (2010) foi implementada a mineração de dados via *support vector machine* (SVM) para selecionar mais eficientemente clientes suspeitos de fraudes. O método extrai informações do histórico de consumo mensal. Os dados de entrada para o SVM são informações da curva de carga dos consumidores e informações adicionais relacionadas às PNTs. O método foi testado em uma distribuidora da Malásia e elevou a taxa de sucesso das inspeções de 3% para 60%.

Em Nagi et al. (2011) continua-se o estudo realizado em Nagi et al. (2010) no qual implementa-se uma técnica de mineração de dados combinada com um *fuzzy inference system* (FIS) visando minorar os custos com inspeções. O FIS permite incorporar uma etapa de pós-processamento na metodologia e a inclusão do conhecimento de especialistas em perdas via regras *fuzzy* do tipo *SE-ENTÃO*. Com tal melhoria, houve um acréscimo na taxa de sucesso das inspeções de 60% para 72%.

Em León et al. (2011) implementa-se um *framework* para detecção das PNTs o qual é dividido em quatro seções: seleção de dados, pré-processamento, descrição e mineração de dados. Apresentam-se dois estimadores estatísticos inovadores para considerar a variabilidade e a análise de tendências no histórico de consumo. Tal estudo apresenta também um modelo preditivo baseado em *generalized rule induction* (GRI). Por meio da análise preditiva descobrem-se regras em uma base de dados. Esse modelo é combinado com um outro método classificador de árvore de busca binária. Houve acréscimo na taxa de identificação de fraudes de 7% para 20%.

Em Ramos et al. (2011) implementa-se uma técnica de reconhecimento de padrão via *optimum path forest* (OPF) para aperfeiçoar a detecção de UCs irregulares. Destaca-se a superioridade e maior simplicidade desta técnica frente a outras da área de sistemas inteligentes que são igualmente utilizadas para detecção de PNTs. A saber: *artificial neural networks* (ANN), SVM, *genetic algorithms* (GAs), *artificial neural networks multilayer perceptron* (ANN-MLP), *self-organizing maps* (SOM). Ao longo do estudo, apresentam-se inúmeros re-

sultados que demonstram a maior eficiência da técnica OPF desenvolvida pelos autores do estudo para detecção das perdas.

Em Angelos et al. (2011) apresenta-se uma técnica computacional para detecção e identificação de anormalidades no consumo de UCs da rede de distribuição. Utiliza-se uma técnica de clusterização *fuzzy C-means* (FCM) para encontrar UCs com perfis de consumo similares. Após esta etapa, utiliza-se um sistema de classificação *fuzzy* utilizando matriz de pertinência *fuzzy* e a distância Euclidiana entre os centros dos *clusters*. Essa distância é normalizada e corresponde a um índice que possui valor elevado para potenciais fraudadores. A proposta foi validada com dados reais mostrando bom desempenho para consumidores residenciais e comerciais. A aplicação do método é totalmente viável. Ele requer o uso de apenas cinco variáveis de entrada disponíveis nas concessionárias de energia elétrica.

De acordo com Ramos et al. (2012), no entanto, o foco não é na extração de características, mas na seleção daquelas que são mais representativas, que possuem maior poder preditivo e que serão as entradas das técnicas de reconhecimento de padrões também denominados sistemas classificadores. Para tal, os autores empregam três algoritmos evolutivos: *particle swarm optimization* (PSO), *gravitational search algorithm* (GSA) e *harmony search* (HS). Os testes a partir de oito características de clientes comerciais e industriais demonstram o melhor desempenho do sistema classificador OPF na detecção de clientes suspeitos quando antes é realizada a seleção das características mais relevantes. O método PSO-OPF superou HS-OPF e GSA-OPF, sendo o HS-OPF o algoritmo com menor tempo computacional.

Em Huang, Lo e Lu (2013) emprega-se estimação de estados para se obter a distribuição de carga nos transformadores da distribuição. A partir dessa distribuição, estimam-se alimentadores com mais PNTs.

Em Porras et al. (2015) desenvolve-se um protótipo de um *software* para melhorar o processo de detecção das PNTs na região de Quindío, Colômbia. Nesse *software*, facilita-se a visualização das PNTs por transformadores, alimentadores, subestações, etc. Utilizam-se regras específicas da região sob análise. Por meio da visualização, é possível realizar análises preliminares a partir da visualização do lugar onde ocorrem as perdas.

Em Rodrigues, Ramos e Papa (2015) busca-se detectar as PNTs no contexto das *smart grids*. Para tal, utiliza a técnica de otimização metaheurística *black hole algorithm* (BHA) para caracterização das UCs irregulares. Os resultados demonstram a importância de selecionar as características mais relevantes das UCs irregulares e comprovam a superioridade da técnica BHA em relação às técnicas HS, *genetic algorithm* (GA), PSO e OPF no reconhecimento do perfil de clientes irregulares.

Por fim, em Trevizan et al. (2015), propõe-se um método para identificação das PNTs envolvendo o algoritmo *optimum-path forest* (OPF) e o estimador de estado da distribuição ou *distribution state estimator* (DSE) em consumidores residenciais. Os consumidores são classificados em suspeito e em não suspeitos via OPF. Na fase de refinamento utiliza-se o DSE. O método foi testado em um sistema-teste do *Institute of Electrical and Electronics Engineers* (IEEE) de 123 barras considerando um grupo de consumidores em cada barra. No melhor resultado, a taxa de sucesso das inspeções foi superior a 72,43%.

2.4 CONSIDERAÇÕES FINAIS

Neste capítulo fez-se uma explanação abrangente a respeito da natureza do problema das PNTs no SDEE. Destacaram-se as principais formas de prevenção e combate às perdas e uma retrospectiva histórica a respeito das referências relevantes avaliadas nesse tema.

No capítulo subsequente abordam-se aspectos fundamentais da análise espacial de dados geográficos com ênfase à análise espacial de pontos e análise espacial de dados agregados por áreas. Tais técnicas são fundamentais para incorporar o espaço geográfico de maneira original ao problema das PNTs.

3 ANÁLISE ESPACIAL DE DADOS GEOGRÁFICOS

Neste trabalho, o espaço geográfico é incorporado ao problema das PNTs através das técnicas de análise espacial de dados geográficos ou estatística espacial. Por meio dessas técnicas, localizam-se as PNTs ou as UCs irregulares no espaço geográfico da cidade e relaciona-se a localização das mesmas aos atributos do lugar onde elas estão.

Neste capítulo, objetiva-se abordar, de forma concisa algumas técnicas de análise espacial de dados geográficos de padrões de pontos e de dados agregados por áreas.

A seguir, apresenta-se um excerto de Ferreira (2014, p. 65) a respeito da estatística espacial e da incerteza relacionada aos dados geográficos.

A estatística é, sobretudo, uma ciência da incerteza. Ao utilizá-la talvez seja porque não se está totalmente seguro com relação ao conhecimento do todo. A estatística parte de um pressuposto surpreendente: de que não é possível se conhecer com segurança e exatidão o todo que se estuda; de que, pelo contrário, somente se sabe de maneira provável; e de que em toda certeza há um resíduo de inexatidão. Tal nível de desordem dos dados (principalmente nos dados geográficos) chama-se *entropia*². Nada tão entrópico como mapas, distribuições espaciais, o território e as regiões. Da entropia deriva a incerteza. Incerteza de querer saber sobre o oceano, mas tendo à mão apenas alguns copos de água; de conhecer o deserto, mas a partir de alguns punhados de suas areias; de compreender a cidade por meio de uma amostra de suas casas; e porque não de entender uma região tendo como evidência parte de seus municípios. Esses exemplos carregam consigo certo grau de incerteza. Se na maioria das vezes restam amostras e fragmentos da totalidade quase infinita do espaço, é com eles que devemos seguir em frente.

A descrição supracitada está em conformidade com as características do problema das PNTs. As bases de dados das distribuidoras não são totalmente confiáveis; comumente, possuem registros incompletos de clientes, desatualizados ou cadastrados erroneamente. No entanto, conforme Ferreira (2014), “se na maioria das vezes restam amostras e fragmentos da totalidade (...) é com eles que devemos seguir em frente”. Portanto, *a priori*, é viável tratar as PNTs com a técnica estatística espacial de dados geográficos.

Nesse contexto, introduz-se, neste capítulo, a análise espacial de dados geográficos com foco na análise espacial de pontos e de áreas. Tal ferramenta é comumente utilizada em várias áreas do conhecimento como: epidemiologia, criminologia, geologia, agronomia, ecologia, etc. Busca-se levar em conta o lugar onde ocorre o fenômeno sob estudo. Neste tra-

² A *entropia* é um conceito originalmente da termodinâmica que mensura a desordem das partículas de um sistema físico. Quanto maior a desordem de um sistema, maior será sua entropia. Neste texto, a entropia refere-se à desordem e a imprecisão sempre presente (em maior ou menor grau) nos dados provenientes de bases de dados georreferenciadas.

balho, enfatiza-se a análise espacial de pontos e de áreas as quais serão utilizadas para incorporar o espaço ao problema das PNTs.

3.1 INTRODUÇÃO À ANÁLISE ESPACIAL DE DADOS GEOGRÁFICOS

A compreensão da distribuição espacial de fenômenos ocorridos no espaço geográfico³ é fundamental para compreensão de inúmeros problemas oriundos de diversas áreas do conhecimento tais como: saúde, geologia, agronomia, etc. Esses estudos são viabilizados e facilitados devido ao surgimento e popularização dos sistemas de informações geográficas (SIGs) ou *geographic information system* (GIS). Esses são programas computacionais com interface amigável e de baixo custo – alguns são livres, portanto, sem custo – que permitem a visualização espacial de variáveis tais como: população de doentes por região, renda, taxa de furto de veículos, etc. A partir de um banco de dados e de uma base geográfica (um mapa da região sob análise), o GIS permite a visualização do padrão espacial do fenômeno.

Além da visualização do padrão espacial do fenômeno, podem-se analisar os padrões existentes via considerações objetivas e mensuráveis. Conforme Druck et al. (2004), a ênfase da análise espacial é mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo de forma explícita. Em outras palavras, ela permite a incorporação do espaço à análise que se deseja realizar. A análise espacial é composta por um conjunto de ferramentas – muitas delas já implementadas nos GIS – para explorar e modelar processos que se expressam através de uma distribuição no espaço. Tais processos são denominados fenômenos geográficos. Seguem alguns problemas que são comumente estudados com o uso da análise espacial.

- ❖ Epidemiologistas coletam dados sobre ocorrência de doenças. A distribuição dos casos da doença estão dispostos em algum padrão no espaço? Existe associação com alguma fonte de contágio, algum local contaminado? Varia no tempo?
- ❖ Deseja-se investigar se há concentração espacial na distribuição no furto de veículos. Tais furtos, que ocorrem predominantemente em determinadas áreas da cidade, estão correlacionados com as características socioeconômicas das mesmas?
- ❖ Geólogos desejam estimar a extensão de um depósito mineral em uma região a partir de algumas amostras. É viável estimar a distribuição do mineral na região a partir dessas amostras?

³ Fenômenos que ocorrem no espaço geográfico são aqueles representados por coordenadas geográficas (x,y) .

- ❖ As UCs com fraudes estão dispostas em aglomerados? Existe alguma associação entre as fraudes e as condições socioeconômicas do lugar onde as mesmas ocorrem? Variam com o tempo?

O processo de modelagem em análise espacial é precedido por uma fase de *análise exploratória* na qual se incluem a apresentação visual dos dados em forma de gráficos e mapas e a identificação de padrões de dependência espacial do fenômeno sob estudo, ou seja, avaliar se as características do lugar onde ocorre o fenômeno exerce influência sobre o mesmo (DRUCK et al., 2004).

Muitas das etapas de análise espacial realizadas em um GIS são iniciadas a partir de perguntas espaciais. As perguntas espaciais podem ser agrupadas em categorias conforme as técnicas de análise espacial envolvidas na construção dessas perguntas. As perguntas espaciais da Tabela 4 foram construídas em um contexto de investigação adaptado à área da epidemiologia espacial. No entanto, podem ser ajustadas a outros contextos de pesquisa geográfica, substituindo a palavra tema de cada pergunta – os *casos de dengue*, por exemplo.

Tabela 4 – Exemplos de perguntas que podem dar início ao procedimentos de análise espacial em GIS e categorias da análise espacial nas quais as mesmas estão associadas.

<i>Pergunta Espacial</i>	<i>Categorias de Análise Espacial</i>			
	<i>Localização</i>	<i>Distribuição</i>	<i>Associação</i>	<i>Mudança</i>
Onde ocorrem casos de dengue?	X			
Existe regularidade na distribuição espacial dos casos de dengue		X		
Por que o padrão espacial da distribuição dos casos de dengue exibe regularidade?			X	
Quais variáveis socioeconômicas estão associadas espacialmente aos casos de dengue em São Paulo?			X	
Os casos de dengue ocorrem agrupados em regiões ou em <i>clusters</i> de municípios?	X	X		
Os casos de dengue sempre ocorreram neste mesmo lugar?				X

Fonte: Adaptado de Ferreira (2014, p. 60).

Em Druck et al. (2004) e Fine et al. (2013), relata-se o trabalho pioneiro de John Snow que foi o primeiro a incorporar intuitivamente o espaço às suas análises. Em 1854, a cidade de Londres passava por mais uma epidemia de cólera. Pouco se sabia dos mecanismos causais da doença. Conforme se mostra na Figura 2, os locais dos óbitos por cólera e dos poços foram modelados como processos pontuais e identificados no mapa da cidade por pontos e cruzes, respectivamente. Observa-se visualmente uma maior concentração dos óbitos nos arredores de um poço localizado na *Broad Street* (região central) cuja água estava contaminada pela bactéria transmissora da doença. Portanto, os óbitos estavam distribuídos espacialmente em aglomerado ou *cluster* ao redor do poço contaminado. Esta é uma situação clara na qual a relação espacial dos dados contribuiu significativamente para melhor compreensão do fenômeno sob estudo.

Figura 2 – Mapa da cidade de Londres com óbitos por cólera e poços de água representados por pontos e cruzes, respectivamente.



Fonte: Druck et al. (2004).

3.1.1 Tipos de Dados em Análise Espacial

Os problemas em análise espacial são caracterizados por três tipos de dados.

- ❖ *Padrões pontuais*: fenômenos expressos por ocorrências identificadas como pontos localizados no espaço – eventos pontuais. Exemplos: localização de crimes, ocorrência de doenças, localização de espécies vegetais, etc.
- ❖ *Áreas com taxas agregadas*: dados associados a levantamentos populacionais como censos demográficos e estatísticas de saúde que se referem a indivíduos localizados em pontos específicos no espaço. No entanto, por razões de confidencialidade, tais dados são agregados em unidades de análise, delimitadas por polígonos fechados tais como: setores censitários⁴, município, etc.
- ❖ *Superfícies contínuas*: estimadas a partir de um conjunto de amostras de campo – regular ou irregularmente distribuídas. Esses dados são comumente resultantes de levantamentos de recursos naturais representados, por exemplo, em mapas geológicos, topográficos e ecológicos.

A representação dos dados em superfícies contínuas pressupõe que os mesmos podem ocorrer em qualquer parte do espaço geográfico. Esse não é o caso das PNTs que podem surgir apenas onde existem UCs. Portanto, neste trabalho, os dados de perdas são caracterizados via análise espacial em padrões pontuais e em áreas com taxas agregadas.

3.1.2 Representação Computacional de Dados Geográficos

O GIS é um *software* que realiza o tratamento computacional dos dados geográficos e armazena a geometria e os atributos dos dados georreferenciados que estão localizados na superfície terrestre e que são representados em uma projeção cartográfica (DRUCK et al., 2004).

Os principais componentes que constituem um GIS se relacionam de forma hierárquica. No nível mais próximo do usuário, a interface define como o sistema é operado e controlado. No nível intermediário, o GIS possui ferramentas de processamentos de dados espaciais (entrada, edição, análise, visualização, saída). Internamente um banco de dados geográficos armazena e recupera dados espaciais.

O modelo de sistema de gerenciamento de banco de dados (SGBD) mais comum é georrelacional. Nesse, os atributos descritivos de cada objeto são organizados em tabelas cujas linhas e colunas correspondem aos dados e aos nomes dos atributos, respectivamente. Ca-

⁴ O setor censitário é a menor unidade territorial das cidades, um polígono com limites físicos identificáveis e com dimensão adequada para levantamento das informações do censo demográfico (IBGE, 2010).

da linha da tabela contém valores associados a um evento ou objeto geográfico. Tais objetos possuem um identificador único que possibilita uma ligação lógica entre os atributos dos objetos e suas representações geométricas.

Na Tabela 5 apresentam-se os tipos de representações geométricas mais comuns em análise espacial que variam conforme os tipos de dados utilizados.

Tabela 5 – Tipos de representações geométricas em análise espacial.

<i>Representação</i>	<i>Descrição</i>
Pontos 2D	Par ordenado (x,y) de coordenadas espaciais. Um ponto indica um local de ocorrência de um evento.
Polígonos	Conjunto de pares ordenados $\{(x,y)\}$ de coordenadas espaciais, sendo o último ponto idêntico ao primeiro, formando uma região fechada no plano.
Amostras	Pares ordenados $\{(x,y,z)\}$ onde z indica o valor do fenômeno estudado na localização indicada pelas coordenadas espaciais (x,y) .
Grade Regular	Matriz onde cada elemento está associado a um valor numérico. Em geral tal matriz representa uma região da superfície terrestre. Possui comumente espaçamentos regulares nas direções horizontal e vertical.
Imagem	Matriz onde cada elemento associa-se a um valor inteiro (de 0 a 255 no padrão RGB ⁵). Os valores numéricos da grade são escalonados para o intervalo de apresentação da imagem.

Fonte: Adaptado de Druck et al. (2004).

3.2 CONCEITOS BÁSICOS EM ANÁLISE ESPACIAL

3.2.1 Dependência Espacial

A noção de dependência espacial parte de Tobler (1979) que enuncia a primeira lei da geografia, segundo a qual, todas as coisas são parecidas; no entanto, coisas mais próximas se parecem mais do que coisas mais distantes. Nesse contexto, Cressie (1991) afirma que a dependência espacial está presente em todas as direções e torna-se mais fraca à medida que se aumenta a dispersão na localização dos dados.

Em termos gerais, diz-se que a maior parte dos fenômenos; quer sejam naturais, quer sejam sociais, apresentam uma relação que depende da distância. Por exemplo, se há um foco de criminalidade em uma região da cidade, é mais provável que locais próximos também tenham altos índices de violência.

⁵ Sistema de cores aditivas denominada escala RGB – *red, green, blue*.

No caso das PNTs, nem sempre a lei de Tobler é obedecida. Se um cliente pratica fraude não necessariamente seu vizinho também praticará. Logo, não há evidência de que a presença de uma UCs com fraude estimule as UCs vizinhas na prática de igual delito. Uma subárea distante, por exemplo, pode exercer maior influência do que uma subárea fronteira. Portanto, a abordagem tradicional da análise espacial de considerar subáreas vizinhas como sendo apenas subáreas fronteiriças será adaptada a fim de melhor representar o problema das PNTs.

A dependência espacial é representada computacionalmente por meio da autocorrelação espacial. A correlação é uma medida estatística com o intuito de mensurar o relacionamento entre duas variáveis aleatórias. O prefixo “auto” indica que a correlação é realizada com a mesma variável aleatória medida em locais distintos do espaço. Verifica-se como varia a dependência espacial a partir da comparação dos valores de uma amostra e de seus vizinhos. Existem vários indicadores para mensurar a autocorrelação espacial como o índice de Moran (para análise espacial de áreas agregadas) e o variograma (para análise espacial de superfícies contínuas), por exemplo. Todos são casos particulares de uma estatística de produtos cruzados conforme (2).

$$\Gamma = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(d) \xi_{ij} \quad (2)$$

Dada uma distância d , w_{ij} são os elementos da matriz de contiguidade espacial entre as variáveis z_i e z_j informando, por exemplo, se elas estão separadas por uma distância inferior a d . ξ_{ij} fornece uma medida de correção entre estas variáveis aleatórias. Os valores obtidos são comparados com os valores que seriam produzidos se não houvesse associação espacial entre as variáveis. Valores significativos dos índices de autocorrelação espacial indicam dependência espacial e, por isso, os modelos inferenciais, nesses casos, devem necessariamente considerar o espaço explicitamente.

3.2.2 Inferência Estatística em Dados Espaciais

Uma consequência da dependência espacial é que as inferências estatísticas não serão tão eficientes quanto no caso de amostras independentes de mesmo tamanho. A dependência espacial leva a uma perda do poder explicativo. Isso se reflete em variâncias maiores para as estimativas, em níveis menores de significância em testes de hipóteses e a um ajuste

pior para modelos estimados frente a dados de mesma dimensão que exibam independência (DRUCK et al., 2004).

A perspectiva mais apropriada é considerar os dados espaciais, não como um conjunto de amostras independentes, mas como uma única realização de um processo estocástico. Na visão amostral tradicional, cada observação traz uma informação; em um processo estocástico, todas as observações são utilizadas de forma conjunta para descrever o padrão espacial do fenômeno estudado.

3.2.3 Estacionaridade e Isotropia

Os principais conceitos estatísticos que definem a estrutura espacial dos dados relacionam-se aos efeitos de primeira e de segunda ordem. O efeito de primeira ordem é o valor esperado, ou seja, a média global do processo no espaço. Efeito de segunda ordem corresponde à covariância entre as áreas S_i e S_j .

O processo é dito *estacionário* se os efeitos de primeira e de segunda ordem são constantes na região sob análise, isto é, não há tendência.

Um processo é *isotrópico* se, além de estacionário, a covariância depende apenas da distância entre os pontos e não da direção entre os mesmos.

Um processo estocástico Z é estacionário de segunda ordem se a esperança de $Z(s)$ é uma constante m em toda região A conforme expressão (3).

$$E\{Z(s)\} = m \quad (3)$$

Ademais, a estrutura de covariância espacial depende unicamente do vetor diferença entre os pontos $d = s - s'$ conforme (4).

$$C(d) = E\{Z(s).Z(s + d)\} - E\{Z(s)\}.E\{Z(s + d)\} \quad (4)$$

A hipótese de estacionaridade de um processo espacial pode ser verificada via análises exploratórias e estatísticas descritivas que consideram explicitamente a localização espacial. Quando a covariância varia simultaneamente com a distância e a direção ela é dita *anisotrópica*. Quando a dependência espacial é a mesma em todas as direções, o fenômeno é *isotrópico*, conforme supracitado.

3.2.4 Modelos inferenciais

A análise espacial é composta por um conjunto de procedimentos encadeados cuja finalidade é a escolha de um modelo inferencial que considere explicitamente o relacionamento espacial presente no fenômeno (DRUCK et al., 2004).

O procedimento inicial da análise espacial é a análise exploratória que se constitui basicamente na visualização dos dados, em geral, por meio de mapas. Observa-se, dessa forma, a distribuição das variáveis sob estudo, identificam-se padrões na distribuição espacial e padrões atípicos – denominados *outliers*. Através desses procedimentos de exploração inicial dos dados, elaboram-se hipóteses a partir das observações realizadas e seleciona-se o modelo inferencial que melhor se adequa aos dados.

Os modelos inferenciais espaciais são categorizados em três grupos: processos pontuais, variação contínua e variação discreta. Para resolução de um problema espacial pode-se empregar um deles, a interação de alguns ou mesmo o uso de todos. Uma das vantagens dos modelos contínuos, por exemplo, é que a inferência não se limita a áreas arbitrariamente definidas. Em contrapartida, modelos discretos permitem mais facilmente estimar parâmetros de associação entre as variáveis.

Conforme Druck et al. (2004), a escolha final será do pesquisador que sabe não existir “o modelo certo”, mas que busca um modelo que se ajuste aos dados e que tenha maior potencial para contribuir para a compreensão do fenômeno sob estudo.

Neste estudo, as PNTs são representadas como pontos no espaço (processo pontual) e como dados agregados por áreas. Esses modelos de inferências são abordados com mais detalhes nas seções subsequentes. Uma abordagem completa a respeito dos modelos espaciais de variação contínua, como em Diggle e Ribeiro (2007), foge ao escopo do presente trabalho.

3.3 ANÁLISE ESPACIAL DE PADRÕES DE PONTOS

Um processo de pontos é um processo estocástico no qual se observam a localização de alguns eventos⁶ de interesse no interior de uma região limitada A . Em Diggle (1983) define-se um processo pontual como um mecanismo estocástico que gera um conjunto finito de eventos. A localização dos eventos gerados por um processo de pontos na área sob análise A é dito *padrão de pontos*. Em alguns casos, variáveis socioeconômicas, por exemplo, podem ser associadas à localização dos eventos observados no espaço geográfico.

⁶ O termo *evento* refere-se a qualquer fenômeno sob estudo que pode ser localizado no espaço geográfico bidimensional por coordenadas geográficas s (DRUCK et al., 2004).

Na análise de padrões de pontos, o objeto de interesse é a própria localização espacial dos eventos em estudo. Avalia-se a distribuição espacial dos pontos e testam-se hipóteses a respeito do padrão observado que pode ser: aleatório, aglomerado (ou *cluster*), regularmente distribuído ou disperso. Se o padrão de eventos pontuais observado apresentar desvios significativos em relação a uma distribuição aleatória é evidência de uma distribuição espacial diferente da completa aleatoriedade e que merece ser objeto de mais análises.

Conforme Druck et al. (2004), os dados de distribuições pontuais possuem as seguintes características:

- ❖ a área dos eventos não é uma medida válida – apesar de que em muitos casos os eventos ocupam algum espaço;
- ❖ em geral, os pontos não estão associados a valores. Considera-se somente o lugar onde ocorrem os eventos;
- ❖ em alguns estudos os pontos podem estar associados a atributos de identificação. Quando tal atributo é objeto de estudo, por meio da comparação da distribuição espacial desses atributos, denomina-se *processo pontual marcado*.

3.3.1 Análise Exploratória de Padrões de Pontos: Caracterização de Distribuições de Pontos

A análise de padrões de pontos é focada na distribuição espacial dos eventos observados. São feitas inferências a respeito do processo subjacente que gera tais pontos. Investiga-se a distribuição dos eventos no espaço e a existência de interação entre eles. A partir da representação dos padrões de pontos na área de estudo é possível ter uma estimativa inicial da distribuição dos pontos e elaborar hipóteses a respeito da distribuição espacial dos eventos.

Os processos pontuais são definidos como um conjunto de pontos irregularmente distribuídos em uma área e cuja localização foi gerada por um mecanismo estocástico. A localização dos pontos é o objeto de estudo que visa compreender o mecanismo gerador dos mesmos. Para caracterização, o processo estocástico é descrito em termos de efeitos de primeira e de segunda ordem.

Efeitos de primeira ordem são efeitos globais ou de larga escala que correspondem a variações no valor médio do processo no espaço. Nesse caso, a medida de interesse é a *intensidade* do processo, isto é, o número de eventos por unidade de área. Efeitos de segunda ordem são locais ou de pequena escala e representam a *dependência espacial* no processo advindo de uma estrutura de correlação espacial.

Sejam n pontos $\{s_1, \dots, s_n\}$ em uma região A onde ocorrem os eventos sob análise. Seja, por exemplo, o fenômeno em estudo os furtos de carros ocorridos em uma cidade. Deseja-se verificar se há um padrão espacial para esse tipo de crime, busca-se por sub-regiões S em A com maior probabilidade de ocorrência do mesmo.

Modela-se o processo pontual considerando sub-regiões S em A por meio da esperança $E[N(S)]$ e da covariância $C[N(s_i), N(s_j)]$, onde $N(S)$ é o número de eventos em S . Objetiva-se encontrar as localizações prováveis de ocorrência de determinados eventos. Para tal, considera-se o valor limite da quantidade de eventos por área. Tal limite corresponde à esperança de $N(S)$ para uma região infinitesimal ds em torno do ponto s de coordenadas (x, y) , quando a mesma tende a zero. Essa esperança é uma propriedade de primeira ordem denominada *intensidade* e é definida em (5).

$$\lambda(s) = \lim_{|ds| \rightarrow 0} \left\{ \frac{E[N(ds)]}{|ds|} \right\} \quad (5)$$

As propriedades de segunda ordem podem ser definidas de forma análoga, considerando a intensidade conjunta $\lambda(s_i, s_j)$ entre duas regiões infinitesimais $|ds_i|$ e $|ds_j|$ que contêm os pontos s_i e s_j , conforme (6).

$$\lambda(d(s_i), d(s_j)) = \lim_{\substack{|ds_i| \rightarrow 0 \\ |ds_j| \rightarrow 0}} \left\{ \frac{C[N(ds_i), N(ds_j)]}{|ds_i|, |ds_j|} \right\} \quad (6)$$

Se o processo é estacionário, $\lambda(s)$ é uma constante então $\lambda(s) = \lambda$; se for isotrópico, $\lambda(s_i, s_j)$ reduz-se a $\lambda(|h|)$, sendo $|h|$ a distância entre os dois pontos. Se o processo é não estacionário, isto é, a intensidade média varia na região A , a modelagem da estrutura de dependência $\lambda(s_i, s_j)$ deve incorporar a variação de $\lambda(s)$.

A maior parte das técnicas em análise espacial de pontos presume um comportamento estacionário e isotrópico do processo aleatório subjacente aos eventos analisados.

3.3.1.1 Aleatoriedade Espacial Completa

No estudo de um processo pontual, um dos primeiros testes realizados é o teste da aleatoriedade espacial completa ou *complete spatial randomness* (CSR). Nesse teste, o evento sob estudo é comparado a distribuições aleatórias que seguem o padrão CSR (usualmente um

processo de Poisson) a fim de levantar o padrão da distribuição espacial dos eventos sob análise.

Sob CSR, os eventos são independentes e estão distribuídos aleatoriamente e uniformemente sobre a área de estudo. Logo, não há regiões onde há maior (ou menor) probabilidade de os eventos ocorrerem e a presença de um dado evento não modifica a probabilidade de outros ocorrerem em sua vizinhança.

A fim de mensurar o grau de semelhança entre os eventos sob análise com o padrão CSR, utilizam-se vários métodos tais como: função G (distância para o evento mais próximo) e função F (distância a partir de um ponto até o evento mais próximo).

3.3.2 Análise Estatística de um Processo Espacial de Pontos

O padrão de pontos pode ser descrito por meio da densidade espacial estatística obtida a partir dos dados observados. Uma função alternativa para mensurar a distribuição espacial dos eventos é a intensidade $\lambda(s)$ do processo de pontos, que é proporcional à densidade espacial. A constante de proporcionalidade entre ambas é o número esperado de eventos do processo pontual na área A . Então, para dois processos com a mesma densidade espacial, mas com diferentes intensidades, o número de eventos observados será maior naquele com maior intensidade (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008).

A intensidade e a densidade espacial são propriedades de primeira ordem, porque mensuram a distribuição dos eventos em toda a região de estudo. Observa-se que a intensidade e a densidade espacial não fornecem informação a respeito da interação entre dois pontos arbitrários. Tal interação é mensurada pelas propriedades de segunda ordem que representam a tendência de os eventos aparecerem em aglomerados, dispersos, independentemente (ou aleatoriamente) ou regularmente espaçados.

De acordo com Bivand, Pebesma e Gómez-Rubio (2008, p. 163), em algumas situações é difícil separar propriedades de primeira e de segunda ordem. Por exemplo, um grupo de eventos situam-se em uma localização específica por que a intensidade é maior nesse local ou por que os eventos estão em aglomerado? Em geral, assume-se que as interações entre os pontos ocorrem em pequena escala, enquanto variações em larga escala são representadas pela intensidade do processo pontual.

3.3.2.1 Processos de Poisson

O processo de Poisson fornece uma abordagem adequada a uma variedade de problemas espaciais de pontos. O mesmo classifica-se em processo de Poisson homogêneo ou

homogeneous Poisson process (HPP) e em processo de Poisson não homogêneo ou *inhomogeneous Poisson processes* (IPP). O HPP e o IPP admitem que os eventos pontuais ocorram independentemente e que estejam distribuídos conforme a intensidade especificada. A principal diferença entre ambos é que o HPP assume que a função de intensidade é constante; enquanto que no IPP, a mesma varia espacialmente. Nesse sentido, o IPP é uma generalização do HPP, ou o HPP é um caso particular do IPP.

O HPP é um tipo de processo de pontos onde os eventos estão distribuídos independente e uniformemente na região A . A localização de um ponto não afeta a probabilidade de outros aparecerem em sua vizinhança e não há região onde há maior probabilidade de os eventos aparecerem. O HPP é também estacionário e isotrópico. É estacionário porque a intensidade é constante e a intensidade de segunda ordem depende apenas da posição relativa entre dois pontos. É isotrópico porque a intensidade de segunda ordem não varia com a rotação. Em resumo, o HPP tem intensidade constante e a intensidade de segunda ordem depende apenas da distância h entre os pontos, independentemente da posição relativa entre eles.

As restrições supracitadas são representadas por uma intensidade constante do processo de pontos, tal que $\lambda(s) = \lambda > 0, \forall s \in A$. O HPP é, portanto, a definição formal de um processo de pontos que é CSR.

Todavia, ressalta-se que a suposição de que o processo de pontos encontra-se distribuído de forma não homogênea é mais realista. Por exemplo, as distribuições da população em uma cidade ou de árvores em uma floresta são influenciadas por diversos fatores que as tornam não homogêneas.

O IPP é uma generalização do HPP, no qual a intensidade é não constante. O mesmo princípio de independência entre os eventos permanece; no entanto, a variação espacial pode ser mais diversa, pois os eventos têm maior probabilidade de ocorrerem em determinadas regiões. Nesse caso, a intensidade é uma função $\lambda(s)$ que varia espacialmente.

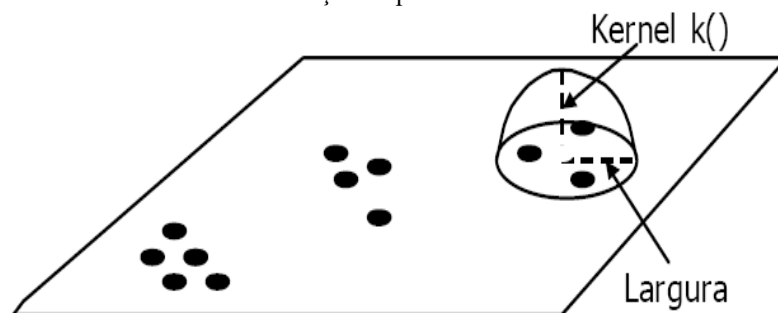
3.3.2.2 Estimação da Intensidade: Kernel Não Paramétrico

A intensidade de um processo de pontos HPP é constante. Nesse caso, o problema de estimar a intensidade é o problema de estimar uma função constante λ tal que o número esperado de eventos na região A ($\int_A \lambda ds$) é igual ao número observado de casos. Uma forma de estimar a intensidade é $n/|A|$, sendo $|A|$ a área da região A e n o número de pontos distribuídos em A .

Para um IPP, pode-se estimar a intensidade de várias formas. Pode ser não parametricamente através de uma função de *kernel* suavizado ou parametricamente por meio de uma função específica cujos parâmetros são estimados por maximização da probabilidade do processo de pontos.

Uma alternativa para analisar o comportamento de padrões de pontos não parametricamente é estimar a intensidade pontual do processo em toda região de estudo via função de *kernel*. Ajusta-se uma função bidimensional sobre os eventos considerados, cuja superfície é proporcional à intensidade das amostras por unidade de área. Por meio de tal função efetua-se a contagem de todos os pontos no interior de uma região de influência, ponderando-os pela distância de cada um em relação à localização de interesse, conforme Figura 3.

Figura 3 – Estimador de intensidade de distribuição de pontos.



Fonte: (DRUCK et al., 2004).

Sejam $s_1, \dots, s_n$ as localizações de n eventos observados em uma região A e s é uma localização genérica no plano bidimensional cujo valor de intensidade deseja-se estimar. O estimador de intensidade, interpolador não paramétrico ou estimador da função de *kernel* em (7) é calculado a partir dos n eventos s_i inscritos em um círculo de raio b (largura de banda ou *bandwidth*) em torno de s a uma distância $\|s - s_i\|$ entre a posição de s e da i -ésima amostra s_i contida no círculo (DIGGLE, 1983; DRUCK et al., 2004).

$$\hat{\lambda}(s) = \frac{1}{b^2} \sum_{i=1}^n \kappa\left(\frac{\|s - s_i\|}{b}\right) / q(\|s\|) \quad (7)$$

Sendo $\kappa(\alpha)$ uma função de *kernel* simétrica ou bivariada. $q(\|s\|)$ é uma função de correção de borda para compensar a perda de dados que há quando s está muito próximo dos limites da região A . A largura de banda b regula o nível de alisamento ou de suavização da função de *kernel*. Pequenos valores de b produzem muitos picos, enquanto valores grandes produzem uma função muito suave. Existem vários tipos de funções de *kernel*. Em (8) apresenta-se a função de *kernel* quártico ou *biweight* em duas dimensões.

$$\kappa(\alpha) = \begin{cases} \frac{3}{\pi} (1 - \|\alpha\|^2)^2, & \text{se } \alpha \in (-1, 1), \\ 0, & \text{caso contrário} \end{cases}, \quad (8)$$

sendo $\|\alpha\|^2$ o quadrado da norma Euclidiana no ponto $\alpha = (\alpha_1, \alpha_2)$ que é igual a $\alpha_1^2 + \alpha_2^2$.

Existem inúmeros métodos para a seleção da largura de banda b ; no entanto, não há um método para escolher um valor para o caso geral (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008). Em geral, testam-se inúmeros valores e escolhe-se um valor que seja plausível conforme o processo sob estudo. Berman e Diggle (1989) propõem um critério para seleção da largura de banda b baseado na minimização do erro quadrático médio da função de alisamento de *kernel* quando o processo de pontos é estacionário.

Quando se estima a intensidade via função de *kernel*, a escolha chave não é a escolha da função de *kernel*, mas da largura de banda, b . Bivand, Pebesma e Gómez-Rubio (2008, p. 167) demonstraram empiricamente que a intensidade do processo de pontos para diferentes funções de *kernel* produzem estimativas semelhantes para uma mesma largura de banda; no entanto, para uma mesma função de *kernel*, diferentes valores de largura de banda produzem resultados totalmente distintos.

No estimador de intensidade não ponderado, todos os pontos interpolados s_i têm igual ponderação que corresponde ao inverso do número total de eventos, n^{-1} . Em contraposição, no estimador de intensidade ponderado, há uma ponderação específica para cada evento ou ponto s_i . Em (9) apresenta-se o estimador de intensidade ponderado, no qual, para cada localização s_i há um peso específico w_i .

$$\hat{\lambda}(s) = \frac{\sum_{i=1}^n w_i \left[1 - \left(\frac{s - s_i}{b} \right)^2 s_i \right]}{\sum_{i=1}^n w_i \left[1 - \left(\frac{s - s_i}{b} \right)^2 \right]} \quad (9)$$

3.3.3 Propriedades de Segunda Ordem

As propriedades de segunda ordem mensuram a força e o tipo de interações entre os eventos do processo de pontos. Objetiva-se determinar se há atração ou competição (ou repulsão) entre os eventos. A intensidade de segunda ordem entre dois pontos i e j representa a probabilidade de qualquer par de eventos ocorrerem nas vizinhanças de ambos eventos.

3.3.3.1 Função G: Distância para o Evento (ou Vizinho) mais Próximo

A função G mensura a distribuição de distâncias a partir de um evento arbitrário i para o evento mais próximo j . Tais distâncias são definidas como $d_i = \min_j \{d_{ij}, \forall j \neq i\}$ sendo $i = 1, \dots, n$. A função G é definida em (10).

$$\hat{G}(r) = \frac{\#\{d_i : d_i \leq r, \forall i\}}{n} \quad (10)$$

O numerador de (10) representa o número de eventos ($\#$) inscritos em um círculo com centro em i e raio r ; n é o número total de pontos inscritos no círculo. Sob CSR, a função G teórica é obtida via (11):

$$G(r) = 1 - e^{-\lambda \pi r^2} \quad (11)$$

onde λ é a *intensidade* (propriedade de primeira ordem) e representa o número médio de eventos por unidade de área.

A semelhança com o padrão CSR pode ser avaliada em um gráfico com a função G empírica $\hat{G}(r)$ – em (10) – versus a função G teórica – em (11) – para um conjunto discreto de distâncias $r = \{0, \dots, r_{max}\}$. Efetuam-se simulações de Monte Carlo para criar uma “nuvem de curvas” no padrão CSR as quais formam um envelope no plano. Tais curvas possuem intensidade estimada $\hat{\lambda}$ na região sob estudo. Verifica-se se a curva obtida via função empírica – equação (10) – está no interior do envelope. Se este for o caso, o padrão de pontos sob estudo encontra-se aleatoriamente distribuído e o espaço não exerce influência sobre a distribuição dos pontos. Nesse caso, a análise espacial em nada contribui para melhor compreensão do problema sob estudo. Se a curva estiver acima do envelope, os pontos encontram-se distribuídos espacialmente em aglomerado ou em *cluster*; caso contrário, estão dispersos.

Outra função para avaliar a semelhança ao padrão CSR, análoga à função G é a função F . A partir dela, mensura-se a distribuição de todas as distâncias a partir de um ponto arbitrário do plano até o evento mais próximo. Tal função é conhecida também como *função do espaço vazio*, pois mensura a média dos espaços entre os eventos. Mais detalhes sobre a função F em Bivand, Pebesma e Gómez-Rubio (2008, p. 162).

3.3.3.2 Função K

A função K de Ripley (1976) permite a avaliação das propriedades de segunda ordem quando o processo espacial é HPP. Mensura-se o número de eventos a uma distância d de qualquer evento conforme (12):

$$K(d) = \lambda^{-1} E[N_0(d)], \quad (12)$$

sendo $E[.]$ o operador esperança e $N_0(d)$ representa o número de eventos até uma distância d ao redor de um evento arbitrário. Ripley (1976) propõe um estimador para estimar a expressão (12) computacionalmente o qual é apresentado em (13):

$$\hat{K}(d) = (n(n-1))^{-1} |A| \sum_{i=1}^n \sum_{j \neq i} w_{ij}^{-1} |\{x_j: d(x_i, x_j) \leq d\}|, \quad (13)$$

onde w_{ij} são pesos proporcionais à área no interior de A de um círculo cujo centro é x_i e raio $d(x_i, x_j)$, que é a distância entre x_i e x_j . O valor teórico da função K para um HPP é $K(d) = \pi d^2$. A partir da comparação entre o valor teórico e o valor estimado, pode-se avaliar o tipo de interação entre os pontos. Como tais interações ocorrem em pequena escala, interessa apenas pequenos valores de d . Se $\hat{K}(d) > \pi d^2$ (valor estimado superior ao teórico) conclui-se que há atração entre os pontos; caso contrário, há competição entre eles. Se as curvas da função K para valores estimados e teóricos para $d = \{0, d_{max}\}$ forem similares, significa que os padrões de pontos sob análise são compatíveis com o padrão CSR, isto é, os eventos são independentes e aleatoriamente distribuídos no espaço.

3.3.4 Processo Pontual Marcado: Estudo Casos-Controles

Uma das aplicações mais relevantes da análise espacial de pontos é a comparação entre dois processos espaciais. Em epidemiologia, por exemplo, é comum existirem dois tipos de pontos na mesma região: os *casos* de certa doença e os *controles*, que em geral representam a distribuição espacial da população. Esse tipo de padrão de pontos é denominado como padrão de pontos marcado, pois cada ponto é atribuído a um grupo.

3.3.4.1 Estudo de Casos-Controles

O interesse primário é na distribuição espacial dos *casos*; no entanto, fatores de risco que afetam tal distribuição devem ser levados em conta. Considerar apenas a distribuição espacial dos *casos* para levantar áreas de alta incidência não é adequado, pois a distribuição dos *casos* é influenciada pela distribuição da população. Para solucionar tal impasse, estima-se a distribuição da população que é representada por um conjunto de *controles*. Tal conjunto é uma amostra aleatória da população de risco. Compara-se o conjunto de *controles* aos *casos* e estima-se a variação espacial do risco.

Neste trabalho realiza-se um estudo de *casos-controles*, onde as UCs irregulares são os *casos* e os *controles* são amostrados aleatoriamente entre as UCs regulares da cidade.

3.3.4.2 Variação Espacial do Risco Relativo

A base de dados é formada por um conjunto de n_1 *casos* e n_0 *controles*. Ambos processos de pontos são realizações de dois IPPs com intensidades $\lambda_1(s)$ e $\lambda_0(s)$, respectivamente. Se a distribuição dos *casos* e *controles* forem a mesma, então resulta que a razão das intensidade é igual a uma constante de proporcionalidade que é a razão entre n_1 e n_0 . Resulta em: $\lambda_1(s) = (n_1/n_0)\lambda_0(s)$.

Sejam $\{s_1, s_2, \dots, s_{n_1}\}$ as localizações na região A dos n_1 *casos* e $\{s_{n_1+1}, s_{n_1+2}, \dots, s_n\}$ as localizações dos n_0 *controles*. Define-se o logaritmo do risco relativo $\rho(s)$ na localização s conforme Kelsall e Diggle (1995a; 1995b) em (14).

$$\rho(s) = \log \left\{ \frac{\lambda_1(s)}{\lambda_0(s)} \right\} \quad (14)$$

Sob a hipótese nula de distribuição espacial igual entre *casos* e *controles*, a razão entre as intensidade de ambos é constante sendo: $\rho_0 = n_1/n_0$.

3.3.4.3 Estimador de Regressão Binária

Em Kelsall e Diggle (1998) propõem-se um estimador de regressão binária para estimar a probabilidade de existir um *caso* em uma localização s . Tal estimador pode ser estendido para incorporação de variáveis.

A probabilidade pode ser estimada considerando uma variável binária Y_i que nomeia *casos* ($Y_i = 1$) e *controles* ($Y_i = 0$) em um conjunto de $n = n_1 + n_0$ eventos. Condiçãoada à localização do ponto s , Y_i é uma realização da variável Y_i de Bernoulli com probabilidade conforme (15).

$$P(Y_i = 1 | S_i = s_i) = P(s_i) = \frac{\lambda_1(s_i)}{\lambda_1(s_i) + \lambda_0(s_i)} \quad (15)$$

3.3.4.4 Teste Global do Risco

O teste global do risco e a identificação de áreas de baixo e de alto risco são efetuados via método de simulação de Monte Carlo (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008; KELSALL e DIGGLE, 1998).

O teste de Monte Carlo baseia-se no fato de que os *casos* e *controles* estão distribuídos igualmente sob hipótese nula – H_0 . Assim, supondo que os *casos* tornem-se *controles* e vice-versa; então o novo conjunto *casos-controles* ainda terá a mesma distribuição espacial e a mesma probabilidade $P(s, x)$. Se isso não ocorrer, a renomeação entre *casos* e *controles* produzirá valores diferentes.

Em (16) testa-se a significância estatística. Assume-se que a variação da probabilidade em (15) é não homogênea e igual a $\hat{P}_i(s_j)$. Em (16) avalia-se $P(s)$ em uma grade regular de pontos $\{s_j, j = 1, \dots, n\}$ com n quadrículas. Calcula-se a soma dos valores $\{[\hat{P}(s_j) - \hat{P}_i(s_j)]^2, i = 1, \dots, k\}$ onde k é o número de renomeações entre *casos* e *controles*, $|c|$ é a área de cada quadrícula, $\hat{P}(s_j)$ é o valor estimado da probabilidade inicial (estimada a partir do conjunto original *casos-controles*, sem renomeações) e $\hat{P}_i(s_j)$ com $i = 1, \dots, k$ são as probabilidades estimadas por renomeações entre *casos* e *controles*.

$$\hat{T}^i = |c| \sum_{j=1}^n [\hat{P}(s_j) - \hat{P}_i(s_j)]^2 \quad (16)$$

Neste teste, avaliam-se k valores do teste estatístico em (16) por renomeações entre *casos* e *controles*. Os *casos* tornam-se *controles* e vice-versa, mantendo-se n_1 e n_0 fixos. Calcula-se a nova probabilidade $\hat{P}_i(s_j)$, $i = 1, \dots, k$ para cada novo conjunto de *casos* e *controles* obtido aleatoriamente. Dessa maneira, tem-se uma série de valores $T^1, \dots, T^k$ sob hipótese nula H_0 . Se T^0 é o valor de T para o conjunto de *casos-controles* original (sem renomeações) obtida a partir de (17) onde $\hat{P}_0 = n_1/(n_1 + n_0)$. A significância (também denominada *pvalor*) pode ser obtida via: $pvalor = (t + 1)/(k + 1)$, onde t é o número de valores T^i com $i = 1, \dots, k$ que são maiores do que T^0 , isto é, $t = \#(T^i > T^0)$.

$$\hat{T}^0 = |c| \sum_{i=1}^n [\hat{P}(s_j) - \hat{P}_0]^2 \quad (17)$$

O teste avalia para cada ponto s_j da grade de pontos a proporção de valores $\hat{P}_i(s_j)$ que são menores do que $\hat{P}(s_j)$ onde $\hat{P}_i(s_j)$, $i = 1, \dots, k$ são as probabilidades estimadas por renomeações entre *casos* e *controles*.

3.3.4.5 Regressão Binária via Modelo Aditivo Generalizado

A abordagem semiparamétrica de modelos aditivos generalizados ou *generalised additive models* (GAM) permite a inclusão de variáveis que caracterizam a região para estimar a variação espacial das PNTs por meio de regressão logística padrão. A variação espacial residual é modelada por meio de uma função de suavização ou *spline*.

Apresenta-se a inclusão dos efeitos das variáveis no modelo, conforme Kelsall e Diggle (1998) em (18). Nessa expressão, x é um vetor das variáveis observadas na localização s , β são os coeficientes das variáveis x que modelam seus efeitos e $g(s)$ é uma função suave das coordenadas espaciais s que levam em consideração outros fatores não explicados pelas variáveis x .

$$\text{logit}\{P(s, x)\} = \log \left\{ \frac{P(s, x)}{1 - P(s, x)} \right\} = \beta x + g(s) \quad (18)$$

Em (18) estima-se uma superfície de probabilidades $P(s, x)$ considerando a influência das variáveis x e de uma variação espacial devida a outros fatores não considerados como variáveis e representados por $g(s)$.

Se as variáveis x fossem eliminadas de (18), a expressão resultante seria outro modo de estimar uma superfície de probabilidade (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008). Se o risco é assumido constante na região, então $g(s) = 0$ e o modelo em (18) reduz-se a um modelo de regressão logística padrão. Portanto em (18), conforme Shimakura et al. (2001), tem-se um modelo de regressão logística estendido por um componente aditivo $g(s)$ que varia suavemente no espaço.

O procedimento para estimação de β e $g(s)$ baseia-se em métodos iterativos usuais de modelos aditivos generalizados (WOOD, 2006).

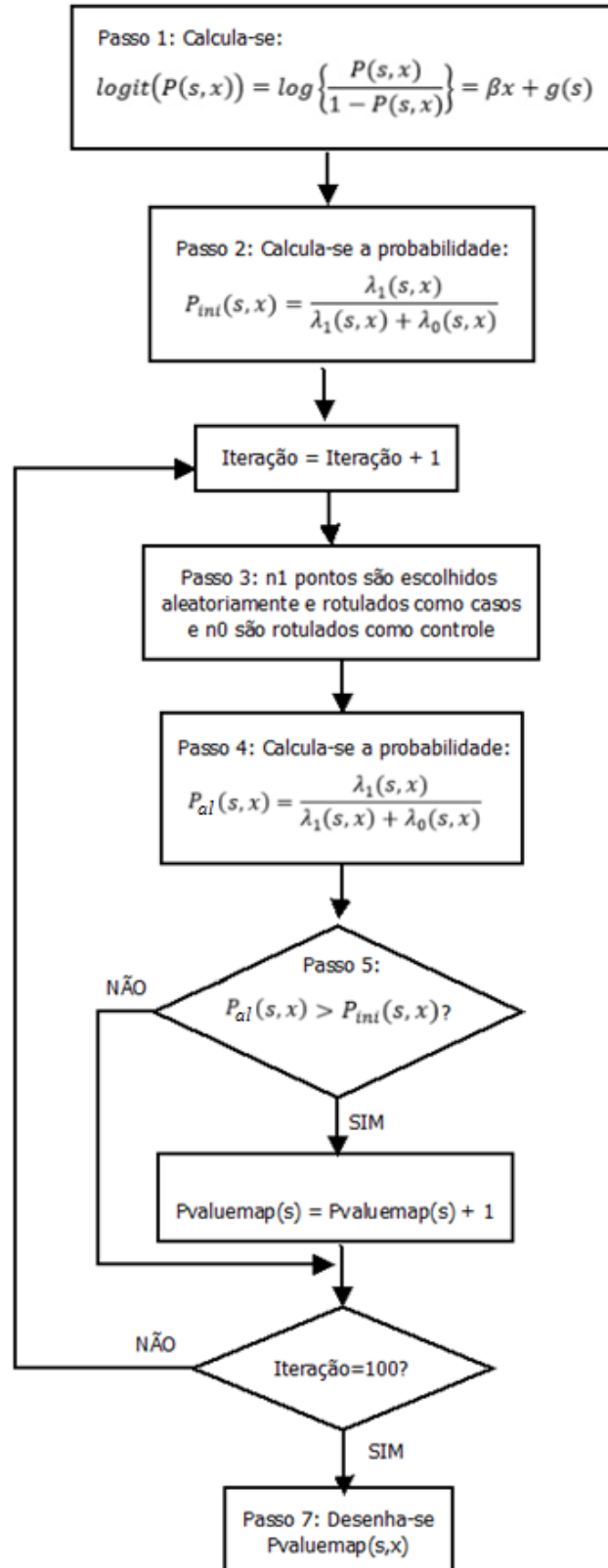
3.3.4.6 Mapa de Probabilidade de Perdas

Nesta seção, detalham-se os passos para obtenção dos mapas de probabilidade de perdas a partir do estimador de regressão em (15) aplicado em um estudo de *casos-controles* e ponderado pelas probabilidades advindas de um modelo de regressão espacial GAM.

Na Figura 4 mostra-se o fluxograma completo com todos os passos para obtenção do mapa de probabilidades de perdas. Os dados de entrada são as UCs flagradas em irregularidades (*casos*); as UCs obtidas aleatoriamente a partir da base de clientes da concessionária

(*controles*) e as variáveis obtidas no censo demográfico IBGE (2010) e da rede de distribuição. Seguem os passos da metodologia.

Figura 4 – Fluxograma do modelo para geração do mapa de probabilidade de perdas.



Fonte: Adaptado de Faria, Melo e Padilha-Feltrin (2014b).

- ❖ *Passo 1:* Executa-se o GAM – expressão (18) – para determinar a probabilidade de cada ponto da base de dados *casos-controles* considerando sua localização espacial s e as variáveis independentes x da subárea na qual a localização s está inserida.
- ❖ *Passo 2:* Cálculo da probabilidade inicial $P_{ini}(s, x)$ via equação (15). $\lambda_1(s, x)$ e $\lambda_0(s, x)$ são as intensidade ponderadas calculadas via expressão (9).
- ❖ *Passo 3:* Uma nova base de dados *casos-controles* é produzida via sorteio; n_1 pontos são rotulados como *casos* e n_0 pontos são rotulados como *controles*.
- ❖ *Passo 4:* Recalcula-se a probabilidade $P(s, x)$ calculada no *Passo 2* e obtém-se $P_{al}(s, x)$ para o novo conjunto aleatório obtido no *Passo 3* de *casos-controles*. Utiliza-se novamente a equação (15) para determinar a probabilidade de cada ponto da base de dados *casos-controles* considerando sua localização espacial s e a influência das variáveis x .
- ❖ *Passo 5:* Se $P_{al}(s, x) > P_{ini}(s, x)$ então incrementa-se $P_{valuemap}(s)$. Os passos de 3 a 5 repetem-se até um número predefinido de iterações. $P_{al}(s, x)$ é a probabilidade calculada no *Passo 4* e $P_{ini}(s, x)$ é a probabilidade calculada no *Passo 2* para o conjunto *casos-controles* original.
- ❖ *Passo 6:* Esboça-se o gráfico do mapa de probabilidade armazenado em $P_{valuemap}(s)$.

3.4 ANÁLISE ESPACIAL DE DADOS AGREGADOS POR ÁREAS

A análise espacial de dados de áreas é composta por uma coleção de métodos para análise de dados espaciais cuja localização está associada a áreas delimitadas por polígonos. Tal análise é útil quando se lida, por exemplo, com eventos agregados por municípios, bairros ou setores censitários, onde não se dispõe da localização exata dos eventos, mas somente de um valor agregado por área. O IBGE particiona os municípios do Brasil em polígonos menores denominados setores censitários. A fim de garantir a confidencialidade dos dados coletados, o IBGE não divulga os dados dos indivíduos pesquisados; por isso, agrega-os por setores censitários. Para cada setor são disponibilizadas publicamente, a cada decênio⁷, centenas de variáveis socioeconômicas tais como: renda média, número de indivíduos alfabetizados, número de pessoas atendidas por serviços básicos de água tratada e coleta de lixo, etc. Portanto,

⁷ O censo demográfico brasileiro é realizado pelo IBGE a cada dez anos. Neste trabalho, utilizam-se dados do último censo demográfico realizado em 2010 (IBGE, 2010).

a análise espacial de dados de área é uma técnica viável para o aproveitamento das informações do censo demográfico.

Nos modelos inferenciais de variação discreta, o espaço geográfico A é constituído por um conjunto fixo de unidades espaciais. O modelo de distribuição comumente utilizado considera um processo estocástico $\{Z_i; i = 1, \dots, n\}$, composto por um conjunto de variáveis aleatórias. Pretende-se construir uma aproximação para a distribuição conjunta dessas variáveis $Z = \{Z_1, \dots, Z_n\}$ onde cada variável aleatória está associada a uma das áreas e possui uma distribuição a ser estimada. Se o processo for considerado estacionário, o valor esperado de Z_i é a média global da região e a estrutura de covariância depende apenas da distância ou da estrutura da vizinhança entre as áreas.

Na análise espacial de áreas, os dados são majoritariamente oriundos de levantamentos populacionais tais como censos demográficos, estatísticas de saúde, de crimes, etc. Supõe-se que há homogeneidade interna, isto é, mudanças importantes somente ocorrem nos limites. Sabe-se que nem sempre tal premissa se cumpre. As unidades de levantamento são definidas frequentemente por critérios políticos (como no caso de Estados e Municípios) ou operacionais (como nos setores censitários), por exemplo. Não há qualquer garantia de que a distribuição do evento seja homogênea no interior das unidades ou das áreas sob estudo. Segundo Druck et al. (2004), em países com grandes contrastes sociais é frequente que grupos sociais distintos estejam agregados em uma mesma região de coleta – favelas e áreas nobres, por exemplo – resultando em indicadores distorcidos que representam a média entre população muito distintas.

A forma usual de apresentação de dados agregados por áreas é o uso de mapas coloridos denominados mapas temáticos (ou mapas coropléticos) que contêm o padrão espacial do fenômeno sob análise. Além da visualização em mapas, é possível analisar os dados espaciais agregados e responder a perguntas tais como: o padrão observado é aleatório ou apresenta agregação definida? Há agrupamentos de áreas com padrões diferenciados dentro da região sob análise?

De acordo com Druck et al. (2004), o primeiro passo é a escolha de um modelo inferencial a ser utilizado. A hipótese mais comum é supor que as áreas são diferenciadas e que cada uma possui uma identidade própria. Em estatística, tal fato implica que cada área possui uma distribuição de probabilidade distinta das demais, o conhecido *modelo espacial discreto*.

3.4.1 Distribuição dos Dados em Áreas

O modelo mais utilizado para dados de área é o modelo de variação espacial discreta. Seja um processo estocástico $Z_i, i = 1, \dots, n$. Z_i é a realização do processo espacial na área i e n é o total de áreas A_i . Objetiva-se construir uma aproximação para a distribuição conjunta das variáveis aleatórias $Z = \{Z_1, \dots, Z_n\}$. Z_i é uma variável aleatória que descreve a contagem ou indicador associada à área A_i . Supõe-se que a variável aleatória Z_i , que descreve o número de ocorrências em cada área, possa ser associada a uma distribuição de probabilidade de Poisson. Conforme abordado na análise espacial de padrões de pontos (seção 3.3), a distribuição estatística de Poisson é mais adequada a fenômenos que envolvem contagem de eventos, como é o caso dos dados agregados por áreas.

3.4.2 O Problema da Unidade de Área Modificável e o Problema de Escala

Os dados agregados por área possuem uma limitação segundo a qual, para uma dada população sob análise, a definição espacial das fronteiras das áreas influi nos resultados obtidos. As estimativas obtidas são funções do modo como as áreas foram agrupadas; pode-se, por exemplo, obter outros resultados apenas alterando as fronteiras das áreas. Este é o problema da unidade de área modificável.

Os dados disponibilizados pelo IBGE estão agregados por área; no entanto, o objeto de estudo refere-se a características e relacionamentos individuais – que neste estudo são as UCs irregulares. Logo, para viabilizar o uso das informações do IBGE e da análise espacial de áreas, as UCs irregulares são agregadas por áreas, ou seja, contabilizam-se o número de UCs irregulares por setor censitário.

Além da influência das fronteiras das áreas na análise, conforme supracitado, há também a influência da escala que está relacionada à dimensão da unidade de área considerada. De acordo com Druck et al. (2004), a redução de escala (áreas maiores) tende a homogeneizar os dados, reduzir a flutuação aleatória e reforçar correlações que aparentam ser mais fortes do que em áreas menores (que possuem maior escala). Em contraposição, o aumento da escala (áreas menores) tende a considerar a heterogeneidade dos dados em torno da média, aumentar a flutuação aleatória e atenuar correlações. Quanto maior o “zoom” (ou a escala) mais aleatório torna-se o fenômeno sob análise. Ressalta-se que não existe uma escala ideal, mas apenas modelos que se prestam ao que se deseja esclarecer. O problema da escala é um efeito inerente aos dados agregados por áreas. Não pode ser removido, tampouco ignorado.

3.4.3 Matriz de Ponderação, de Vizinhança ou Matriz de Proximidade Espacial

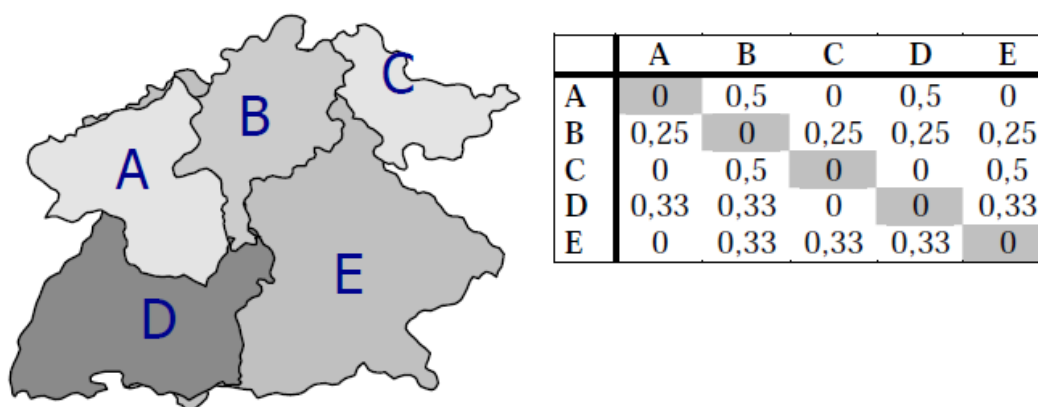
A fim de estimar a variabilidade espacial dos dados, utiliza-se a matriz de proximidade espacial, matriz de vizinhança ou matriz de ponderação. Seja um conjunto de n áreas $\{A_1, \dots, A_n\}$; constrói-se a matriz de ponderação $\mathbf{W}_{(n \times n)}$ onde cada um dos elementos w_{ij} representa uma *medida de proximidade* ou medida de ponderação entre as áreas A_i e A_j . Seguem alguns critérios comumente utilizados para cálculo dessa matriz:

- ❖ $w_{ij} = 1$, se o centroide de A_i está a uma determinada distância do centroide de A_j ; caso contrário, $w_{ij} = 0$;
- ❖ $w_{ij} = 1$, se A_i tem um (ou mais) pontos de tangência a A_j ; caso contrário, $w_{ij} = 0$;
- ❖ $w_{ij} = L_{ij}/L_i$, onde L_{ij} é o comprimento da fronteira comum entre A_i e A_j . L_i é o perímetro de A_i .

A fim de simplificar os cálculos dos índices de correlações espaciais, as linhas da matriz de ponderação $\mathbf{W}_{(n \times n)}$ são normalizadas de modo que a somatória de cada linha seja unitária. Na Figura 5 há um exemplo ilustrativo de uma matriz de proximidade espacial normalizada na qual os valores da matriz refletem o segundo critério de adjacência supracitado.

Observa-se que a matriz de proximidade espacial pode ser generalizada para vizinhos de maior ordem $\mathbf{W}^{(2)}, \dots, \mathbf{W}^{(n)}$ (vizinhos dos vizinhos). Na Figura 5, por exemplo, as áreas A e C são vizinhas na matriz de ponderação de segunda ordem – A é vizinho de B e B é vizinho de C . Neste trabalho utilizam-se apenas matrizes de vizinhança de primeira ordem.

Figura 5 – Matriz de ponderação espacial de primeira ordem normalizada.



Fonte: (DRUCK et al., 2004).

As técnicas de análise espacial são empregadas majoritariamente em problemas relacionados à epidemiologia, ecologia, levantamento de recursos naturais, agricultura, etc.

Nessas áreas de estudo, em geral, é válida a primeira lei da geografia de Tobler (seção 3.2.1) cuja consequência é que a dependência espacial está presente em todas as direções e torna-se mais fraca à medida que se aumenta a dispersão na localização dos dados.

Esta lei é comumente empregada para construção da matriz de ponderação $W_{(n \times n)}$. Conforme os três critérios supracitados e outros inúmeros critérios da literatura avaliada, citados em Bivand, Pebesma e Gómez-Rubio (2008, p. 237-255), por exemplo, em geral, os elementos da matriz de ponderação w_{ij} tendem a ser inversamente proporcionais às distâncias entre os centroides das áreas A_i e A_j . Essa relação faz todo sentido na área da epidemiologia. No estudo de John Snow (seção 3.1), por exemplo, observou-se que a maior parte dos casos de cólera estavam concentrados próximos aos poços de água contaminada. Quanto maior a distância do indivíduo a tais poços, menor o risco de o mesmo contrair a cólera.

No entanto, conforme discutido na seção 3.2.1, no problema das PNTs, não há evidência clara da aplicação da lei de Tobler. As UCs vizinhas de UCs com fraudes no medidor de energia também o farão? Logo, não há evidência concreta de que os critérios encontrados na literatura avaliada para construção da matriz de ponderação representem adequadamente o problema das PNTs.

Portanto, umas das contribuições do presente trabalho é a implementação de um novo critério para construção da matriz de vizinhança, a *vizinhança semântica*. Tal vizinhança baseia-se no grau de semelhança das áreas com respeito às suas respectivas variáveis socioeconômicas e da rede de distribuição. Dessa maneira, áreas vizinhas não precisam ter necessariamente fronteiras em comum; ademais, é possível que áreas distantes tenham medida de proximidade w_{ij} maior do que áreas próximas. As técnicas e adaptações realizadas para construção matriz de vizinhança serão abordadas em detalhes no capítulo 4.

3.4.4 Análise Exploratória

Uma etapa imprescindível da análise exploratória de áreas consiste na identificação da estrutura de correlação espacial que melhor descreva os dados. Estima-se a magnitude da autocorrelação espacial entre as áreas. O resultado da correlação depende fundamentalmente da definição da vizinhança adotada contida na matriz de ponderação definida na seção anterior.

A identificação do grau de dependência entre duas variáveis é uma importante ferramenta para auxílio à formulação de hipóteses sobre a dependência espacial no fenômeno sob estudo e o estabelecimento de relações de causa e efeito.

O grau de correlação global entre duas variáveis pode ser estimado via coeficientes numéricos que indicam se há relação entre ambas variáveis e qual a intensidade e a significância desta relação.

3.4.4.1 Coeficiente de Correlação de Spearman r_s

O coeficiente de correlação de Spearman r_s é uma alternativa para determinar a relação entre duas variáveis. Baseia-se na ordenação ou *ranking* dos valores das duas variáveis.

O cálculo de r_s é feito via organização dos valores das duas variáveis em ordem crescente (ou decrescente), substituindo cada valor original da série de dados por um número inteiro positivo que representa a posição hierárquica ou ordinal do valor original na série de dados. O valor desse coeficiente de correlação é calculado via (19):

$$r_s = 1 - \frac{6 \sum d^2}{n(n+1)(n-1)} \quad (19)$$

sendo n o número de unidades de observação, ou o número de amostras e d^2 é o quadrado da diferença entre a posição ordinal ou *ranking* do valor das duas variáveis sob análise em uma mesma unidade geográfica.

O coeficiente de Spearman ou coeficiente de correlação ordinal não é sensível a assimetrias na distribuição nem a presença de *outliers*, não exigindo que os dados venham de populações com distribuição normal.

Segundo Ferreira (2014, p. 104), o coeficiente de Spearman é muitas vezes mais útil do que o coeficiente de correlação de Pearson. Tal coeficiente permite identificar em um conjunto de variáveis aquelas que estão mais associadas entre si.

Uma alternativa para avaliar o nível de significância do coeficiente de correlação é através do teste t de *Student*. Formulam-se duas hipóteses sobre r_s :

- ❖ H_0 : o valor do coeficiente de correlação ocorreu por mero acaso.
- ❖ H_1 : o valor do coeficiente de correlação é maior do que se poderia esperar, caso acontecesse por acaso.

Para o teste das hipóteses supracitadas, utiliza-se a distribuição t de *Student* com $n - 2$ graus de liberdade. A equação (20) estabelece a relação entre os parâmetros envolvidos no cálculo da função t :

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{(1-r_s^2)}} \quad (20)$$

sendo n o número de amostras.

Em seguida, compara-se o valor t calculado com os valores críticos da distribuição t de *Student* para diversos graus de liberdade para obter o valor do nível de significância. Para mais detalhes a respeito dos testes estatísticos recomendam-se textos clássicos de estatística como Bussab e Morettin (2010), por exemplo.

Observa-se que os coeficientes de correlação variam entre -1 e 1 . Quando $r_s = -1$, diz-se que há correlação negativa perfeita entre duas variáveis; nesse caso, à medida que aumenta o valor de uma delas, o valor da outra decresce na mesma proporção; são inversamente proporcionais. Quando $r_s = 1$ há uma correlação positiva perfeita entre as variáveis que são diretamente proporcionais. Na Tabela 6 agrupam-se os intervalos do coeficiente de correlação em sete categorias a fim de atribuir uma intensidade à correlação.

Tabela 6 – Intervalos do coeficiente de correlação e magnitude da correlação.

<i>Intervalos do Coeficiente de Correlação de Spearman r_s</i>	<i>Magnitude da Correlação</i>
$0 < r_s < 0,3$	Fracamente positiva
$0,3 < r_s < 0,8$	Moderadamente positiva
$0,8 < r_s < 1,0$	Altamente positiva
$-0,3 < r_s < 0$	Fracamente negativa
$-0,8 < r_s < -0,3$	Moderadamente negativa
$-1,0 < r_s < -0,8$	Altamente negativa
$r_s = 0$	Não há correlação

Fonte: Adaptado de Ferreira (2014, p. 99).

3.4.4.2 Técnicas de Corte da Variável para Classificação e Construção de Legendas em Mapas Temáticos

Uma das formas gráficas mais adequadas à representação da distribuição de frequência de dados é o agrupamento desses valores em uma série de intervalos discretos denominados classes. As classes podem ser dispostas em um diagrama de barras verticais (histograma de frequência) ou utilizadas para construção de mapas temáticos no qual se diferencia cada classe pela coloração.

A fim de identificar como se distribuem espacialmente os valores das variáveis, as mesmas são agrupadas em classes segmentadas por intervalos de valores de cada variável.

Conforme Ferreira (2014, p. 69), qualquer processo de classificação é parcialmente arbitrário, principalmente no que se refere à escolha do número ideal de classes segundo o qual a série completa dos dados originais será segmentada. A quantidade de classes e o valor da amplitude dos intervalos das respectivas classes dependem da técnica utilizada para a classificação da série de dados da variável em estudo. A técnica de Sturges, por exemplo, uma das mais comuns em estatística, fornece o número de classes k a partir do critério de intervalos iguais para todas as classes. Há vários outros critérios para definição das classes como: intervalos iguais, quantis, quebras naturais na distribuição dos dados da variável, intervalos de desvio padrão, etc.

Para construção das legendas dos mapas temáticos deste trabalho, optou-se pelo critério de *intervalos de desvio padrão*. Segundo Ferreira (2014, p. 261), “esta técnica é a que melhor considera a distribuição estatística da variável a ser mapeada”. Calcula-se a média ($\bar{X}$) e o desvio padrão (σ) da distribuição; em seguida, segmenta-se a série de dados em intervalos de classes cujos limites são proporções do desvio padrão dos valores da variável. Cada classe é posicionada em relação à distância até a média de toda a série de dados. As classes posicionadas abaixo da média têm limites medidos em valores negativos de desvio padrão ($-\sigma$) e as classes posicionadas acima da média têm limites medidos em valores positivos de desvio padrão ($+\sigma$). Generalizando, o conjunto das k classes de uma série de dados pode ser obtido da seguinte forma: $k = \{\dots, (\bar{X} - 2\sigma), (\bar{X} - \sigma), \bar{X}, (\bar{X} + \sigma), (\bar{X} + 2\sigma), \dots\}$. Alternativamente, pode ser utilizado $1/2 \sigma$ ou $1/4 \sigma$ em lugar de σ para séries de dados pouco esparsas.

Em Druck et al. (2004) observa-se que esta técnica é mais apropriada ao uso em séries de dados que se ajustam à distribuição normal. No entanto, em Ferreira (2014, p. 262), ressalta-se que é possível utilizá-la também em outros casos, com o objetivo de mostrar em mapas, as *distorções locais* de algumas unidades locais em relação à média regional. Tais distorções podem ser tão importantes quanto a regra de adequação à distribuição normal.

3.4.5 Modelos de Regressão

Os modelos de regressão são muito comuns em estudos de dados espaciais de áreas. São ferramentas da estatística que utilizam o relacionamento que há entre duas ou mais variáveis de modo que uma delas (variável dependente – y) possa ser descrita ou seu valor estimado a partir das demais (variáveis independentes ou variáveis explicativas – x).

Segundo Druck et al. (2004), na presença de autocorrelação espacial nos dados, as estimativas do modelo devem incorporar tal estrutura espacial, porque a dependência entre as observações reduz o poder explicativo do modelo – ver seção 3.2.2.

Em uma análise de regressão linear, objetiva-se quantificar a relação linear entre uma variável dependente y e um conjunto de variáveis independentes ou explicativas x conforme equação matricial em (21):

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2), \quad (21)$$

onde Y é um vetor ($n \times 1$) de variáveis dependentes de observações extraídas de cada uma das n áreas; X é uma matriz ($n \times k$) com k variáveis explicativas tomadas das n áreas; β é um vetor ($k \times 1$) com os coeficientes da regressão e ε é o vetor ($n \times 1$) de erros aleatórios com distribuição normal, média nula e variância constante σ^2 .

Na análise de regressão objetiva-se:

- ❖ encontrar um bom ajuste entre valores preditos pelo modelo e os valores observados da variável dependente;
- ❖ descobrir quais variáveis explicativas contribuem de forma significativa para este relacionamento linear.

Para atingir esses objetivos supracitados, a hipótese padrão é que as observações são não correlacionadas, e, por consequência, os resíduos ε do modelo são igualmente independentes e não correlacionados com a variável dependente Y .

Todavia, em dados espaciais onde está presente a dependência espacial, é improvável que a hipótese padrão de observações não correlacionadas seja verdadeira. Os resíduos da regressão continuam apresentando autocorrelação espacial presente nos dados que se manifestam por diferenças regionais significativas nas relações do modelo.

O estudo dos resíduos da regressão a procura de sinais da estrutura espacial é uma etapa importante na regressão espacial. Testes de autocorrelação espacial podem indicar que as observações estão mais correlacionadas do que seria esperado sob uma condição de independência. Na presença de autocorrelação, deve-se especificar um modelo que não despreze as interferências causadas pela mesma.

Os modelos de regressão espacial são categorizados em globais e locais. Modelos globais modelam a estrutura espacial com apenas um único parâmetro; modelos locais dividem o espaço em sub-regiões, em cada sub-região, são realizadas regressões espaciais.

3.4.5.1 Modelos de Regressões Espaciais Globais

A inclusão dos efeitos espaciais nos modelos de regressão tradicionais é feita de variadas formas. Os modelos de regressão espaciais globais são mais simples e restritos. Eles supõem que é possível capturar toda a estrutura de correlação espacial em apenas um único parâmetro que é incluso no modelo de regressão tradicional em (21).

No modelo espacial autorregressivo misto ou *spatial autoregressive* (SAR), por exemplo, (também conhecido por *spatial lag model*) atribui-se a autocorrelação espacial ignorada à variável dependente y . A ideia básica é incorporar a autocorrelação espacial como uma componente do modelo. A dependência espacial é considerada via adição de um novo termo ao modelo de regressão na forma de uma relação espacial para a variável dependente y conforme (22):

$$Y = \rho WY + X\beta + \varepsilon, \quad (22)$$

sendo W a matriz de ponderação (ou matriz de vizinhança) e ρ é o coeficiente espacial autorregressivo. O produto WY expressa a dependência espacial em Y . A hipótese nula para a não existência de autocorrelação é representada via $\rho = 0$. Nesse caso, retorna-se ao modelo de regressão tradicional em (21).

Há outro modelo de regressão espacial com parâmetros globais que considera os efeitos espaciais como ruídos denominado *modelo do erro espacial* ou *conditional autoregressive* (CAR) – também denominado *spatial error model*. Nesse contexto, os efeitos da autocorrelação espacial são associados ao termo de erro ε conforme (23):

$$Y = X\beta + \varepsilon, \quad \varepsilon = \lambda W + \xi, \quad (23)$$

sendo λW a componente do erro que leva em consideração os efeitos espaciais; λ é o coeficiente autorregressivo e ξ é a componente do erro com variância constante e não correlacionada. A hipótese nula para a não existência de autocorrelação é representada via $\lambda = 0$.

Conforme Druck et al. (2004), os modelos de regressão espacial com efeitos globais (como os supracitados, por exemplo) partem de uma suposição de que o processo espacial subjacente aos dados sob análise é estacionário. Isso possibilita que os padrões de autocorrelação espacial existentes nos dados possam ser capturados em apenas um único parâmetro. Todavia, para dados censitários de médio e grande porte, a natureza do processo espacial é tal que diversos padrões de associação espacial podem estar presentes. Portanto, modelos de re-

gressões locais cujos parâmetros variam no espaço podem ser mais adequados para representar a heterogeneidade espacial comumente existente nos dados socioeconômicos.

3.4.5.2 Modelos de Regressões Espaciais Locais

Os modelos de regressões espaciais locais são mais adequados para representar a heterogeneidade espacial em processos espaciais não estacionários em comparação aos modelos de regressões globais.

Entre os inúmeros modelos de regressões espaciais locais destacam-se os modelos com variação espacial discreta e os modelos com variações locais contínuas sem forte tendência global.

Na primeira categoria há os modelos de regressão de regimes espaciais de forma discreta. Tais modelos particionam o espaço em k sub-regiões estacionárias. Cada sub-região possui seu padrão espacial próprio. Realizam-se regressões em separado em cada sub-região conforme (24):

$$Y_{k'} = X_{k'}\beta_{k'} + \varepsilon_{k'}, \quad k' = \{1, \dots, k\} \quad (24)$$

Apesar de cada um dos k regimes possuírem seus próprios coeficientes β_k ; todo o conjunto de observações disponíveis é utilizado em cada regressão. Os regimes espaciais são estimados via técnicas de análise exploratória.

Na categoria dos modelos de variações locais contínuas destaca-se a regressão geograficamente ponderada ou *geographically weighted regression* (GWR). Baseia-se no ajuste de um modelo de regressão a cada ponto observado, ponderando as demais observações como função da distância a este ponto. Cada ponto considerado possui seus próprios coeficientes de ajuste. Tais parâmetros podem ser apresentados visualmente a fim de identificar os relacionamentos espaciais entre as variáveis. O modelo padrão GWR é descrito via equação (25):

$$Y(s) = \beta(s)X + \varepsilon, \quad (25)$$

sendo $Y(s)$ a variável aleatória dependente representando o processo no ponto s , $\beta(s)$ são os parâmetros estimados em s . Os parâmetros $\beta(s)$ do modelo são estimados via mínimos quadrados para o caso não espacial em (26):

$$\beta = (X^T X)^{-1} X^T Y, \quad (26)$$

que é generalizada usando um método de ajuste local conforme (27):

$$\beta(s) = (X^T W(s) X)^{-1} X^T W(s) Y \quad (27)$$

No GWR básico, originário no artigo seminal de Brunson, Fotheringham e Charlton (1996), o ajuste da ponderação local w_{ij} é feito de modo que os pontos mais próximos exerçam maior influência; análogo aos estimadores de densidade por *kernel*. Para tal ajuste, pode-se, por exemplo, utilizar o *kernel* Gaussiano apresentado em (28):

$$w_{ij}(s, b) = \frac{1}{2\pi b} e^{\left(-\frac{d_{ij}^2}{2b^2}\right)}, \quad (28)$$

sendo b o raio de influência, d_{ij} a distância entre a localização considerada s e o j -ésimo ponto.

Nesta seção apresentaram-se modelos de regressões globais e locais típicos que podem ampliar sobremaneira a compreensão acerca de padrões espaciais associados aos dados de área. Neste trabalho, utiliza-se o modelo de regressão local GWR para capturar a heterogeneidade espacial das PNTs. Nesta seção, mostraram-se aspectos básicos do GWR o qual será abordado com mais detalhes no capítulo 4. Ressalta-se que para melhor representar as PNTs é preciso realizar algumas modificações nos critérios para construção da matriz de ponderação do GWR básico. Conforme supracitado, a determinação da ponderação w_{ij} via *kernel* Gaussiano (equação 28) aumenta a influência de áreas vizinhas próximas e reduz a influência daquelas que estão distantes. No capítulo 4 define-se a *vizinhança semântica* que amplia a noção de vizinhança tradicional baseada comumente em áreas fronteiriças. Com esta nova abordagem, é possível, por exemplo, que áreas distantes exerçam maior influência do que áreas próximas e vice-versa.

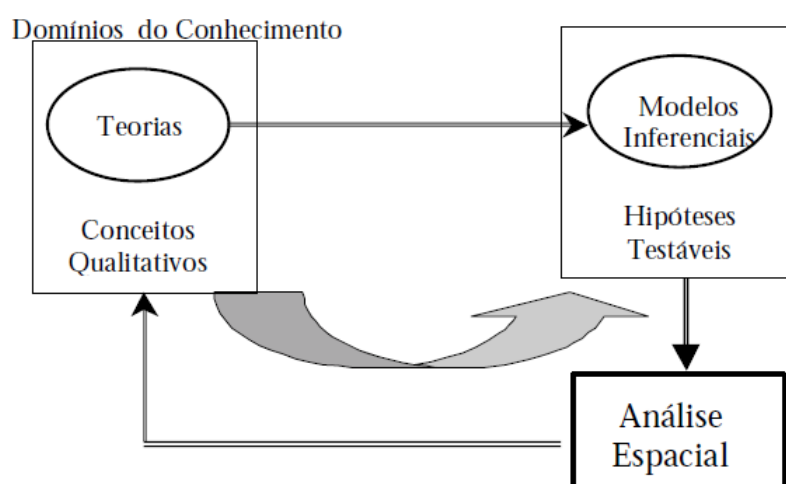
3.5 CONSIDERAÇÕES FINAIS

Neste capítulo destacaram-se os principais conceitos da análise espacial de dados geográficos. Deu-se ênfase à análise espacial de padrões de pontos e análise espacial de áreas; técnicas, que serão empregadas para modelar espacialmente as PNTs.

Na Figura 6 apresenta-se o fluxo de produção de conhecimento via técnicas de análise espacial de dados geográficos. Os especialistas das áreas de estudo na quais se utilizam técnicas de análise espacial desenvolvem teorias sobre os fenômenos – conceitos qualitativos. Tais teorias incluem hipóteses sobre o comportamento espacial dos dados. De posse

dessas teorias, formulam-se modelos inferenciais quantitativos que podem ser submetidos a testes de validação por meio das técnicas de análise espacial. O resultado numérico pode dar suporte à aceitação ou à rejeição dos conceitos qualitativos iniciais. Em outras palavras, as técnicas quantitativas da análise espacial devem estar a serviço do conhecimento dos especialistas e nunca devem ser utilizadas como um fim em si. Portanto, é preciso interpretar os resultados quantitativos obtidos via análise espacial e adequá-los ao contexto de cada problema sob estudo.

Figura 6 – Fluxo de produção de conhecimento em análise espacial.



Fonte: (DRUCK et al., 2004).

Segundo Bivand, Pebesma e Gómez-Rubio (2008, p. 308), a disponibilidade de implementação de um método não significa que o mesmo seja o melhor. O analista é responsável pela escolha do método e implementação em relação à situação específica, aos requerimentos e aos dados disponíveis.

No próximo capítulo abordam-se o modelo de regressão local GWR e aspectos fundamentais da modelagem espaço-temporal, ou modelagem dinâmica com ênfase às cadeias de Markovianas. Ambas técnicas serão utilizadas conjuntamente para produção de mapas de probabilidades de perdas.

4 MODELOS GEOGRAFICAMENTE PONDERADOS E MODELAGEM DINÂMICA

Na primeira parte deste capítulo, apresenta-se uma introdução aos modelos geograficamente ponderados ou *geographically weighted models* (GWMs). Os GWMs são uma família de modelos ponderados localmente que visam explorar a heterogeneidade espacial. Podem ser utilizados, por exemplo, como ferramenta para análise exploratória, regressão espacial local, para predição espacial, etc.

Neste trabalho utilizam-se tais modelos para realizar uma análise exploratória nos dados e, principalmente, para produção de mapas de probabilidade de perdas via regressão espacial geograficamente ponderada ou *geographically weighted regression* (GWR) – técnica que pertence à família dos GWMs. Conforme Rodrigues (2012), na regressão global utiliza-se um único modelo para descrever todo o fenômeno espacial sob análise; no entanto, os relacionamentos entre as variáveis podem não ser espacialmente constantes.

Segundo Gollini et al. (2015), a modelagem geograficamente ponderada ou *geographically weighted* (GW) envolve muitas técnicas que são aplicadas quando há uma provável heterogeneidade ou não estacionariedade no processo espacial sob análise.

Na segunda parte do capítulo, apresenta-se uma breve introdução às estruturas básicas e técnicas para modelagem dinâmica ou modelagem espaço-temporal das PNTs. Ênfase maior é dada às cadeias Markovianas que serão utilizadas neste estudo para construção de mapas de probabilidade de perdas no futuro.

4.1 INTRODUÇÃO AOS MODELOS GEOGRAFICAMENTE PONDERADOS

Os GWMs são uma coleção ou uma família de modelos estatísticos locais não estacionários. Conforme Bivand, Pebesma e Gómez-Rubio (2008), os valores dos coeficientes ajustados em modelos globais para todos os dados podem não representar as variações locais detalhadamente. Logo, os GWMs são adequados quando os dados não são plenamente descritos por modelos estacionários globais e, em algumas regiões, um ajuste local provê uma melhor descrição.

Tal técnica emprega uma janela de ponderação deslizando, onde se encontra uma coleção de modelos locais em cada localização alvo. Observa-se que não se procura uma variação local no espaço dos dados, mas move-se uma janela ponderada sobre os dados, estiman-

do um conjunto de coeficientes em cada ponto ajustado – em geral, os pontos de ajuste são os centroides das áreas (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008, p. 306).

A variação locacional dos coeficientes é um indicador de não estacionaridade. Os parâmetros do modelo ou saídas estimadas são mapeadas a fim de que a heterogeneidade espacial possa ser explorada e avaliada (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008; GOLLINI et al., 2015; LU et al., 2014).

Em suma, no GWM convencional, um modelo individual é calibrado em uma dada localização, todas as localizações vizinhas são ponderadas via função de *kernel* que decai com o aumento da distância. O tamanho da janela de ponderação é controlado via largura de banda (ou *bandwidth*) da função de *kernel*. Pequenas larguras de bandas conduzem a rápidas variações espaciais nos resultados (tendendo a um comportamento aleatório), enquanto que uma grande largura de banda conduz a um resultado que se aproxima dos resultados de modelos espaciais globais (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008, p. 167).

Conforme comentado no capítulo 3, neste trabalho, há uma modificação na matriz de ponderação do GWR convencional a fim de adaptar tal técnica ao problema das PNTs.

4.1.1 Definição da Matriz de Ponderação Convencional

A chave dos GWMs é a matriz de ponderação que define a dependência espacial nos dados (LU et al., 2014, p. 3). Conforme definida na seção 3.4.3, a matriz de ponderação $W(s_i)$ é uma matriz $n \times n$ (onde n é o número de amostras ou de observações) denotando a ponderação geográfica associada a cada ponto de observação i localizado na coordenada s_i do plano bidimensional.

Existem três elementos fundamentais para construção da matriz $W(s_i)$. A saber:

- ❖ tipo de distância;
- ❖ tipo de função de ponderação de *kernel*;
- ❖ largura de banda da função de *kernel*.

Com respeito à distância, em (29) apresenta-se a expressão para a família de distâncias de Minkowski d_{xy} de ordem p entre dois pontos $X = (x_1, x_2)$ e $Y = (y_1, y_2) \in \mathbb{R}^2$, ou seja, pontos localizados no plano bidimensional. Por exemplo, se $p=2$ obtém-se a distância Euclidiana; se $p=1$, obtém-se a distância Manhattan.

Em Lu et al. (2014) e em Gollini et al. (2015) há uma discussão detalhada a respeito das funções de *kernel* mais utilizadas. A saber: *kernel* Gaussiano (em (28)); *bi-square* (em (30)); *box-car*; exponencial, etc. Em todas as funções de *kernel*, o parâmetro largura de banda b controla a taxa de decaimento.

$$d_{xy} = \left(\sum_{i=1}^2 |x_i - y_i| \right)^{1/p} \quad (29)$$

Em (30) apresenta-se a função de *kernel bi-square* onde b é a largura de banda e d_{ij} é a distância entre os pontos i e j . Em (30), quando $d_{ij} \rightarrow 0$ acarreta $w_{ij} \rightarrow 1$ (máxima ponderação); quando $d_{ij} \rightarrow b$, $w_{ij} \rightarrow 0$ (mínima ponderação). Em todas as funções de *kernel* supracitadas, há uma tendência da redução de w_{ij} à medida que a distância d_{ij} aumenta.

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b} \right)^2 \right)^2 & , \text{ se } d_{ij} \leq b \\ 0 & , \text{ caso contrário} \end{cases} \quad (30)$$

Conforme apresentado nesta seção, a construção da matriz de ponderação baseia-se fundamentalmente na seleção da largura de banda b para uma função espacial de *kernel* ponderado que decai com a distância. Dessa forma, áreas vizinhas mais próximas exercem maior influência do que áreas vizinhas mais distantes.

Neste estudo propõe-se uma modificação na matriz de ponderação que não mais será construída via funções de *kernel*. Será construída via técnica para detecção de aglomerados, isto é, detecção de áreas que possuem atributos semelhantes; áreas essas que serão doravante consideradas áreas vizinhas. Neste trabalho utilizam-se os mapas auto-organizáveis de Kohonen ou *self organization maps* (SOM) para detecção de áreas com atributos semelhantes. No Apêndice A, apresenta-se uma breve introdução ao SOM.

4.1.2 Redefinição da Matriz de Ponderação para o problema das PNTs: a *Vizinhança Semântica*

No GWM convencional, nos trabalhos avaliados em análise espacial, áreas vizinhas são também áreas fronteiriças, isto é, possuem limites geográficos em comum ou estão próximas entre si. Ressalta-se que não necessariamente áreas vizinhas devem ser áreas fronteiriças. O ponto central do GWM convencional é que áreas mais próximas exercem maior influência do que áreas mais distantes, conforme enunciado pela primeira lei de Tobler – seção 3.2.1. Em Bivand, Pebesma e Gómez-Rubio (2008, p. 239-251), por exemplo, apresentam-se inúmeras possibilidades de vizinhanças entre subáreas para modelagem espacial. O número de vizinhos de cada subárea varia de acordo com o critério de vizinhança

escolhido. No entanto, todos os critérios são baseadas na proximidade entre vizinhos contíguos.

Neste trabalho define-se um novo estilo de vizinhança denominada *vizinhança semântica*. Por meio dela, subáreas distantes entre si, não contíguas, podem ser consideradas como vizinhas. Essa vizinhança baseia-se não mais na distância Euclidiana entre as áreas, mas na similaridade entre os atributos das mesmas. Logo, na *vizinhança semântica*, as áreas vizinhas são aquelas que possuem atributos semelhantes independente da distância entre elas.

A *vizinhança semântica* possui maior aderência ao problema das PNTs. Não há evidência de que áreas distantes devam exercer menor influência do que áreas próximas como ocorre no estilo de vizinhança tradicional baseado em funções de *kernel*. A *vizinhança semântica* expande a noção de vizinhança tradicional (baseada unicamente na distância Euclidiana) e possibilita que áreas distantes exerçam maior influência do que áreas próximas e vice-versa. Adicionalmente, este novo estilo de vizinhança possibilita o estudo de subáreas “desconhecidas”, que não receberam inspeções, via comparação dos atributos das mesmas com os atributos de áreas que receberam inspeções.

Os atributos das áreas avaliados neste trabalho são algumas variáveis socioeconômicas e variáveis técnicas da rede de distribuição. Para detecção das áreas com características semelhantes, utilizam-se os mapas auto-organizáveis de Kohonen. Outras técnicas igualmente adequadas para detecção de aglomerados ou *clusters* (como *k-means*, por exemplo) poderiam ser utilizadas. Uma das vantagens importantes do SOM é que o número de agrupamentos não precisa ser arbitrado *a priori*, como parâmetro de entrada. O próprio SOM reordena os vetores de peso no espaço n -dimensional das n variáveis de entrada de modo a obter os *clusters* conforme a disposição espacial do conjunto de amostras de entrada. Tais *clusters* podem ser agrupados *a posteriori* para produção do mapa de contexto⁸ o qual agrupa os *clusters* em classes representativas do problema em estudo.

Neste trabalho utiliza-se o mapa bidimensional de Kohonen, topologia hexagonal apresentado na Figura 7. Os elementos w_{ij} da matriz de ponderação $\mathbf{W}_{(n \times n)}$ definida na seção 3.4.3 são calculados após análise do mapa de contexto.

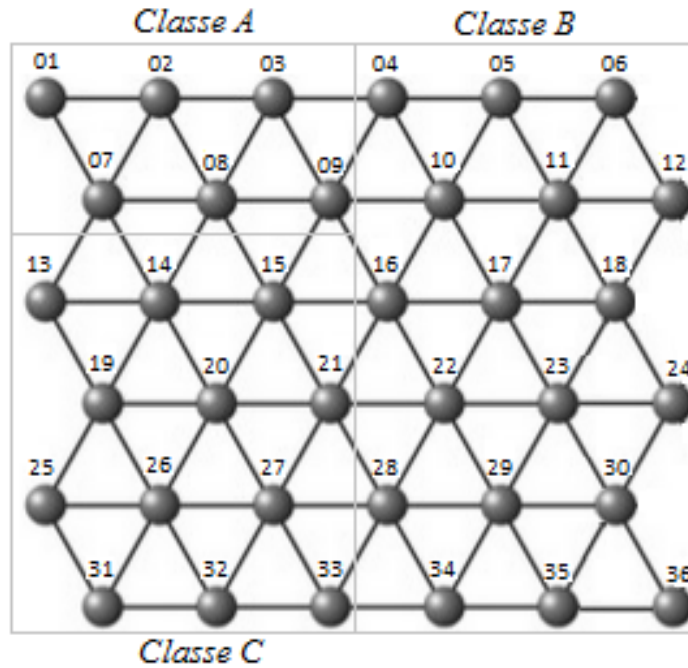
Sejam os neurônios da *classe A* indicados no mapa de contexto da Figura 7 – neurônios 1, 2, 3, 7, 8 e 9. Após as fases de treinamento e de operação do SOM, associa-se cada amostra de entrada $\mathbf{x}_k$ a cada um dos neurônios da grade neural. A cada neurônio da grade neural associa-se um vetor de pesos $\mathbf{w}$. Entre os neurônios pertencentes à *classe A*, por exem-

⁸ Ver Apêndice A para mais detalhes a respeito do mapa de contexto do SOM.

plo, o vetor de pesos $\mathbf{w}^{(j')}$ do neurônio j' que está associado a um maior número de amostras de entrada é nomeado neurônio representante da *classe A*.

O elemento w_{ij} da matriz de ponderação $\mathbf{W}_{(n \times n)}$ representa a ponderação entre as áreas i e j representadas pelos vetores de variáveis de entrada $\mathbf{x}_i$ e $\mathbf{x}_j$, respectivamente. Se ambas pertencem à mesma classe, ou seja, se ambas estiverem associadas a algum dos neurônios pertencentes à *classe A*, $w_{ij} \neq 0$; caso contrário, $w_{ij} = 0$. Supõe-se que as áreas i e j estejam associadas a algum dos neurônios da *classe A*. Dessa forma, w_{ij} representa um grau de aderência ou de semelhança das áreas i e j à *classe A* a qual é representada pelo neurônio j' .

Figura 7 – Mapa neural bidimensional de Kohonen topologia hexagonal 6×6 – 36 neurônios.



Fonte: Elaboração do próprio autor.

Em (31), w_{ij} é o complementar da maior norma Euclidiana da diferença entre as variáveis de entrada $\mathbf{x}_i$ e $\mathbf{x}_j$ e o vetor de pesos do neurônio j' representante da *classe A*, designado por $\mathbf{w}^{(j')}$. Portanto, entre as amostras de entradas $\mathbf{x}_i$ e $\mathbf{x}_j$, seleciona-se aquela que maximiza a norma Euclidiana da diferença entre elas e o vetor de pesos do neurônio representante da classe a qual ambas as amostras pertencem; portanto, seleciona-se aquele $\mathbf{x}_k$ onde $k = \{i, j\}$ que resulta no menor valor de w_{ij} .

Observa-se que quanto maior a afinidade ou a semelhança das entradas $\mathbf{x}_i$ e $\mathbf{x}_j$ ao vetor de pesos do neurônio representante da classe a qual ambas amostras pertencem, maior será o valor de $w_{ij} \in [0,1]$, sendo que $w_{ij} = w_{ji}$.

$$w_{ij} = 1 - \max \|x_k - w^{(j')}\|, \quad \text{sendo } k = \{i, j\} \quad (31)$$

4.1.3 Análise Exploratória: Resumo Estatístico Geograficamente Ponderado

O resumo estatístico geograficamente ponderado é um precursor vital para aplicação de um GWM (GOLLINI et al., 2015; LU et al., 2014).

Sejam os atributos ou variáveis z e y em uma subárea i onde w_{ij} são os elementos da matriz de ponderação $W(s)$. Define-se em (32)-(35) métricas para a análise estatística geograficamente ponderada ou GW. A saber: a média GW, o desvio padrão GW, o coeficiente de correlação de Pearson GW e a covariância GW, respectivamente.

$$\mu(z_i) = \frac{\sum_{j=1}^n w_{ij} z_j}{\sum_{j=1}^n w_{ij}} \quad (32)$$

$$s(z_i) = \sqrt{\frac{\sum_{j=1}^n w_{ij} (z_j - \mu(z_i))^2}{\sum_{j=1}^n w_{ij}}} \quad (33)$$

$$\rho(z_i, y_i) = \frac{c(z_i, y_i)}{s(z_i)s(y_i)} \quad (34)$$

$$c(z_i, y_i) = \frac{\sum_{j=1}^n w_{ij} [(z_j - \mu(z_i))(y_j - \mu(y_i))]}{\sum_{j=1}^n w_{ij}} \quad (35)$$

O desvio padrão GW em (33) destaca áreas com alta variabilidade para uma dada variável; áreas que podem ser melhor exploradas via regressão GW, por exemplo. A correlação GW em (34) provê uma avaliação preliminar da relação não estacionária entre a variável dependente e uma variável independente de uma regressão GW. A correlação GW também provê uma avaliação da multicolinearidade⁹ local entre duas variáveis independentes em uma regressão GW.

⁹ A multicolinearidade é um problema comum em regressões onde as variáveis independentes (ou explicativas) possuem relações lineares exatas ou aproximadamente exatas entre si.

4.1.4 Regressão Geograficamente Ponderada

O GWR decorre de modelos de regressão ponderados localmente e não paramétricos para ajuste de curvas e interpolação (GOLLINI et al., 2015, p. 37).

A regressão geograficamente ponderada ou GWR tradicional é descrita em Fotheringham, Brunson e Charlton (2002) e envolve a seleção da largura de banda para uma função espacial de *kernel* ponderado isotrópico, tipicamente um *kernel* Gaussiano com uma largura de banda fixa ou escolhida via técnica de validação cruzada (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008; BRUNSDON; FOTHERINGHAM; CHARLTON, 1996).

O GWR é adequado para modelar dados espaciais não estacionários. Por meio de funções de *kernel*, permite-se que os parâmetros do modelo variem espacialmente, produzindo superfícies não paramétricas (RODRIGUES, 2012). Conforme Bivand, Pebesma, Gómez-Rubio (2008, p. 307), tal técnica é útil para indicar regiões não estacionárias no mapa, isto é, regiões onde os coeficientes da regressão ponderados localmente afastam-se de seus valores globais. Como os dados são geograficamente ponderados, observações vizinhas exercem maior influência na estimação do conjunto local dos coeficientes da regressão do que observações não vizinhas.

Ressalta-se que, conforme discutido na seção 4.1.2, o processo para obtenção da matriz de ponderação foi modificado neste estudo a fim de implementar a *vizinhança semântica* que se adequa melhor ao problema das PNTs.

O GWR viabiliza a modelagem espacial de dados não estacionários. Um processo espacial é estacionário se sua distribuição de probabilidade é invariante no espaço. A estacionaridade é condição necessária para aplicação de modelos de regressão globais. Conforme discutido na seção 3.4.5.1, a condição da estacionaridade é muito restritiva e apenas em condições muito particulares pode-se afirmar que um modelo global seja capaz de representar adequadamente todas as regiões do processo espacial sob análise. (FOTHERINGHAM; BRUNSDON; CHARLTON, 2002).

A regressão geograficamente ponderada é a aplicação mais comum em modelos GW no qual se inclui o GWR. Na regressão GW, explora-se a variação espacial entre as variáveis dependentes e as variáveis independentes. Tal exploração consiste no mapeamento dos coeficientes estimados das regressões locais e os associam a (pseudo) *t*-valores a fim de determinar evidência de não estacionaridade. Em (36) apresenta-se a extensão do modelo de regressão global para o modelo de regressão GW:

$$y_i = \beta_{i0} + \sum_{k=1}^m \beta_{ik} x_{ik} + \epsilon_i, \quad (36)$$

sendo y_i a variável dependente na localidade i ; x_{ik} o valor da k -ésima variável independente na localidade i ; m é o número de variáveis independentes; β_{i0} é o parâmetro intercepto na localidade i ; β_{ik} é o coeficiente de regressão local para a k -ésima variável independente na localidade i e ϵ_i é o erro aleatório, independente e identicamente distribuídos com distribuição normal, média nula e variância σ^2 constante – $N(0, \sigma^2)$ – (FOTHERINGHAM; BRUNSDON; CHARLTON, 2002).

O modelo estima o relacionamento em um ponto i , onde os coeficientes de cada regressão são estimados via método dos mínimos quadrados ponderados conforme (37):

$$\hat{\beta}_i = (\mathbf{X}^T \mathbf{W}(s) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(s) \mathbf{y}, \quad (37)$$

sendo $\mathbf{X}$ a matriz das variáveis independentes com uma coluna de uns (1s) para o intercepto; $\mathbf{y}$ é o vetor de variáveis dependentes; $\hat{\beta}_i = (\beta_{i0}, \dots, \beta_{im})^T$ é o vetor dos $m + 1$ coeficientes de regressão locais; e $\mathbf{W}_i$ é a matriz que denota a ponderação geográfica para cada dado observado no ponto i de coordenadas s . No GWR convencional tal ponderação é conforme seção 4.1.1; para o problema das PNTs, a ponderação será conforme discutido na seção 4.1.2.

No GWR, a visualização das relações entre a variável dependente y_i e as k variáveis independentes (ou explicativas) x_{ik} pode ser feita via mapa com superfícies contendo as estimativas dos parâmetros locais β_{ik} e os erros padrão ϵ_i associados. Dessa forma, a determinação de padrões espaciais locais e o entendimento das causas dos mesmos tornam-se mais viáveis.

4.2 INTRODUÇÃO À MODELAGEM DINÂMICA

O GIS foi inicialmente construído para armazenar, organizar, recuperar e modificar informações sobre a distribuição espacial de recursos naturais e dados demográficos, por exemplo, de forma estática. A principal fonte de inspiração para construção desses *softwares* foram os mapas cartográficos que fornecem uma representação estática e bidimensional da realidade. No entanto, de acordo com Druck et al. (2004), em geral, os fenômenos espaciais são inerentemente dinâmicos; logo, as representações estáticas convencionais utilizadas em GIS não os capturam de forma completa.

Segundo Heppenstall et al. (2012), talvez não seja possível observar diretamente a dinâmica estocástica que rege um problema do mundo real, mas este processo produz consequências observáveis a partir das quais o sistema pode ser estudado e modelado.

Nesse contexto, conforme Druck et al. (2004), um dos grandes desafios da ciência da informação espacial é o desenvolvimento de técnicas e de abstrações que sejam capazes de fornecer representações mais próximas da realidade de fenômenos espaciais dinâmicos.

Um modelo espacial dinâmico é uma representação matemática (e computacional) de um processo pertencente ao mundo real no qual uma localização da superfície terrestre muda em resposta a variações de forças dirigidas (BURROUGH, 1998).

Os GIS são construídos supondo homogeneidade, uniformidade e universalidade do espaço, do tempo e do modelo matemático que representa o fenômeno. No entanto, conforme Couclelis (1997), para incorporar a dinâmica em processos espaciais com maior realismo, tais suposições rígidas devem ser flexibilizadas.

4.3 COMPONENTES PARA MODELAR A DINÂMICA DE FENÔMENOS ESPACIAIS

Os principais componentes para modelar um fenômeno espacial dinamicamente são apresentados na Tabela 7. O espaço pode ser classificado quanto à estrutura em distribuído regular, isto é, dividido em subáreas iguais (em geral retangulares) ou distribuído de forma diferenciada, irregularmente. Em geral, as vizinhanças são estacionárias, ou seja, iguais para todos os pontos. Essa estacionaridade pode ser superada personalizando as vizinhanças conforme a localização de cada ponto no espaço. Pode existir mais de uma função de transição, o tempo pode ser representado por intervalos variáveis e podem ser incluídas variáveis externas na modelagem – o que configura um sistema aberto.

Tabela 7 – Requisitos para modelar a dinâmica ou a variação espaço-temporal de eventos espaciais.

<i>Espaço</i>	<i>Estrutura</i>	Regular	Irregular
	<i>Propriedades</i>	Uniforme	Não Uniforme
<i>Vizinhança</i>		Estacionária	Não Estacionária
<i>Função de Transição</i>		Universal	Não Universal
<i>Tempo</i>		Regular	Irregular
<i>Sistema</i>		Fechado	Aberto

Fonte: Adaptado de Couclelis (1997).

Nas próximas seções abordam-se aspectos fundamentais da modelagem dinâmica: representação do fenômeno no espaço, no tempo e o modelo dinâmico a ser empregada para modelagem espacial das PNTs.

4.3.1 Representação do Espaço

O espaço é o conceito chave da geografia e da ciência da informação espacial. O espaço é classificado comumente em espaço absoluto e espaço relativo. O espaço absoluto também dito espaço Cartesiano ou Newtoniano é um espaço representado por um polígono fechado na superfície terrestre dentro do qual há coisas e eventos. O espaço relativo é o espaço constituído pelas relações espaciais, por fluxos e conexões entre coisas e os eventos (COUCLELIS, 1997).

Em geral, fenômenos geográficos, estabelecem relações entre si que independem das relações espaciais geométricas típicas tais como: relações topológicas, direcionais e de distância. Esses fenômenos têm relações como fluxos de pessoas, de materiais, conexões funcionais de influência, etc.

Os GIS atuais são abstrações dos mapas relacionados à noção cartográfica do espaço absoluto. No entanto, em processos dinâmicos, a noção do espaço relativo é fundamental para representar fluxos e conexões entre entidades do sistema (DRUCK et al., 2004).

4.3.2 Representação do Tempo

O tempo juntamente com o espaço é um dos pilares da modelagem dinâmica de fenômenos espaciais. O tempo pode ser classificado nos seguintes aspectos: ordem temporal, variação temporal e granulidade conforme Tabela 8.

A ordem temporal refere-se ao modo como o tempo flui que pode ser: linear, ramificado ou cíclico. Na ordem temporal linear, o tempo flui sequencialmente, ou seja, cada instante de tempo tem apenas um sucessor e um antecessor. Na ordem temporal ramificada; cada instante de tempo pode ter múltiplos instantes sucessores e antecessores. A ordem temporal cíclica é utilizada para modelar eventos recorrentes.

Em relação à variação temporal há duas possibilidades: tempo contínuo e discreto. Uma variável temporal contínua é usada em processos cujas medidas de tempo têm níveis arbitrários, isto é, pode-se aumentar a precisão da variável tanto quanto se queira. A variável discreta por outro lado é usada quando o tempo é medido em certos intervalos e a variação é descontínua entre os mesmos.

Tabela 8 – Classificação das principais estruturas temporais para modelagem dinâmica.

<i>Ordem Temporal</i>	<i>Variação Temporal</i>	<i>Granulidade</i>
Linear	Discreta	Instante
Ramificada	Contínua	Intervalo
Cíclica	-	Período

Fonte: Adaptado de Druck et al. (2004).

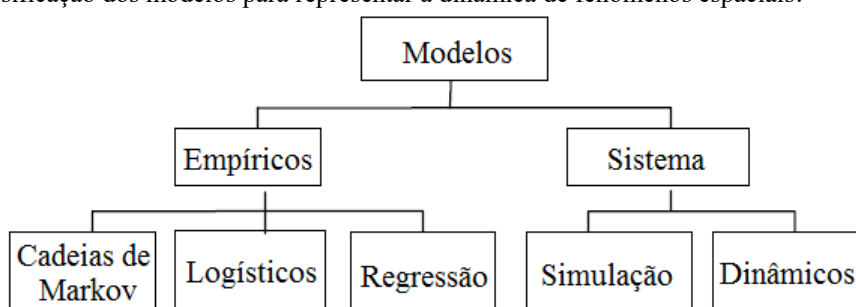
A granulidade refere-se à menor duração de tempo para cada sistema. Um instante de tempo representa um ponto particular no tempo; um intervalo é o tempo entre dois instantes e um período é uma sequência de intervalos de tempo.

4.3.3 Modelos de representação

Modelos espaciais dinâmicos representam a evolução de fenômenos espaciais ao longo do tempo. Pretende-se responder as questões chave “onde”, “porque” e “quando”. Um modelo que responde a tais questões é capaz de localizar o fenômeno no espaço geográfico, de explicar o porquê do mesmo estar distribuído espacialmente dessa maneira e ainda prever sua evolução espaço-temporal ou a dinâmica que rege o mesmo.

Um modelo é composto basicamente por três elementos: variáveis, relacionamentos e processos. A ênfase dada a cada um desses elementos gera a diferenciação entre os modelos (DRUCK et al., 2004).

Na Figura 8 mostra-se a classificação de alguns modelos em empíricos e de sistemas. Conforme Heppenstall et al. (2012), modelos empíricos focalizam os relacionamentos entre as variáveis do modelo supondo que os relacionamentos passados permanecerão no futuro. Os modelos de sistemas são baseados em descrições matemáticas de processos que interagem entre si. Enfatizam-se, nesse caso, as interações entre todos os componentes do sistema.

Figura 8 – Classificação dos modelos para representar a dinâmica de fenômenos espaciais.

Fonte: Adaptado de Druck et al. (2004).

4.4 MODELOS EMPÍRICOS

Os modelos empíricos são compostos basicamente por três elementos: configuração inicial, função de transição e configuração de saída. Em geral, a configuração inicial é obtida via dados históricos do fenômeno sob estudo. Para representar a função de transição, podem ser utilizadas equações diferenciais que incluem pelo menos um termo de derivada no tempo. Nesse caso, o processo é determinístico. Caso sejam utilizadas variáveis aleatórias para explicar o processo o mesmo é dito estocástico-probabilístico.

Os modelos empíricos são caracterizados pela simplicidade dos modelos, pelo número reduzido de variáveis e pelo baixo tempo computacional em sua execução. Devido a isso, são eficientes em fazer previsões; no entanto, são limitados para explicar aspectos causais do fenômeno.

Seguem três modelos empíricos: cadeias de Markov, modelos de regressão e modelos logísticos de difusão.

4.4.1 Cadeias de Markov de Tempo Discreto

As cadeias de Markov são um caso particular de um processo estocástico cujos estados de transição são discretos (ZAPATA, 2010). São descritas conforme (38):

$$\Pi_{(t+1)} = \mathbf{P}^t \cdot \Pi_{(t)}, \quad (38)$$

sendo $\Pi_{(t)}$ é o vetor de estados do sistema no tempo t , $\Pi_{(t+1)}$ é o vetor de estados do sistema após o intervalo $t + 1$ e $\mathbf{P}^t$ é a matriz de transição que contém as probabilidades de transição entre os estados do sistema. A matriz de transição representa a probabilidade de um determinado estado i permanecer o mesmo ou mudar para o estado j a cada passo de tempo t . As probabilidades de transição são obtidas a partir de amostras relativas a um intervalo de tempo.

As cadeias Markovianas de primeira ordem admitem que o estado futuro do sistema dependa apenas do estado presente e das probabilidades de transição. Tal modelo não ignora o passado; no entanto, considera que toda informação do passado esteja concentrada no estado presente do sistema. Caracteriza-se como um processo estacionário, pois as probabilidades de transição não mudam com o tempo.

As principais vantagens desse modelo envolvem a simplicidade operacional e matemática, a facilidade de implementação com dados provenientes de GIS e por não necessitar de uma grande quantidade de dados.

Em relação à implementação em análise espacial, através das cadeias de Markov é possível localizar os fenômenos no espaço e prever onde os mesmos ocorrerão no futuro – admitindo que o processo seja estacionário. Todavia, são limitadas em explicar o porquê dos fenômenos estarem localizados em determinada região do espaço geográfico. Ademais, tal modelo não suporta a inclusão direta de variáveis exógenas.

As limitações das cadeias Markovianas serão contornadas, neste trabalho, com o uso combinado das mesmas com os modelos de regressões espaciais de dados por pontos GAM (seção 3.3.4.5) e de dados agregados por áreas GWR (seção 4.1.4) para modelar as PNTs a nível espaço-temporal.

4.4.2 Metodologia para Estimação Espaço-Temporal das Perdas Não Técnicas via GAM e Markov

Nesta seção apresenta-se a metodologia para estimar a variação espaço-temporal das PNTs, considerando o lugar onde elas ocorrem. A metodologia segue duas etapas: na primeira, estima-se um mapa de probabilidades de perdas por subáreas da cidade no momento presente. Este mapa é produzido a partir de um GAM (vide seção 3.3.4.5) com a inclusão de variáveis socioeconômicas e da rede de distribuição. Na segunda etapa, para estimar as perdas no futuro, gera-se um novo mapa de perdas por meio das cadeias de Markov que consideram possíveis mudanças no valor presente das probabilidades de perdas. Tais mapas de perdas indicam as subáreas ou setores censitários da cidade mais vulneráveis a ocorrência de PNTs.

Neste estudo pretendem-se avaliar a localização das PNTs no presente e no futuro por subáreas da cidade, explicitando quais variáveis criam um ambiente propício às perdas.

4.4.2.1 Vulnerabilidade às Perdas Não Técnicas no Presente

Segundo Zapata (2010), em alguns casos, é possível que o espaço de estado do fenômeno aleatório em estudo seja contínuo; no entanto, interessa considerar apenas alguns estados particulares. Nesse estudo, a vulnerabilidade das subáreas do município às PNTs será representada por três estados de perda: *regular (R)*, *atenção (A)* e *crítico (C)*. Tais estados representam vulnerabilidade baixa, intermediária e alta às PNTs, respectivamente.

Uma base de dados *casos-controles* é construída – vide seção 3.3.4. Essa base corresponde aos dados de entrada de um GAM a partir do qual se obtém as probabilidades $P(s, x)$ relacionadas ao risco de PNTs na localização s .

Após execução do GAM, a cada ponto da base de dados de *casos-controles* associa-se uma probabilidade $P(s, x)$ referente ao risco de PNTs. O estado de perda atual de cada

subárea do município é determinado por meio do agrupamento e classificação das probabilidades contidas em cada subárea.

Sejam m pontos da base de dados *casos-controles* que estão contidos em uma determinada subárea. Determina-se o percentual desses m pontos que estão nos estados: *regular* ($P(s, x) \in [0; 0,1]$); *atenção* ($P(s, x) \in]0,1; 0,3]$) e *crítico* ($P(s, x) \in [0,3; 1]$). Tais percentuais representam uma probabilidade para cada estado de perda e estão encerradas em um vetor $[\%P_R \quad \%P_A \quad \%P_C]_{1 \times 3}^A$ associado a cada subárea A . Os termos $\%P_R$, $\%P_A$ e $\%P_C$, representam o percentual das m probabilidades da subárea A que estão nos estados de perdas *regular*, *atenção* e *crítico*, respectivamente. O estado de perda atual do setor A é determinado via classe modal, isto é, determina-se o estado de perda no qual as probabilidades são mais frequentes.

Para fins ilustrativos, seja a zona sul de uma cidade qualquer dividida por suas subáreas, como se mostra na Figura 9. A subárea em vermelho contém três cruzeiros em branco que representam UCs da base de dados *casos-controles* ($m = 3$). Supõem-se, por exemplo, que as probabilidades associadas a cada um desses pontos resultantes da execução do GAM sejam: 0,02 (estado *regular*), 0,09 (estado *regular*) e 0,28 (estado de *atenção*). Portanto, $2/3$ e $1/3$ das probabilidades estão nos estados de perda *regular* e de *atenção*, respectivamente, e $0/3$ das probabilidades encontram-se no estado *crítico*. As probabilidades para cada estado de perda para a subárea em vermelho da Figura 9 estão encerradas em um vetor $[2/3 \quad 1/3 \quad 0]$. Conclui-se, portanto, que essa subárea encontra-se no estado de perda atual *regular*, pois as probabilidades GAM são mais frequentes neste estado de perda. Observa-se que a subárea em vermelho possui limites em comum com outras quatorze subáreas vizinhas a ela (destacadas em verde).

Figura 9 – Mapa da zona sul de uma cidade qualquer com destaque para uma de suas subáreas (em vermelho) que contém três pontos da base de dados *casos-controles* e quatorze subáreas vizinhas (em verde).



Fonte: Elaboração do próprio autor.

4.4.2.2 Vulnerabilidade às Perdas Não Técnicas no Futuro

Conforme Bailey e Gatrell (1995), não existe apenas uma única solução para modelar fenômenos espaciais dinâmicos. Neste trabalho, utilizam-se as cadeias de Markov para estimar subáreas ou setores censitários vulneráveis às PNTs no futuro. Algumas considerações são feitas para viabilizar o uso das cadeias de Markov para modelar a variação espaço-temporal das PNTs. Para um horizonte de estudo de curto prazo, é viável considerar que as PNTs são estacionárias. As perdas são influenciadas pelas condições socioeconômicas da cidade que se modificam lentamente. Dessa maneira, o fenômeno das perdas é considerado aproximadamente estacionário e os estados do passado recente assemelham-se ao presente.

4.4.2.3 Matriz de Transição

De posse do estado de perda atual de cada subárea do município, utilizam-se as matrizes de transição para estimar os estados de perda no futuro.

Na Figura 10 mostra-se a estrutura genérica da matriz de transição com todos os elementos considerando os estados de perda: *regular*, *atenção* e *crítico*. P_{ij} é a probabilidade de uma subárea permanecer no mesmo estado de perda (para $i = j$) ou de mudar de estado (para $i \neq j$) após uma transição fixa de tempo de um ano.

A matriz de transição para a subárea A é obtida a partir de suas probabilidades de perdas e das probabilidades de perda de seus vizinhos. A taxa de sucesso (TS) das inspeções para a subárea A será utilizada como uma aproximação para a probabilidade de perda, conforme equação (39).

$$TS_A(Perdas) = 100 \times \frac{N^{\circ} UCs Irregulares Encontradas}{N^{\circ} UCs Inspeccionadas}, \quad (39)$$

sendo TS_A a *Taxa de Sucesso* percentual das inspeções na subárea ou setor censitário A . A *Taxa de Sucesso* representa uma estimativa para a probabilidade anual das perdas na subárea A .

Figura 10 – Matriz de transição dos estados de perdas por subáreas do município.

Estados de Perdas	Regular (R)	Atenção (A)	Crítico (C)
Regular (R)	P_{RR}	P_{RA}	P_{RC}
Atenção (A)	P_{AR}	P_{AA}	P_{AC}
Crítico (C)	P_{CR}	P_{CA}	P_{CC}

Fonte: Elaboração do próprio autor.

Deseja-se obter os elementos das matrizes de transições para todas as subáreas. Supõe-se, por exemplo, que a subárea destacada em vermelho na Figura 9 possua probabilidades anuais de perdas definidas a partir da equação (39) de 0,05 (estado *regular*); 0,09 (estado *regular*) e 0,27 (estado de *atenção*) nos anos de 2009, 2010 e 2011, respectivamente. Entre as transições anuais possíveis (de 2009 para 2010 e de 2010 para 2011); em uma delas, o estado *regular* permaneceu o mesmo – transição de 2009 para 2010. Na outra transição possível, de 2010 para 2011, o estado *regular* muda para o estado de *atenção*. Observa-se que o ano de 2012 é reservado para avaliar o desempenho da metodologia.

No exemplo da Figura 9, há quinze subáreas envolvidas (a subárea alvo e as quatorze subáreas vizinhas) totalizando trinta transições, sendo duas por subárea. Contabiliza-se por subárea, o número de transições que o estado de perda i permaneceu o mesmo ou mudou para o estado j após uma transição anual fixa de tempo.

Após inserção e normalização das probabilidades individuais, tem-se a matriz de transição P em (40) para a subárea destacada em vermelho na Figura 9. Observa-se que a soma de cada uma das linhas da matriz de transição deve ser unitária (HASTING, 1973).

$$P = \begin{bmatrix} 3/5 & 1/10 & 3/10 \\ 0 & 0 & 1 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} \quad (40)$$

4.4.2.4 Determinação do Estado de Perda Futuro

A previsão do estado de perda futuro é encontrada por meio da multiplicação matricial do vetor de estados de perdas atuais pela matriz de transição por subárea. O resultado é um vetor que contém as probabilidades de cada subárea estar no futuro nos estados de perda *regular*, *atenção* e *crítico*. A maior probabilidade do vetor resultante determina o estado de perda previsto para a subárea avaliada. Em (41), estima-se o estado de perda futuro para a subárea destacada no exemplo da Figura 9 considerando o vetor de probabilidades de perdas no presente (seção 4.4.2.1) e a matriz de transição em (40).

$$\begin{bmatrix} 2/3 & 1/3 & 0 \end{bmatrix} \begin{bmatrix} 3/5 & 1/10 & 3/10 \\ 0 & 0 & 1 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} = \begin{bmatrix} 6/15 & 1/15 & 8/15 \end{bmatrix} \quad (41)$$

Nota-se que a subárea destacada na Figura 9 possui probabilidades de estar em 2012 nos estados de perda *regular*, *atenção* e *crítico* de 6/15, 1/15 e 8/15, respectivamente.

Conclui-se que o estado de perda dessa subárea mudará de *regular* (estado de perda atual) para *crítico* (estado de perda futuro).

Neste ponto, pode-se resumir a metodologia apresentada nesta seção para estimar a variação espaço-temporal das PNTs via mapas de probabilidade de perdas.

- ❖ *Passo 1:* Constrói-se a base de dados *casos-controles* a partir dos dados de inspeções georreferenciadas e da base de clientes da concessionária. Essa base é associada às variáveis socioeconômicas do censo demográfico IBGE (2010) e são os dados de entrada para a regressão espacial GAM.
- ❖ *Passo 2:* Executa-se o GAM e constrói-se o mapa de probabilidades de perdas para o presente.
- ❖ *Passo 3:* Determina-se o estado atual de perda de cada subárea ou setor censitário utilizando as probabilidades calculadas no *passo 2* por meio da classe modal. Os estados de perdas atuais de cada subárea são determinados a partir do percentual de probabilidades do mapa que estão contidas em cada subárea e que estão nos estados de perda *regular*, *atenção* e *crítico*. O estado de perda atual da subárea *A* é determinado pela classe modal, isto é, pelo estado de perda no qual as probabilidades GAM são mais frequentes.
- ❖ *Passo 4:* Estima-se a matriz de transição *P* por subárea. Ela é construída a partir de (39) avaliando-se as transições dos estados de perdas anuais por subáreas e de seus respectivos vizinhos.
- ❖ *Passo 5:* A previsão do estado de perda futuro por subárea é encontrada através da multiplicação matricial entre o vetor dos estados de perdas atuais (obtido no *passo 3*) e a matriz de transição (obtida no *passo 4*). O resultado é um vetor 1×3 que contém as probabilidades de cada subárea estar futuramente nos estados de perda *regular*, *atenção* e *crítico*. O estado de perda associado a maior probabilidade desse vetor resultante determina o estado de perda previsto para a subárea avaliada.

4.4.3 Modelos de Regressão

Os modelos de regressão estabelecem relações estatísticas entre o fenômeno sob estudo e as variáveis independentes que exercem influência sobre o mesmo. Logo, o modelo permite a inclusão de variáveis exógenas. Conforme Druck et al. (2004), isso contribui para o entendimento do fenômeno; no entanto, não é suficiente para explicá-lo, visto que, a identificação de um relacionamento estatístico entre duas variáveis não estabelece um relacionamento causal entre ambas (HUFF, 1973). Por exemplo, através de um modelo de regressão pode-

se identificar que há relação entre a criminalidade e a falta de saneamento básico nos domicílios de uma cidade; todavia, o modelo de regressão não explica os mecanismos que ligam essas variáveis.

O modelo de regressão estabelece um relacionamento linear entre variáveis dependentes e independentes conforme (42):

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon, \quad (42)$$

sendo y a variável dependente, x_i são as k variáveis independentes ou explicativas, β_i são os coeficientes da regressão e ε é o componente de erro.

A dimensão espacial é incorporada ao modelo de regressão através da integração do mesmo com um GIS. Uma limitação desse modelo é que o mesmo se aplica somente a processos estacionários (LAMBIN, 1994).

Conforme mencionado na seção anterior, neste trabalho, implementam-se dois modelos de regressão espacial. A saber: GAM e o GWR. No primeiro, as PNTs são modeladas no espaço da cidade a nível de padrões de pontos; no segundo, as perdas são agregadas por subáreas da cidade denominadas setores censitários.

4.4.4 Modelos Logísticos de Difusão

Os modelos logísticos são comumente utilizados para descrever fenômenos onde as variáveis apresentam inicialmente variações em ritmo lento, a seguir o ritmo aumenta, na sequência, volta a reduzir até que um dado nível de saturação seja alcançado. As interações temporais entre as variáveis do modelo são descritas conforme (43):

$$\frac{dQ}{dt} = rQ \left[\frac{U - Q}{U} \right] \quad (43)$$

sendo Q a variável de um fenômeno que cresce ao longo do tempo t ; r a taxa de crescimento e U é uma função de crescimento. Tais modelos enfatizam a velocidade do processo e permitem a inclusão de variáveis causais. Dentre os modelos baseados em funções logísticas destacam-se os modelos de difusão.

Em Druck et al. (2004) listam-se os principais elementos de um modelo espacial de difusão: meio ambiente (isotrópico ou heterogêneo), tempo (contínuo ou discreto), item a

ser difundido (material, pessoas, doença), locais de origem, locais de destino e caminhos a serem percorridos.

Tais elementos interagem entre si por meio de um mecanismo composto por quatro estágios: estágio inicial (início do processo de difusão), estágio de difusão (processo de espalhamento), estágio de condensação (redução do ritmo de espalhamento) e o estágio de saturação (desaceleração ou encerramento do processo de difusão).

Modelos de difusão não explicam as causas do fenômeno espacial, apesar de permitirem a incorporação de variáveis exógenas (variáveis socioeconômicas, por exemplo). No entanto, preveem o comportamento futuro dos eventos. A dimensão espacial é incorporada via interação do modelo com um GIS.

4.5 MODELOS DE SISTEMAS

Os modelos de sistemas buscam descrever o sistema como um todo por meio da representação das interações que existem entre os componentes do mesmo. Representam a dimensão espacial de maneira mais eficiente e precisa como as relações de vizinhança, por exemplo. Seguem duas classes de modelos de sistemas: modelos de simulação de ecossistemas e modelos de simulação da dinâmica espacial.

4.5.1 Modelos de Simulação de Ecossistemas

Os modelos de ecossistemas visam imitar o comportamento do sistema, enfatizando as interações entre os componentes do mesmo. São baseados na composição de ecossistemas complexos que se relacionam em geral por equações diferenciais.

A construção de um modelo de simulação requer que os aspectos relevantes que afetam o fenômeno estejam bem integrados, que seus relacionamentos funcionais estejam bem representados e que o modelo possa prever as mudanças ao longo do tempo (DRUCK et al., 2004).

Esses modelos são adequados para representar processos não estacionários; no entanto, têm a limitação de considerar o espaço como uma entidade homogênea (LAMBIN, 1994).

4.5.2 Modelos de Simulação da Dinâmica Espacial

Os modelos de simulação dinâmica espacial são uma extensão de modelos de ecossistemas para melhor representar a heterogeneidade espacial e incluir processos de tomada de decisão.

Em uma das abordagens mais difundidas, o espaço é representado por uma matriz de células, nas quais, aplicam-se equações matemáticas simultaneamente. Cada célula está conectada às células adjacentes (ou vizinhas) de forma que é possível estabelecer um intercâmbio com as mesmas. Por exemplo, se uma célula possui três vizinhos que estão em um determinado estado i é provável que o estado desta célula também seja i . No entanto, este raciocínio pode ser aperfeiçoado via regras de transição. Tais regras de tomada de decisão são baseadas em abstrações análogas às que ocorrem na mente humana.

4.5.2.1 Autômatos Celulares

Em Wolfram (1994) foram propostos os autômatos celulares como sendo modelos matemáticos simples de sistemas naturais compostos por malhas ou grades de células idênticas e discretas onde cada célula possui um valor pertencente a um conjunto finito. Tais valores evoluem em passos de tempo discreto, respeitando regras determinísticas.

Os autômatos celulares são idealizações discretas das equações diferenciais parciais comumente utilizadas para descrever os sistemas naturais. A partir dessa natureza discreta, pode-se fazer uma analogia com computadores digitais. Nesse contexto, os autômatos podem ser vistos como computadores de processamento paralelo de natureza simplificada (MELO, 2010, p. 50-51).

Nos autômatos celulares, o espaço é representado como uma matriz de célula, onde cada uma dessas células pode assumir diferentes estados ou valores ao longo do tempo. O tempo varia em intervalos discretos. O estado das células muda simultaneamente em função do seu próprio estado, do estado das demais células vizinhas e conforme um conjunto de regras de transição (CARREÑO, 2008; WOLFRAM, 1994).

4.5.2.2 Sistemas Multiagentes

Os sistemas multiagentes (SMAs) são uma expansão do modelo de autômato celular clássico com o intuito de integrar fatores ambientais e socioeconômicos para representar a dinâmica espacial de fenômenos urbanos.

Devido à complexidade do comportamento humano e às peculiaridades do ambiente urbano, é extremamente difícil entender e modelar padrões de crime – como o furto de

energia, por exemplo. No entanto, um maior entendimento do processo e dos motivadores por trás do crime são essenciais para que os pesquisadores sejam capazes de modelá-lo adequadamente.

Antes da década de 1970, os pesquisadores em criminologia eram largamente dominados pelos estudos das vítimas, da lei e dos criminosos e então omitiam um elemento fundamental: o lugar onde os crimes ocorrem. A *criminologia ambiental* surge como uma disciplina para estudar a variação espacial do crime e visa compreender as razões subjacentes para tal variação (HEPPENSTALL et al., 2012, p. 411-414).

Os SMAs são entidades que se adaptam ao meio, reagem a ele e provocam mudanças neste meio (MELO, 2010, p. 22). Eles apresentam vantagens para modelar a complexidade urbana, pois proporcionam um modelo útil e potente, capaz de modelar processos complexos de interação entre pessoas e ambientes urbanos que eles mesmos criam. Fatores subjetivos, reais ou simbólicos podem ser incorporados (MELO, 2010, p. 26-27; WOOLDRIDGE, 2002).

Os SMAs com agentes reativos são comumente utilizados. Tais sistemas são constituídos por um grande número de agentes. São muito simples. Não têm inteligência ou representação de seu ambiente e interagem utilizando um comportamento ação e reação. A inteligência surge conforme tais sistemas trocam de interações entre agentes e o ambiente. Em outras palavras, os agentes não são inteligentes individualmente, mas o comportamento global o é (MELO, 2010, p. 25).

4.6 CONSIDERAÇÕES FINAIS

Neste capítulo apresentaram-se sucintamente os modelos de regressão espacial locais geograficamente ponderados, com ênfase ao GWR para modelar as PNTs agregadas por subáreas. Tais modelos levam em consideração os atributos socioeconômicos e da rede de distribuição de cada subárea.

Na segunda parte do capítulo, abordaram-se sucintamente os principais aspectos para modelar a dinâmica de fenômenos espaciais. Ênfase maior foi dada às cadeias Markovianas que serão combinadas individualmente com os modelos de regressão espacial GAM (para padrões de pontos) e GWR (para dados agregados por áreas) para estimar a variação espaço-temporal das PNTs.

No capítulo subsequente, implementam-se as técnicas de análise espacial e de modelagem dinâmica ou espaço-temporal abordadas nos capítulos 3 e 4, respectivamente, a

partir de dados de inspeções reais de uma cidade de porte médio do interior paulista para produção de mapas de probabilidades de perdas.

5 ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS

Neste capítulo, implementam-se técnicas de análise espaço-temporal abordadas nos capítulos 3 e 4. São utilizados dados de inspeções reais provenientes de um município de porte médio do interior paulista.

O presente capítulo está dividido em três estudos. Primeiramente, é realizada uma análise das características do município em estudo e uma análise exploratória nos dados de inspeções, dados socioeconômicos e da rede de distribuição. No contexto da análise exploratória, emprega-se, por exemplo, a análise espacial de pontos onde as PNTs são representadas como pontos no espaço geográfico do município para construção de mapas de densidade de perdas. No segundo estudo, implementa-se um modelo de regressão aditivo generalizado (GAM) para construção de mapas de probabilidade de perdas. Neste modelo, levam-se em consideração as variáveis socioeconômicas e da rede de distribuição que caracterizam o lugar onde as perdas ocorrem. No estudo final, estima-se a variação espaço-temporal das PNTs. Para tal, utilizam-se as regressões espaciais GAM e GWR; ambas, combinadas individualmente com as cadeias de Markov para produção de mapas de probabilidades de perdas no tempo presente e no futuro.

5.1 CONSIDERAÇÕES INICIAIS

Os estudos realizados neste capítulo foram executados utilizando dados oriundos de um município de porte médio do interior do Estado de São Paulo, Brasil, com aproximadamente 200 mil habitantes dos quais 98% são urbanos e 2% são rurais e cuja área total ultrapassa 550 km². O produto interno bruto (PIB) da região é de quase três bilhões de reais e a renda *per capita* é de quase 900 reais. A principal atividade econômica é o setor de serviços com participação também da indústria e, em menor percentagem, do agronegócio (SISTEMA ESTADUAL DE ANÁLISE DE DADOS – SEADE, 2010).

Neste estudo utilizam-se variáveis do censo demográfico. Tais variáveis são disponibilizadas por setores censitários¹⁰. Os setores analisados neste estudo pertencem à região urbana do município em estudo; a parte rural foi desconsiderada. As distribuidoras inspecionam prioritariamente a população urbana do município devido à maior concentração das

¹⁰ Neste capítulo, a expressão “setores censitários” será frequentemente substituída pela palavra mais abrangente “subáreas” para fins de maior clareza.

PNTs nesta região, à maior densidade populacional, ao menor custo financeiro despendido com inspeções e devido à maior facilidade de acesso das equipes de inspeções (COMETTI; VAREJÃO, 2005).

As simulações deste trabalho foram realizadas majoritariamente no software R[®] versão 2.15.3 e algumas no *software* livre Terraview versão 4.2.0 (TERRAVIEW, 2010). O R[®] é um GIS e um *software* livre para computação estatística e gráfica. É uma implementação da linguagem S. Possui uma série de métodos padrões e inovadores para análise estatística. Nele estão implementadas bibliotecas conhecidas para análise numérica; ademais, o mesmo pode ser conectado a códigos escritos nas linguagens de programação S, C, C++, Fortran e Java (PETERNELLI; MELLO, 2011; R DEVELOPMENT CORE TEAM, 2011; VENABLES et al., 2012).

5.2 DADOS DE ENTRADA

Neste estudo utilizam-se dados reais do histórico de inspeções georreferenciado e da rede de distribuição provenientes de uma distribuidora de energia elétrica brasileira. Utilizam-se também dados oriundos do último censo demográfico do Brasil (IBGE, 2010). Os dados anuais de inspeções referem-se aos anos de 2009 a 2012. Na Tabela 9 há um resumo anual com respeito ao número total de inspeções realizadas, de UCs irregulares (ou com PNTs) e a taxa anual de sucesso das inspeções – equação (1). As UCs irregulares estão distribuídas em sua maioria nas classes de consumidores residencial (*RES*), comercial (*COM*) e industrial (*IND*) – as percentagens entre parênteses correspondem ao percentual de UCs irregulares que se encontram nas classes de consumidores residenciais, comerciais e industriais.

Na Tabela 9 observa-se também que a maior parte das UCs com PNTs pertencem à classe de consumidores residenciais sendo de 80,3%, 64,4%, 67,9% e 83,3% de 2009 a 2012, respectivamente. No entanto, as UCs com PNTs das classes comercial e industrial, apesar de existirem em menor número, causam maior prejuízo financeiro às concessionárias, porque são constituídas majoritariamente por grandes consumidores de energia elétrica.

Observa-se também a inconstância no número de UCs inspecionadas anualmente. Em geral, um ano com muitas inspeções é seguido por um ano com poucas inspeções e vice-versa. Em 2010, por exemplo, há pouco mais de um terço das inspeções realizadas em 2009; em 2012, há pouco mais da metade das inspeções realizadas em 2011. Não há um critério padrão para definir nem a quantidade, nem a localização das UCs inspecionadas. As inspeções, no caso específico do município sob análise, são motivadas por denúncias de leituristas

e da população. Ademais, são feitas análises preliminares em planilhas eletrônicas com regras heurísticas elementares baseadas no conhecimento empírico de especialistas em perdas. Tais regras podem ser, por exemplo, uma UC com consumo (em kWh) nulo ou quase nulo por meses subsequentes ou então uma queda considerável e permanente na curva de consumo mensal. Selecionam-se UCs com perfis de consumo atípicos para serem alvos das inspeções.

Tabela 9 – Distribuição das inspeções e irregularidades em UCs das principais classes consumidoras.

<i>Ano</i>	<i>UCs Inspeccionadas</i>	<i>UCs Irregulares</i>	<i>Taxa de Sucesso das Inspeções (%)</i>	<i>Irregularidades Encontradas nas Principais Classes Consumidoras</i>			
				<i>RES</i>	<i>COM</i>	<i>IND</i>	<i>Demais Classes</i>
2009	3.068	173	5,6	139 (80,3%)	20 (11,6%)	09 (5,2%)	05 (2,9%)
2010	1.298	101	7,8	65 (64,4%)	20 (19,8%)	05 (5,0%)	11 (10,9%)
2011	4.613	505	10,9	343 (67,9%)	101 (20,0%)	16 (3,2%)	45 (8,9%)
2012	2.387	461	19,3	384 (83,3%)	37 (8,0%)	06 (1,3%)	34 (7,4%)
Total	11.366	1.240	10,9	931 (75,1%)	178 (14,4%)	36 (2,9%)	95 (7,6%)

Fonte: Elaboração do próprio autor.

Em último, destaca-se o crescimento anual da taxa de sucesso das inspeções – vide equação (1), seção 2.2. Isso demonstra o aperfeiçoamento dos critérios internos empíricos da concessionária de energia na seleção de UCs a serem inspeccionadas.

No Tabela 10 classificam-se e contabilizam-se as inspeções da Tabela 9 conforme diagnóstico da concessionária de energia para as inspeções realizadas aos longo de quatro anos – de 2009 a 2012. As irregularidades mais recorrentes (citadas na Tabela 1, seção 2.2) são: desvio fora da medição, autorreligação, diversas modalidades de fraudes no medidor e medidor avariado.

Observa-se que as UCs da base de dados de perdas não estão georreferenciadas. Devido a isso, as UCs são agrupadas pelos transformadores aos quais se conectam. Isso, pois, ao contrário das UCs, os transformadores da rede de distribuição estão georreferenciados, viabilizando a utilização das técnicas de análise espacial propostas neste trabalho.

Na Tabela 11 é apresentado um resumo e a descrição dos dados de entrada empregados neste trabalho para modelagem espacial das PNTs. Além dos dados de inspeções

supracitados, há os dados socioeconômicos e da rede de distribuição. Tais dados estão classificados em três grupos: variáveis socioeconômicas (advindas do último censo demográfico IBGE (2010)), variáveis técnicas (advindas da distribuidora de energia) e variáveis derivadas (produzidas via composição das variáveis socioeconômicas e variáveis técnicas).

Tabela 10 – Diagnóstico da concessionária de energia com respeito às inspeções realizadas ao longo de quatro anos.

<i>Tipos de Irregularidades Encontradas</i>	<i>Número de Ocorrências</i>
Desvios fora da medição ou <i>by-pass</i>	57
Autorreligações	237
Medidores Avariados	324
Fraudes no Medidor	472
Outras Irregularidades	150
UCs Normais ou Regulares	10.126

Fonte: Elaboração do próprio autor.

Tabela 11 – Descrições das variáveis analisadas por subáreas.

<i>Variáveis Analisadas</i>		<i>Siglas</i>	<i>Descrições</i>
<i>Variáveis Socioeconômicas</i>	<i>Média de Moradores</i>	<i>MOR</i>	Número médio de moradores em domicílio particular
	<i>Renda Média</i>	<i>REN</i>	Rendimento médio mensal das pessoas de 10 anos ou mais (com e sem renda) em reais
	<i>%UCs Alugados⁽¹⁾</i>	<i>ALU</i>	Percentual de domicílios particulares alugados
	<i>%UCs com Água⁽¹⁾</i>	<i>AGU</i>	Percentual de domicílios particulares com abastecimento de água da rede geral
	<i>%UCs com Lixo Coletado⁽¹⁾</i>	<i>LIX</i>	Percentual de domicílios particulares com lixo coletado
	<i>%UCs com Pavimentação⁽¹⁾</i>	<i>PAV</i>	Percentual de domicílios particulares com pavimentação
	<i>%Alfabetizados⁽²⁾</i>	<i>ALF</i>	Percentual de pessoas alfabetizadas com cinco ou mais anos de idade
	<i>%UCs com Medidor Comum</i>	<i>MEC</i>	Domicílio com energia elétrica de companhia distribuidora e com medidor comum a mais de um domicílio
<i>Variáveis Técnicas</i>	<i>Densidade de Carga</i>	<i>DC</i>	Densidade de Carga por setor censitário em kVA/km²
	<i>Inspeções</i>	<i>INS</i>	Número de UCs inspecionadas
	<i>UCs Irregulares</i>	<i>PER</i>	Número de UCs flagradas em irregularidades
	<i>NTRAFO</i>	<i>NTR</i>	Número de transformadores por setor censitário
<i>Variáveis Derivadas</i>	<i>Taxa de Sucesso de Perdas</i>	<i>TS</i>	Razão entre o número de UCs irregulares e o número de UCs inspecionadas por setor censitário
	<i>PERTRAFO</i>	—	Percentual de UCs irregulares por transformador com relação ao total de UCs conectadas no mesmo
	<i>PERAREA</i>	—	Percentual de UCs irregulares por setor com relação ao total de UCs no mesmo
	<i>CAMPANHA</i>	—	Variável binária que indica se houve ações de campanha por setor

Fonte: Elaboração do próprio autor com adaptações de IBGE (2010).

⁽¹⁾Percentuais com relação ao total de domicílios por setor censitário.

⁽²⁾Percentual com relação ao total de indivíduos por setor censitário.

Em último, observa-se que as variáveis do censo demográfico IBGE (2010) não foram atualizadas, pois a data das inspeções disponíveis (de 2009 a 2012) está próxima da data de realização do último censo demográfico – em 2010. Ademais, nesse ínterim, não há mudanças significativas na configuração socioeconômica do município em estudo como se mostra da seção subsequente.

5.3 O PERFIL MUNICIPAL

Não há uma metodologia geral para resolver o problema das PNTs. Cada cidade possui uma realidade socioeconômica e sociocultural própria. Portanto, o conhecimento a respeito do perfil do lugar sob estudo e das subáreas onde ocorrem as perdas auxilia no entendimento das mesmas.

Na Tabela 12, obtida a partir de SEADE (2010), apresenta-se um resumo comparativo entre o município em estudo e o Estado de São Paulo com respeito a alguns indicadores socioeconômicos. A saber: território, população, saúde, habitação, infraestrutura urbana, educação, rendimento e economia. Todos esses indicadores descrevem o perfil do município analisado neste trabalho.

A partir da análise das informações da Tabela 12, pode-se traçar o perfil do município, espaço no qual ocorre o fenômeno sob estudo. Seguem algumas inferências:

- ❖ Em geral, cidades com alta criminalidade registram alta mortalidade de indivíduos jovens (HEPPENSTALL et al., 2012). Conforme Druck et al. (2004), a mortalidade por homicídios na região sudeste da cidade de São Paulo é a causa de mais de 20% de óbitos de homens jovens entre 15 a 45 anos. Depreende-se, portanto, que o município sob análise possui baixa criminalidade, pois a taxa de mortalidade da população entre 15 e 34 anos dele é inferior à taxa do Estado (taxa essa que representa a média entre todos os municípios que o constitui) sendo de 75 e 122, respectivamente, para município e Estado, por 100 mil habitantes.
- ❖ Os serviços públicos de coleta de lixo, abastecimento de água e esgoto sanitário são eficientes, pois quase a totalidade da população possui acesso a eles. Tal eficiência é traduzida pelo índice de desenvolvimento humano municipal (IDHM) superior à média dos municípios paulistas sendo de 0,8 e 0,78, respectivamente, para município e Estado. As melhores condições de vida combinado com a baixa taxa de natalidade (inferior à taxa média do Estado) leva a um alto índice de envelhecimento sendo de 80 e 64 para município e Estado, respectivamente.

- ❖ A população do município sob análise é predominantemente urbana (95%) com alta densidade demográfica sendo superior ao dobro da densidade do Estado. O crescimento da população encontra-se em um nível de saturação e é inferior à média dos demais municípios sendo de 0,70 e 0,87, para município e Estado, respectivamente.

Tabela 12 – Perfil socioeconômico do município sob análise e comparativo com o Estado de São Paulo (SP).

<i>Indicadores</i>	<i>Descrição</i>	<i>Município</i>	<i>Estado - SP</i>
Área	Em km ²	550	248.223
População	Número de habitantes	200.000	42.673.000
Densidade Demográfica	Habitantes/km ²	380	172
Taxa Geométrica de Crescimento Anual da População	Expressa, o crescimento médio da população em determinado período de tempo – de 2010 a 2015 em percentual ao ano (% a.a.).	0,70	0,87
Grau de Urbanização	Percentual	95	96
Índice de Envelhecimento	Proporção de pessoas com 60 anos ou mais para cada grupo de 100 indivíduos de 0 a 14 anos.	80	64
Taxa de Natalidade	Por 1 mil pessoas	10	15
Taxa de Mortalidade da População entre 15 e 34 anos	Por 100 mil habitantes	75	122
Índice de Desenvolvimento Humano Municipal (IDHM)	Indicador que sintetiza três aspectos: longevidade, educação e renda	0,80	0,78
Renda Per Capita	Em reais	900	854
Coleta de Lixo	Percentual de atendimento	99,0	99,6
Abastecimento de água	Percentual de atendimento	99,0	97,9
Esgoto sanitário	Percentual de atendimento	97,0	89,7
Taxa de analfabetismo da população de 15 anos ou mais	Percentual	4,0	4,3
Participação da Agropecuária	Percentual do valor adicionado pelo setor agropecuário no valor total do município	1,0	1,9
Participação da Indústria	Percentual do valor adicionado pela indústria no valor total do município	17,0	24,9
Participação dos Serviços	Percentual do valor adicionado pelo setor de serviços no valor total do município	82,0	73,1

Fonte: Adaptado de SEADE (2010).

- ❖ O setor de serviços é predominante na economia seguido pela indústria e em menor importância a agropecuária, cujas participações na economia são de 82%, 17% e 1%, respectivamente.

A eficiência no combate às PNTs é avaliada pelas concessionárias de energia por meio do montante de energia recuperada e não pelo total de UCs flagradas em irregularidades. Como o setor industrial possui participação de apenas 17%, no município em questão, há um número reduzido de indústrias. Como em geral o consumo das mesmas é elevado, todas as indústrias são inspecionadas.

De acordo com dados da Tabela 12, a economia local é impulsionada majoritariamente pelo setor de serviços (82%); logo, o município analisado possui muitos comércios tais como: açougues, padarias, distribuidoras de bebidas, supermercados, lojas em geral, etc. Em geral, o consumo de energia em kWh de pequenos estabelecimentos comerciais, microempresas é sobremaneira elevado se comparado ao consumo de residências comuns. Portanto, a classe consumidora crítica no município analisado com respeito às PNTs é justamente a classe comercial.

5.4 ANÁLISE EXPLORATÓRIA

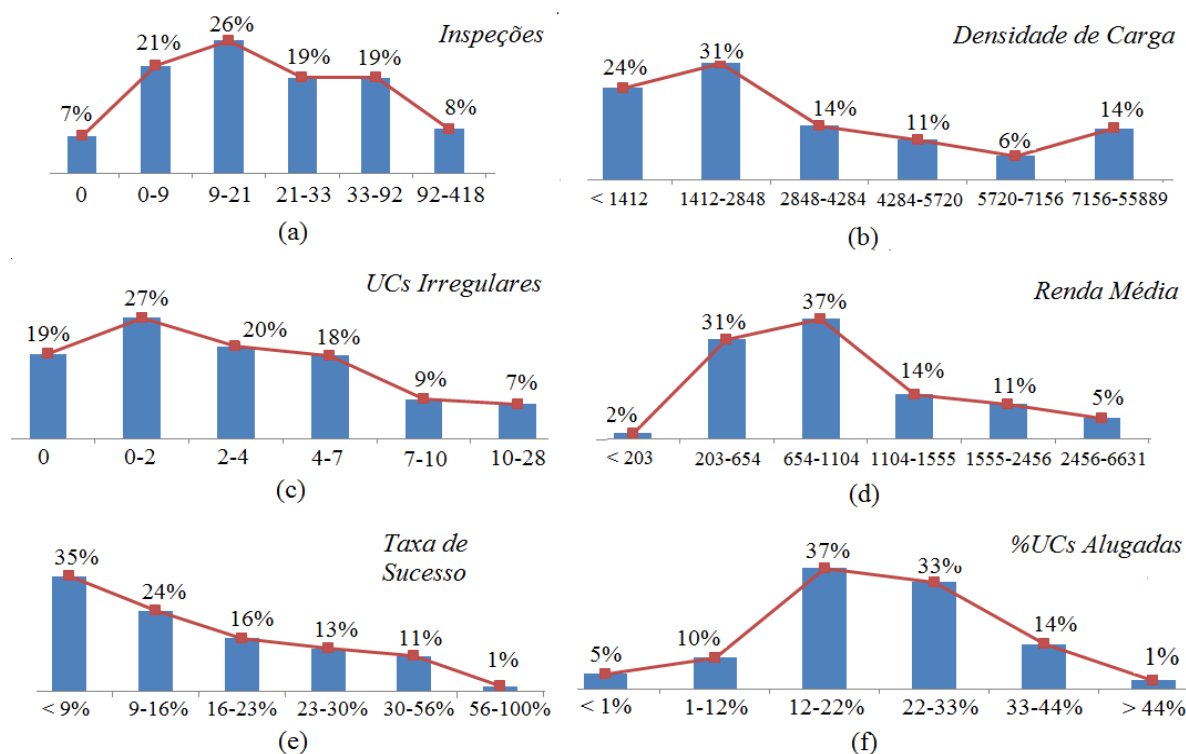
O processo de modelagem espacial é precedido por uma fase de análise exploratória na qual se incluem a apresentação visual dos dados em forma de gráficos e mapas e a identificação de padrões de dependência espacial do fenômeno em estudo. Observa-se, dessa forma, a distribuição das variáveis sob estudo, identificam-se padrões na distribuição espacial e padrões atípicos ou *outliers*. Através desses procedimentos de exploração inicial dos dados, elaboram-se hipóteses a partir das observações realizadas e seleciona-se o melhor modelo inferencial que melhor se adequa aos dados. Averigua-se se há indícios de que o lugar onde ocorre o fenômeno exerce influência sobre o mesmo.

Este é o objetivo dessa seção, isto é, apresentar os dados na forma de mapas e verificar se é viável abordar o problema das PNTs com a análise espacial de dados geográficos. Caso seja viável, a informação geográfica disponível também será considerada informação espacial.

Segundo Ferreira (2014, p. 74), a interpretação da forma dos histogramas em linhas contínuas é uma etapa indispensável e importante para o entendimento de como se comporta a distribuição de uma variável em um conjunto de unidades geográficas. Neste trabalho, as unidades geográficas são os setores censitários. Nesse contexto, nas Figuras 11 e 12 apresentam-se os histogramas de frequência relativa para as variáveis socioeconômicas, técnicas e variáveis derivadas da Tabela 11. Nas Figuras 11 e 12, o eixo das abscissas contém as classes discretas obtidas via técnica de *intervalos de desvio padrão* (vide seção 3.4.4.2). Mostra-se

nesses histogramas o percentual de dados que há em cada classe (acima da barra vertical em azul). Na Figura 11 (a), por exemplo, tem-se o histograma do número de inspeções realizadas por subáreas ao longo de quatro anos. Observa-se que a maioria das subáreas do município em estudo (26% delas) receberam entre 9 e 21 inspeções.

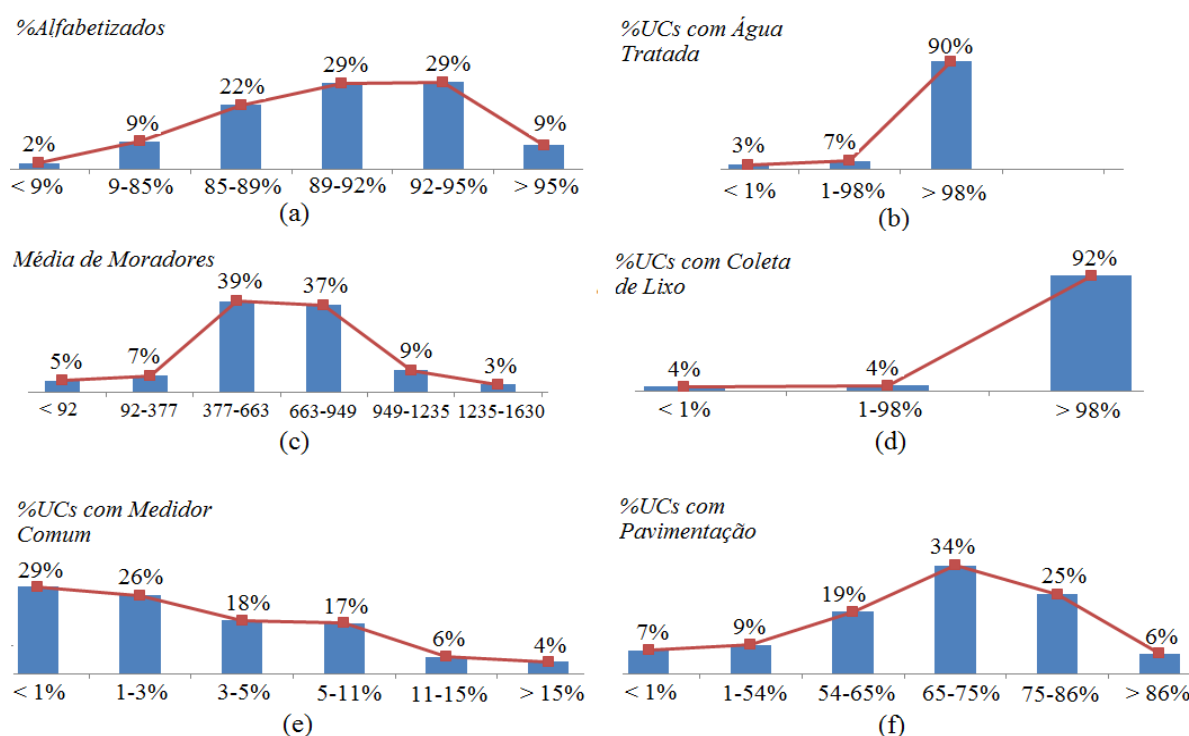
Figura 11 – Histogramas das frequências relativas das variáveis socioeconômicas, técnicas e derivadas, por subáreas (a) *Inspeções*; (b) *Densidade de Carga* (kVA/km²); (c) *UCs Irregulares*; (d) *Renda Média* (em R\$); (e) *Taxa de Sucesso* percentual e (f) *%UCs Alugadas*.



Fonte: Elaboração do próprio autor.

Nas Figuras 13 a 16 apresentam-se mapas temáticos com a distribuição espacial das classes das variáveis socioeconômicas, variáveis técnicas e derivadas da Tabela 11. Essas classes foram apresentadas nos histogramas das Figuras 11 e 12 (no eixo das abscissas) e estão dispostas no espaço geográfico do município em estudo. Associa-se uma cor a cada classe. A distribuição em frequência relativa percentual de cada intervalo encontra-se entre parênteses. Na Figura 13 (a), por exemplo, tem-se a distribuição espacial das inspeções realizadas no município em um mapa de cores ou mapa temático. Essa figura contém as mesmas seis classes para as inspeções que foram mostradas em seu histograma de frequência relativa da Figura 11 (a). Observa-se da Figura 13 (a) que a maioria das subáreas possuem a cor correspondente à classe majoritária (verde cintilante) segundo a qual 26% das subáreas do município receberam entre 9 e 21 inspeções.

Figura 12 – Histogramas de frequências relativas das variáveis socioeconômicas, técnicas e derivadas por subáreas. (a) %Alfabetizados; (b) %UCs com Água Tratada; (c) Média de Moradores; (d) %UCs com Coleta de Lixo; (e) %UCs com Medidor Comum e (f) %UCs com pavimentação.



Fonte: Elaboração do próprio autor.

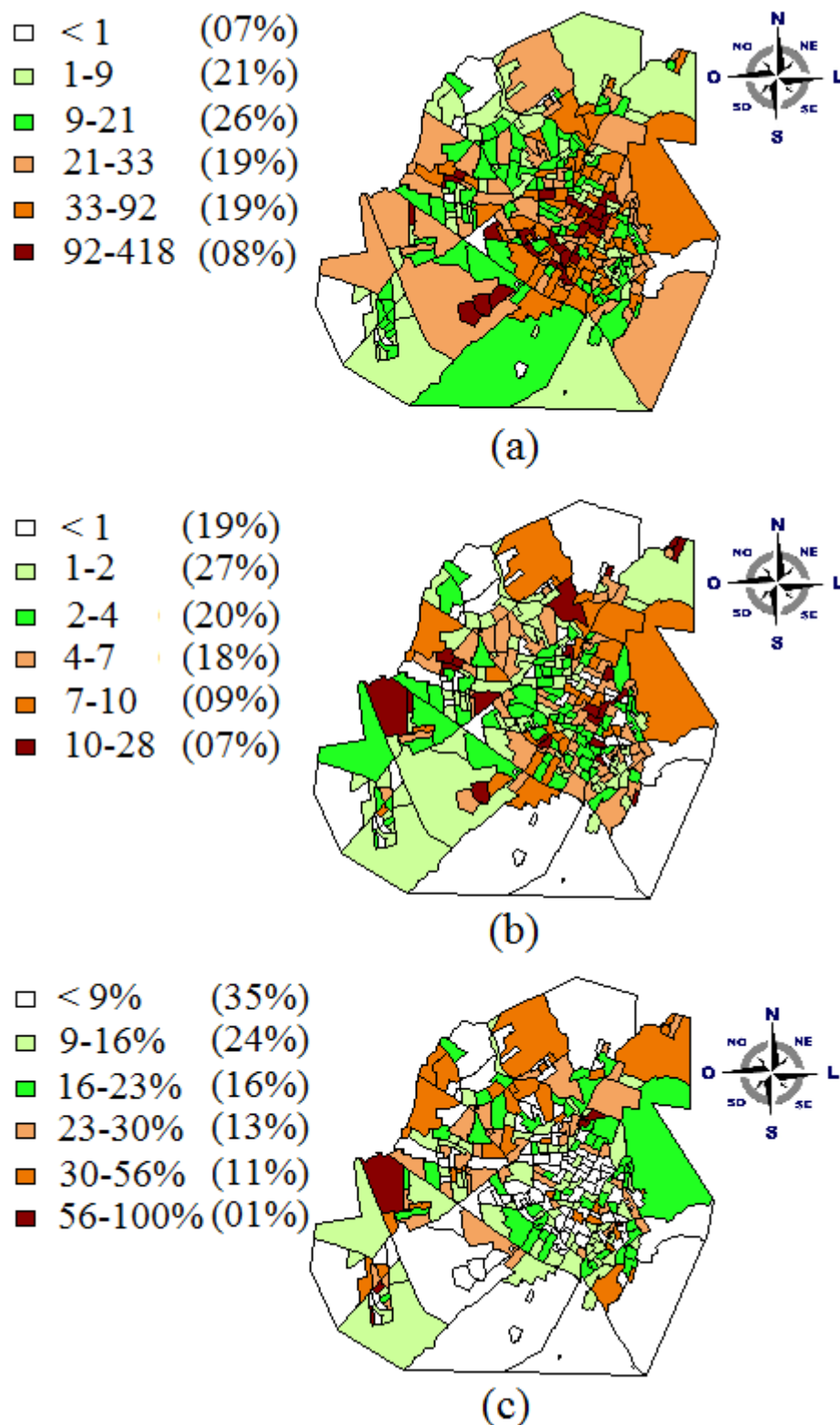
A percepção visual e intuitiva das variáveis obtida através dos histogramas de frequência relativa (Figuras 11 e 12) e dos mapas temáticos com a distribuição espacial das variáveis (Figuras 13 a 16) pode ser avaliada mais objetivamente de modo quantitativo com o auxílio das métricas da estatística descritiva.

Segundo Bussab e Morettin (2010) e Larson e Farber (2010), a estatística descritiva é útil e comumente empregada na fase de análise exploratória dos dados. Constitui-se no ramo da estatística que descreve e avalia certo grupo de dados, seja ele população ou amostra. Existem na estatística descritiva basicamente dois métodos para apresentação dos dados: métodos gráficos e métodos numéricos – envolvendo apresentação de medidas de posição, dispersão, etc. A estatística descritiva preocupa-se em inferir sobre um todo a partir de medidas de uma ou mais características ou atributos (variáveis) que descrevem o todo. Tal suposição é feita a partir de um conjunto de entidades unitárias (setores censitários) formadores do todo (município) que se denominam amostras (FERREIRA, 2014, p. 66).

Nesse contexto, apresentam-se na Tabela 13 algumas métricas clássicas da estatística descritiva relativas à tendência central e à dispersão das variáveis utilizadas neste estudo. Tais métricas estão descritas em Ferreira (2014, p. 77-83) e em outros textos de estatística tais

como: Bussab e Morettin (2010) e Larson e Farber (2010). Seguem algumas deduções a partir da análise dos dados desta Tabela 13 e das Figuras 11-16:

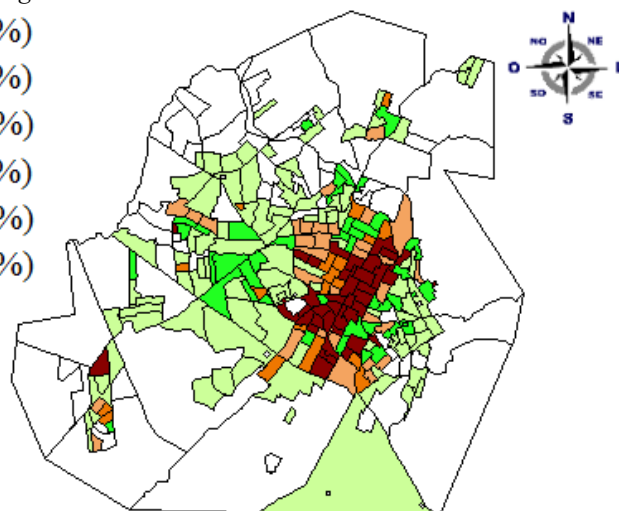
Figura 13 – Mapas temáticos com a distribuição espacial das variáveis por subáreas de 2009 a 2012. (a) *Inspeções realizadas*; (b) *UCs Irregulares* e (c) *Taxa de Sucesso* percentual das inspeções.



Fonte: Elaboração do próprio autor.

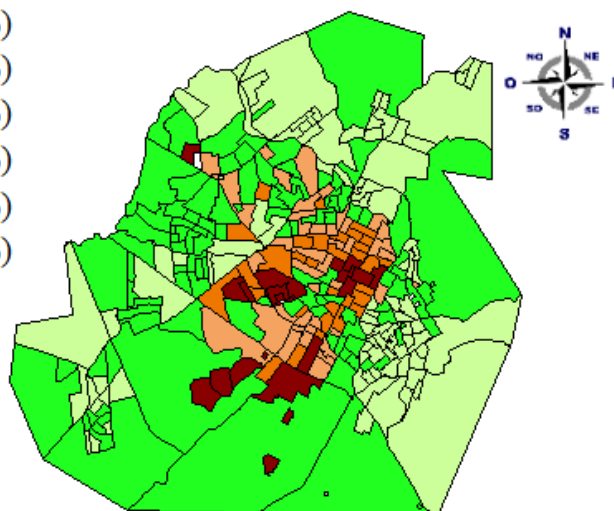
Figura 14 – Mapas temáticos com a distribuição espacial por subárea da (a) *Densidade de Carga* em kVA/km²; (b) *Renda Média* em reais e (c) *%UCs Alugadas*.

- < 1412 (24%)
- 1412-2848 (31%)
- 2848-4284 (14%)
- 4284-5720 (11%)
- 5720-7156 (06%)
- 7156-55889 (14%)



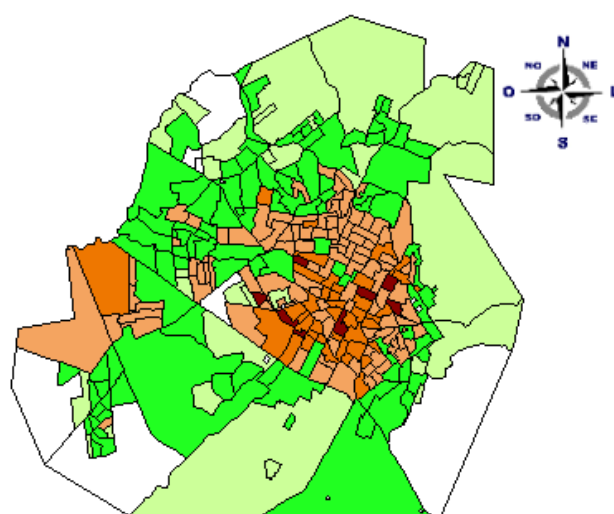
(a)

- < 203 (02%)
- 203-654 (31%)
- 654-1104 (37%)
- 1104-1555 (14%)
- 1555-2456 (11%)
- 2456-6631 (05%)



(b)

- < 1% (05%)
- 1-12% (10%)
- 12-22% (37%)
- 22-33% (33%)
- 33-44% (14%)
- > 44% (01%)

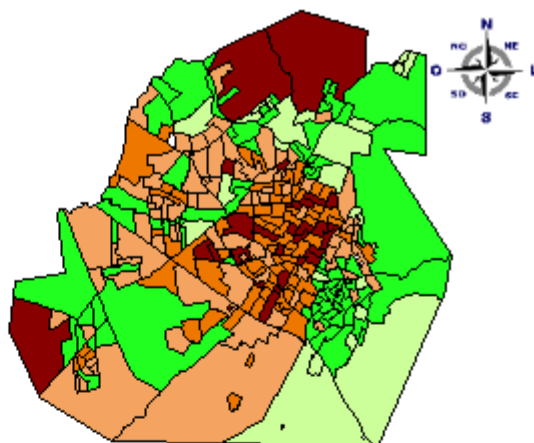


(c)

Fonte: Elaboração do próprio autor.

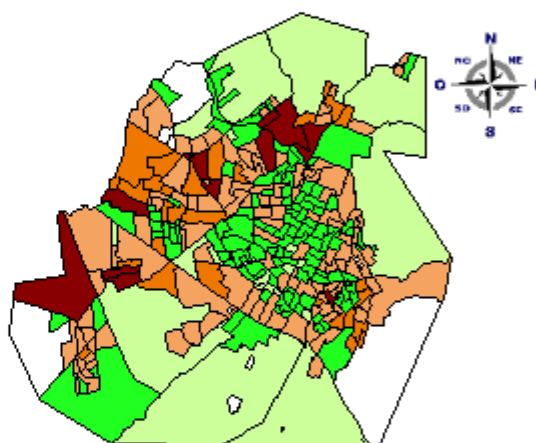
Figura 15 – Mapas temáticos com a distribuição espacial por subáreas. (a) %Alfabetizadas; (b) Média de Moradores e (c) %UCs com Medidor Comum.

- | | | |
|---|--------|-------|
| □ | < 9% | (02%) |
| □ | 9-85% | (09%) |
| ■ | 85-89% | (22%) |
| ■ | 89-92% | (29%) |
| ■ | 92-95% | (29%) |
| ■ | > 95% | (09%) |



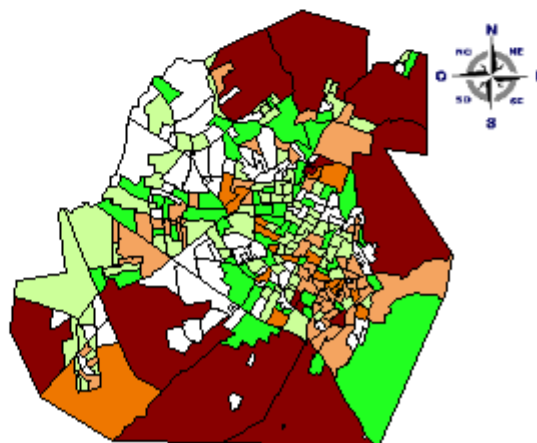
(a)

- | | | |
|---|-----------|-------|
| □ | < 92 | (05%) |
| □ | 92-377 | (07%) |
| ■ | 377-663 | (39%) |
| ■ | 663-949 | (37%) |
| ■ | 949-1235 | (09%) |
| ■ | 1235-1630 | (03%) |



(b)

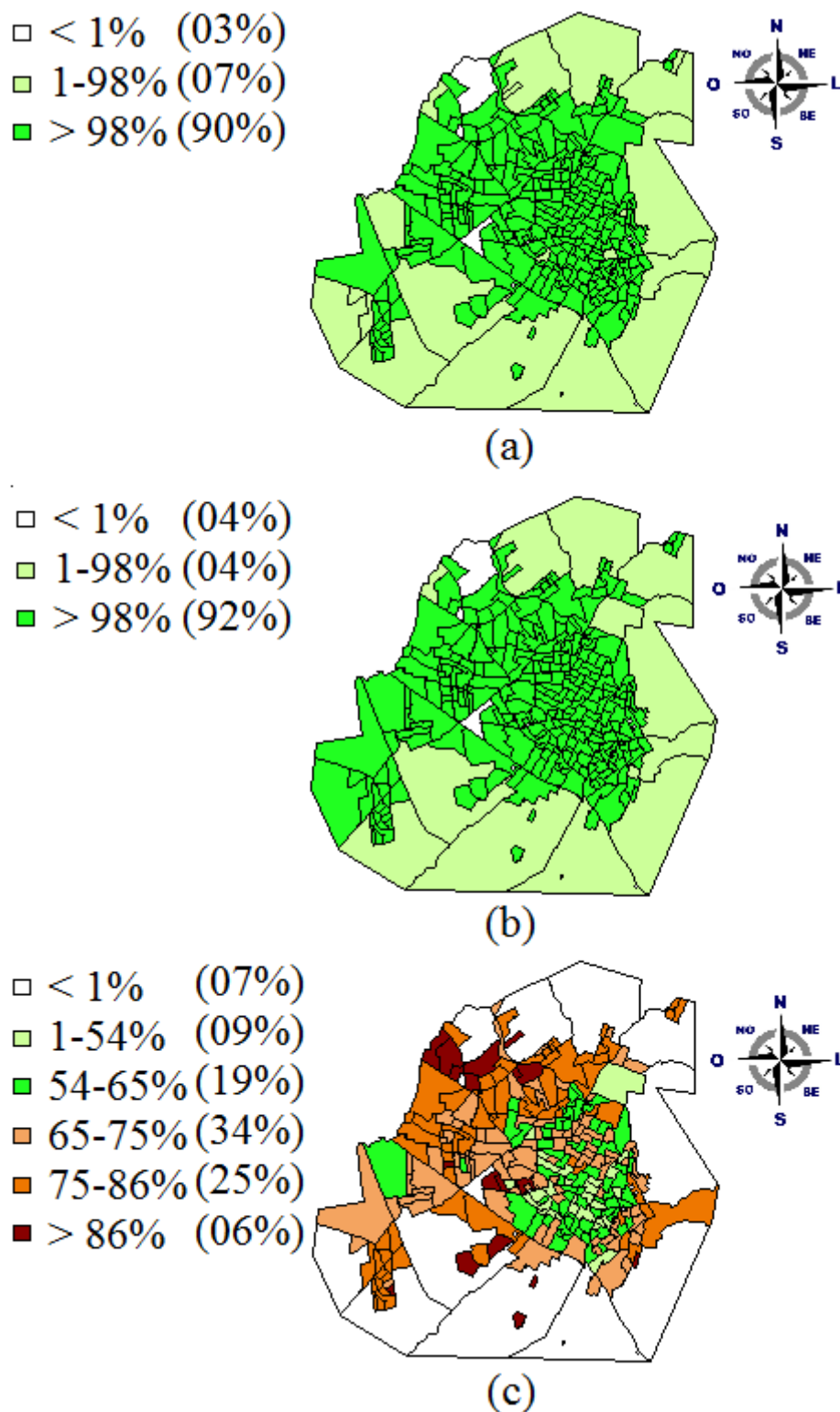
- | | | |
|---|--------|-------|
| □ | < 1% | (29%) |
| □ | 1-3% | (26%) |
| ■ | 3-5% | (18%) |
| ■ | 5-11% | (17%) |
| ■ | 11-15% | (06%) |
| ■ | > 15% | (04%) |



(c)

Fonte: Elaboração do próprio autor.

Figura 16 – Mapas temáticos com a distribuição espacial por subáreas. (a) %UCs com Água Tratada; (b) %UCs com Coleta de Lixo e (c) %UCs com Pavimentação Asfáltica.



Fonte: Elaboração do próprio autor.

Tabela 13 – Valores das medidas de tendência central e de dispersão para as variáveis da Tabela 11 por subáreas: Média ($\bar{X}$); Moda (Mo); Mediana (Md); Desvio Padrão (σ); Coeficiente de Variação (CV); Coeficiente de Assimetria (S) e Curtose (kt).

<i>Variáveis</i>	<i>Medidas de Tendência Central</i>			<i>Medidas de Dispersão</i>			
	$(\bar{X})$	(Mo)	(Md)	(σ)	(CV)	(S)	(kt)
<i>Taxa de Sucesso</i>	16%	14%	13%	15%	94,4	0,53	4,72
<i>UCs Irregulares</i>	3,91	0,43	3,00	3,97	101,7	0,68	6,48
<i>Inspeções</i>	33,06	3,31	19,00	47,59	143,9	0,89	18,60
<i>Densidade de Carga (kVA/km²)</i>	4283,7	2130,0	2495,9	5743,1	134,1	0,93	27,13
<i>NTRAFO</i>	9,04	5,23	6,00	10,17	112,5	0,90	18,92
<i>%UCs Alugadas</i>	22%	18%	22%	11%	48,2	0,04	-0,21
<i>%UCs Medidor Comum</i>	5%	1%	3%	8%	164,2	0,83	21,07
<i>Renda Média (R\$)</i>	1104,3	541,1	816,9	901,2	81,6	0,96	10,28
<i>%UCs com Pavimentação</i>	65%	67%	68%	21%	32,8	-0,55	3,59
<i>%UCs com Água tratada</i>	94%	100%	100%	23%	24,8	-0,82	11,29
<i>%UCs com esgoto</i>	93%	100%	100%	23%	25,2	-0,87	10,92
<i>%UCs com Coleta de lixo</i>	95%	100%	100%	21%	21,9	-0,76	14,78
<i>%Alfabetizados</i>	89%	90%	91%	15%	17,0	-0,38	29,23
<i>Média de Moradores</i>	663,11	770,3	660,0	285,9	43,1	0,03	0,79

Fonte: Elaboração do próprio autor.

- ❖ Todas as distribuições em frequência relativa são unimodais, ou seja, há predomínio de apenas uma classe de frequência relativa nos valores das variáveis conforme se pode constatar nos histogramas de frequências relativas das Figuras 11 e 12.
- ❖ Da Tabela 13, observa-se que a média e a moda são aproximadamente coincidentes e situam-se na mediana do gráfico. Os dados das subáreas estão posicionados próximos à média em sua maioria. A distorção regional entre setores é pequena. Encontram-se nessa situação as variáveis: *%UCs Alugadas* e *Média de Moradores*.
- ❖ Assimetria levemente positiva (média situada um pouco abaixo da mediana $\bar{X} < Md$ e coeficiente de assimetria positivo $S > 0$). Nesse caso, o histograma de frequência relativa está deslocado ligeiramente para a esquerda. As variáveis possuem leve tendência para assumir valores inferiores à média. São elas: *%UCs com Pavimentação*; *%UCs com água tratada*, *%UCs com Esgoto*; *%UCs com coleta de lixo* e *%Alfabetizados*. Como a maior parte da população do município é assis-

tida por serviços públicos básicos, é esperado que a média das variáveis socioeconômicas torne-se um pouco inferior à mediana.

- ❖ Assimetria levemente positiva (média situada um pouco acima da mediana – $\bar{X} > Md$ e coeficiente de assimetria negativo $S < 0$). Neste caso, o histograma de frequência relativa encontra-se deslocado ligeiramente para a direita. Variáveis com leve tendência para assumir valores superiores à média. São elas: *Taxa de Sucesso*; *NTRAFO*; *UCs Irregulares*; *Inspecções*; *Densidade de Carga* e *%UCs Medidor em Comum*.

Além de avaliar as variáveis visualmente e numericamente via mapas temáticos e via métricas da estatística descritiva, respectivamente, é importante avaliar também a correlação entre as variáveis em estudo. Nesse contexto, na Tabela 14 apresenta-se a matriz simétrica dos coeficientes de correlação de Spearman r_s .

Tabela 14 – Matriz simétrica de coeficientes globais de correlação de Spearman (r_s) entre pares de variáveis.

	<i>TS</i>	<i>PER</i>	<i>INSP</i>	<i>DC</i>	<i>NTR</i>	<i>ALU</i>	<i>MEC</i>	<i>REN</i>	<i>PAV</i>	<i>AGU</i>	<i>ESG</i>	<i>LIX</i>	<i>ALF</i>	<i>MOR</i>
<i>TS</i>	1,00	0,67	0,13	0,07	0,23	0,09	0,30	-0,13	0,24	0,65	0,52	0,77	-0,13	0,37
<i>PER</i>	0,67	1,00	0,84	0,51	0,61	0,43	0,38	0,27	0,17	0,74	0,62	0,83	0,25	0,43
<i>INSP</i>	0,13	0,84	1,00	0,54	0,60	0,42	0,21	0,37	-0,08	0,75	0,59	0,84	0,29	0,20
<i>DC</i>	0,07	0,51	0,54	1,00	0,37	0,58	0,06	0,40	-0,20	0,82	0,70	0,89	0,41	-0,12
<i>NTR</i>	0,23	0,61	0,60	0,37	1,00	0,20	0,28	0,40	0,04	0,56	0,45	0,67	0,32	0,23
<i>ALU</i>	0,09	0,43	0,42	0,58	0,20	1,00	0,25	0,39	-0,53	0,80	0,64	0,91	0,37	-0,07
<i>MEC</i>	0,30	0,38	0,21	0,06	0,28	0,25	1,00	-0,18	-0,21	0,60	0,35	0,71	-0,12	0,15
<i>REN</i>	-0,13	0,27	0,37	0,40	0,40	0,39	-0,18	1,00	-0,03	0,74	0,65	0,87	0,77	-0,27
<i>PAV</i>	0,24	0,17	-0,08	-0,20	0,04	-0,53	-0,21	-0,03	1,00	0,73	0,60	0,79	-0,09	0,39
<i>AGU</i>	0,65	0,74	0,75	0,82	0,56	0,80	0,60	0,74	0,73	1,00	0,96	0,99	0,71	0,68
<i>ESG</i>	0,52	0,62	0,59	0,70	0,45	0,64	0,35	0,65	0,60	0,96	1,00	0,98	0,65	0,52
<i>LIX</i>	0,77	0,83	0,84	0,89	0,67	0,91	0,71	0,87	0,79	0,99	0,98	1,00	0,83	0,81
<i>ALF</i>	-0,13	0,25	0,29	0,41	0,32	0,37	-0,12	0,77	-0,09	0,71	0,65	0,83	1,00	-0,26
<i>MOR</i>	0,37	0,43	0,20	-0,12	0,23	-0,07	0,15	-0,27	0,39	0,68	0,52	0,81	-0,26	1,00

Fonte: Elaboração do próprio autor.

Conforme discutido na seção 3.4.4.1, o coeficiente de correlação entre duas variáveis varia entre -1 e 1 , isto é, $r_s \in [-1, 1]$. Um coeficiente de correlação positivo implica em uma relação diretamente proporcional entre as variáveis; um coeficiente negativo implica em uma relação inversamente proporcional entre elas. Observa-se que as variáveis da Tabela 14 estão associadas às siglas que foram definidas na Tabela 11.

Seguem algumas inferências a partir da análise dos dados da Tabela 14:

- ❖ As variáveis *UCs Irregulares (PER)* e *Densidade de Carga (DC)* possuem correlação positiva moderada de 0,51. Logo, há um indício de que se encontram mais perdas em subáreas com elevada densidade de carga (região central do município, por exemplo). As variáveis *Taxa de Sucesso (TS)* e *%UCs alugadas (ALU)* têm fraca correlação positiva de 0,09.
- ❖ As variáveis *UCs Irregulares (PER)* possuem correlação moderada positiva com *%UCs alugadas (ALU)* e *%Medidor Comum (MEC)* de 0,43 e 0,38, respectivamente. Logo, há um indício de que a existência de muitas UCs alugadas favoreça a presença das perdas.
- ❖ Esperava-se uma correlação negativa entre as variáveis socioeconômicas *%UC com esgoto (ESG)*, *%UCs com Água (AGU)*, *%UCs com Lixo Coletado (LIX)*, *%UCs com Pavimentação (PAV)* e as *UCs Irregulares (PER)*, ou seja, quanto menos serviços públicos, mais perdas. Isso não ocorre provavelmente devido ao acesso da quase totalidade da população aos serviços públicos básicos.
- ❖ Observa-se que a correlação entre *UCs Irregulares (PER)* e *Renda Média (REN)* é fracamente positiva – 0,27. A distribuição homogênea da renda é um dos fatores que contribui para a fraca correlação entre *PER* e *REN*.
- ❖ Conforme esperado, há uma forte correlação positiva entre as *Inspeções (INSP)* e as fraudes – *UCs Irregulares (PER)* – de 0,84, ou seja, quanto mais inspeções realizadas em cada subáreas, mais fraudes são encontradas.
- ❖ As perdas estão concentradas em setores centrais que têm extensa rede de distribuição – correlação moderada positiva de 0,61 entre *UCs Irregulares (PER)* e *NTRAFOS (NTR)* – com significativo número de UCs e regiões populosas – correlação moderada positiva de 0,43 entre *UCs Irregulares (PER)* e *Média Moradores (MOR)*.

Segundo Gollini et al. (2015, p. 38), a fraca correlação das variáveis em escala global não implica em fraca correlação a nível local. Ademais, podem existir variáveis que estejam fracamente correlacionadas às perdas, mas que são importantes para melhor compreensão do problema do ponto de vista qualitativo e vice-versa.

Conforme Rodrigues (2012, p. 29), o índice global enfatiza similaridades, pressupondo que todas as subáreas podem ser representados por um único valor. No entanto, a presença de distorções regionais invalida tal suposição. Em certos casos, resultados opostos podem ser obtidos quando os dados são analisados globalmente e localmente. Na seção 5.7,

complementa-se a análise global do coeficiente de correlação via análise espacial a nível local.

Em Miotto (2015), apresenta-se uma importante observação sobre a “falácia estatística”. Segundo essa referência, a correlação não significa causalidade. Em muitos casos, há modelos que não consideram variáveis fundamentais, por não estarem disponíveis, por exemplo. Mesmo assim, tais modelos podem apresentar forte correlação espacial. É preciso cuidado na interpretação de uma aparente correlação espacial que pode ser consequência de um modelo mal especificado (BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008, p.259; HUFF, 1973).

Em último, não se pode afirmar conclusivamente que as variáveis da Tabela 11 são as melhores possíveis para análise do problema em questão. Tal questão foge ao escopo deste trabalho. No entanto, tais variáveis são adequadas em termos qualitativos (estão relacionadas ao problema das PNTs) e em termos quantitativos como indicam as análises exploratórias realizadas nesta seção.

5.4.1 Mapeamento Anual das Inspeções, Irregularidades e Taxa de Sucesso de Perdas

Uma das formas mais usuais de definição da probabilidade é de acordo com (44). Nesta, a probabilidade $P(e)$ de um evento e é a frequência relativa com a qual este evento ocorre em uma série de tentativas sob as mesmas condições (BUSSAB; MORETTIN, 2010):

$$\hat{P}(e) = \frac{N_e}{n}, \quad (44)$$

sendo N_e o número de vezes em que o evento e ocorreu em n tentativas. $\hat{P}(e)$ é uma estimativa da probabilidade $P(e)$, cujo valor teórico é o limite em (45).

$$P(e) = \lim_{n \rightarrow \infty} \frac{N_e}{n} \quad (45)$$

Supõe-se que, enquanto a ocorrência de um evento e em uma observação particular é totalmente incerta, a frequência relativa com a qual este evento ocorre em observações repetidas possui propriedades estáveis. Se não houver evidência de regularidade, a teoria das probabilidades pode não ser aplicável (CAMARGO, 1981).

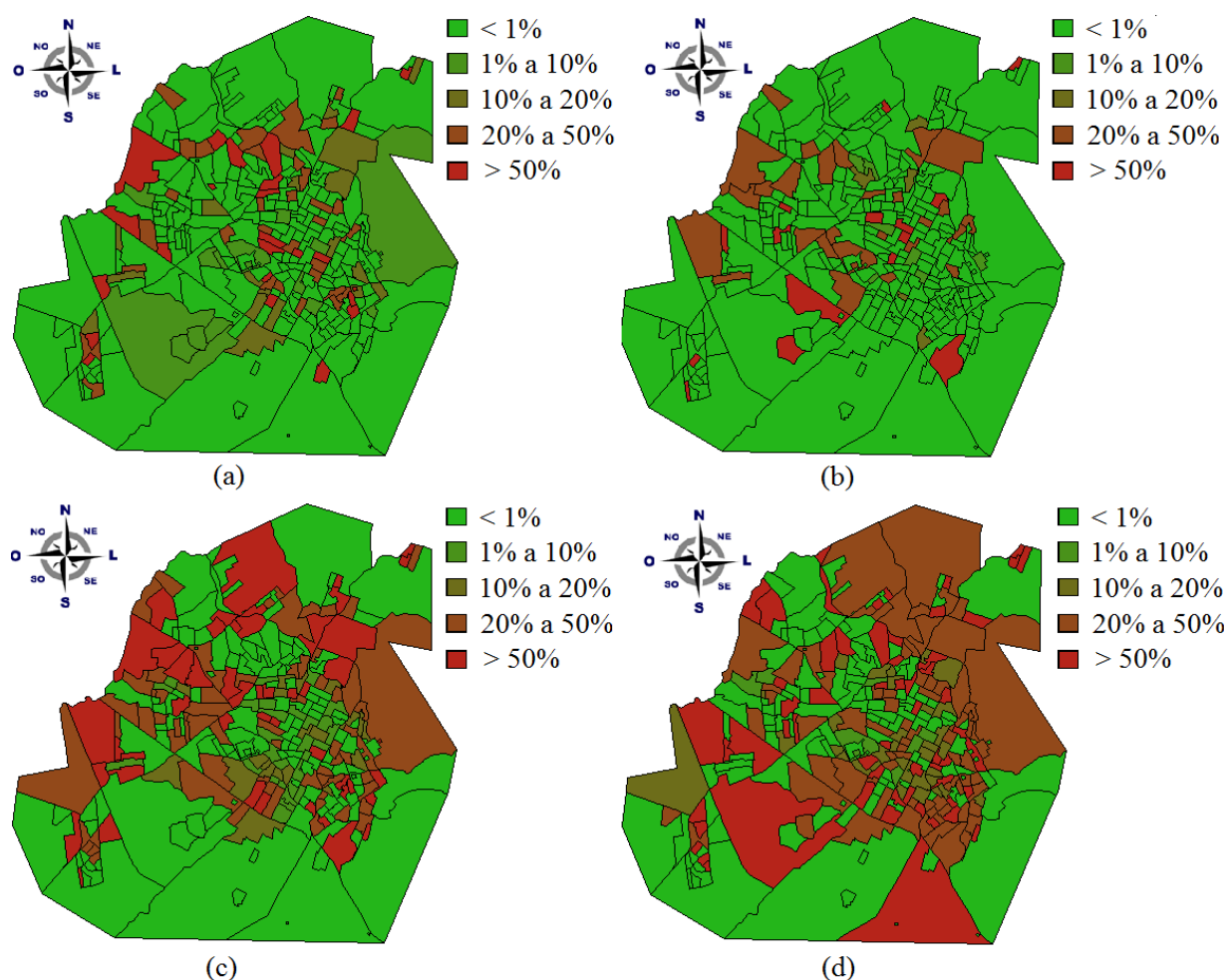
Neste trabalho, N_e representa o número de vezes em que o evento *UCs irregulares* foi encontrado em um total de n *UCs Inspeccionadas* pelas equipes de inspeções em campo da concessionária de energia. Nesses termos, reescreve-se a expressão (44) e obtém-se (46):

$$TS_A(Perdas) = 100 \times \frac{N^{\circ} \text{ UCs Irregulares Encontradas}}{N^{\circ} \text{ UCs Inspeccionadas}}, \quad (46)$$

sendo TS_A a *Taxa de Sucesso* percentual das inspeções na subárea ou setor censitário A e; neste estudo, a *Taxa de Sucesso* representa uma estimativa para a probabilidade anual das perdas na subárea A .

Na Figura 17 emprega-se (46) para cálculo da *Taxa de Sucesso* anual das inspeções por subáreas de 2009 a 2012. Observa-se que, à medida que os anos avançam, os mapas tornam-se mais coloridos. Isso significa que há um acréscimo da taxa de sucesso das inspeções, fruto de uma melhoria dos critérios internos da concessionária de energia para seleção das UCs a serem inspeccionadas.

Figura 17 – Mapas temáticos referentes à *Taxa de Sucesso* anualmente por subárea. (a), (b), (c), (d) Mapas referentes aos anos de 2009, 2010, 2011 e 2012, respectivamente.

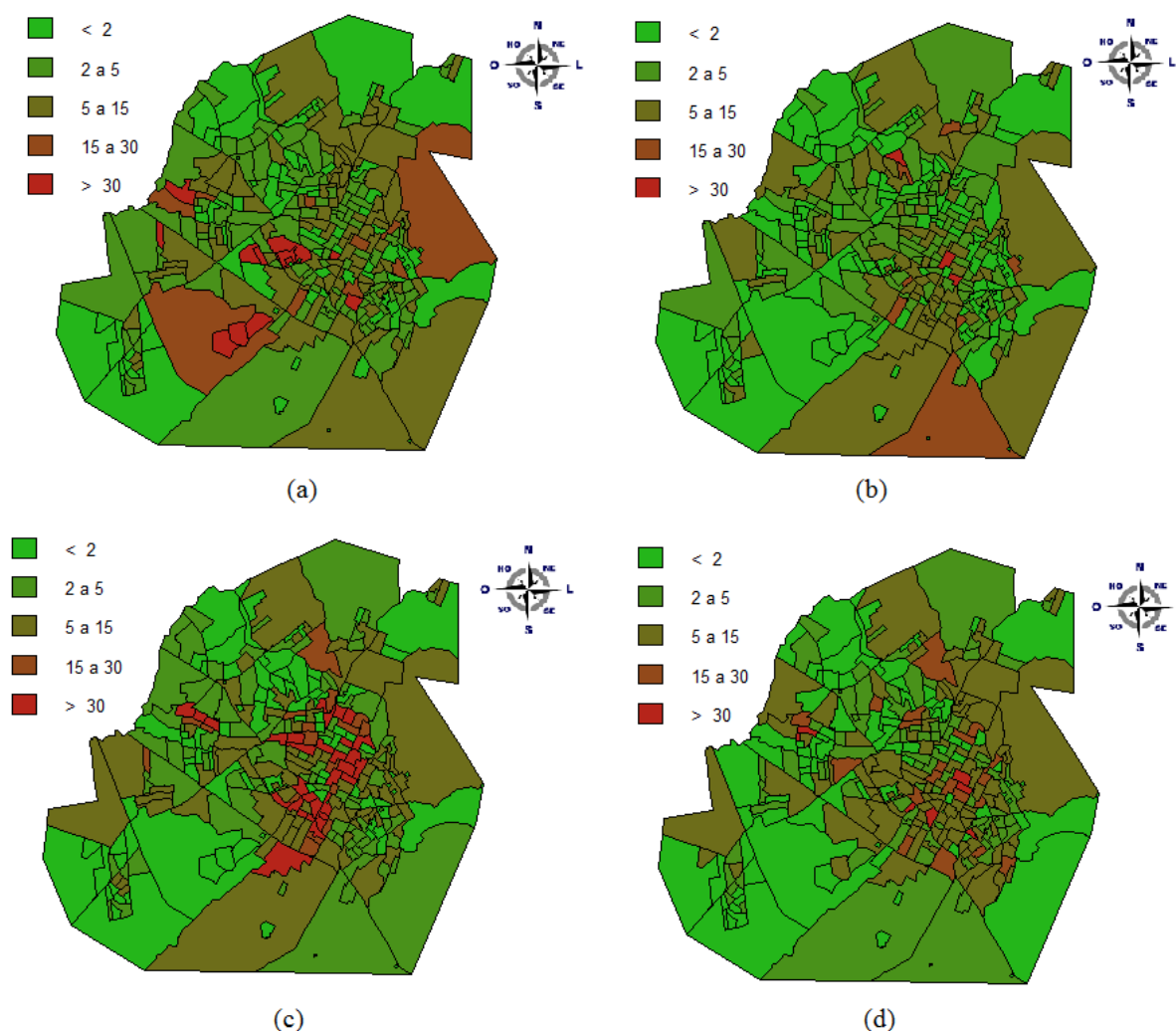


Fonte: Elaboração do próprio autor.

Nas Figuras 17 a 19, os dados de inspeções da Tabela 9 estão distribuídos espacialmente por subáreas do município. Objetiva-se visualizar o padrão espacial anual da taxa de sucesso, das inspeções realizadas e das irregularidades, respectivamente, por subáreas.

Na Figura 18, apresentam-se as inspeções anuais por subáreas. As áreas em tom mais escuro foram alvos de campanhas de inspeções. As campanhas de inspeções caracterizam-se por um número elevado de inspeções em subáreas específicas. Neste trabalho, considera-se que subáreas com mais de trinta UCs inspecionadas foram alvo de campanhas de inspeções.

Figura 18 – Mapas temáticos referentes ao número total de *Inspeções* anualmente por subáreas. (a), (b), (c), (d) Mapas de UCs inspecionadas em 2009, 2010, 2011 e 2012, respectivamente.

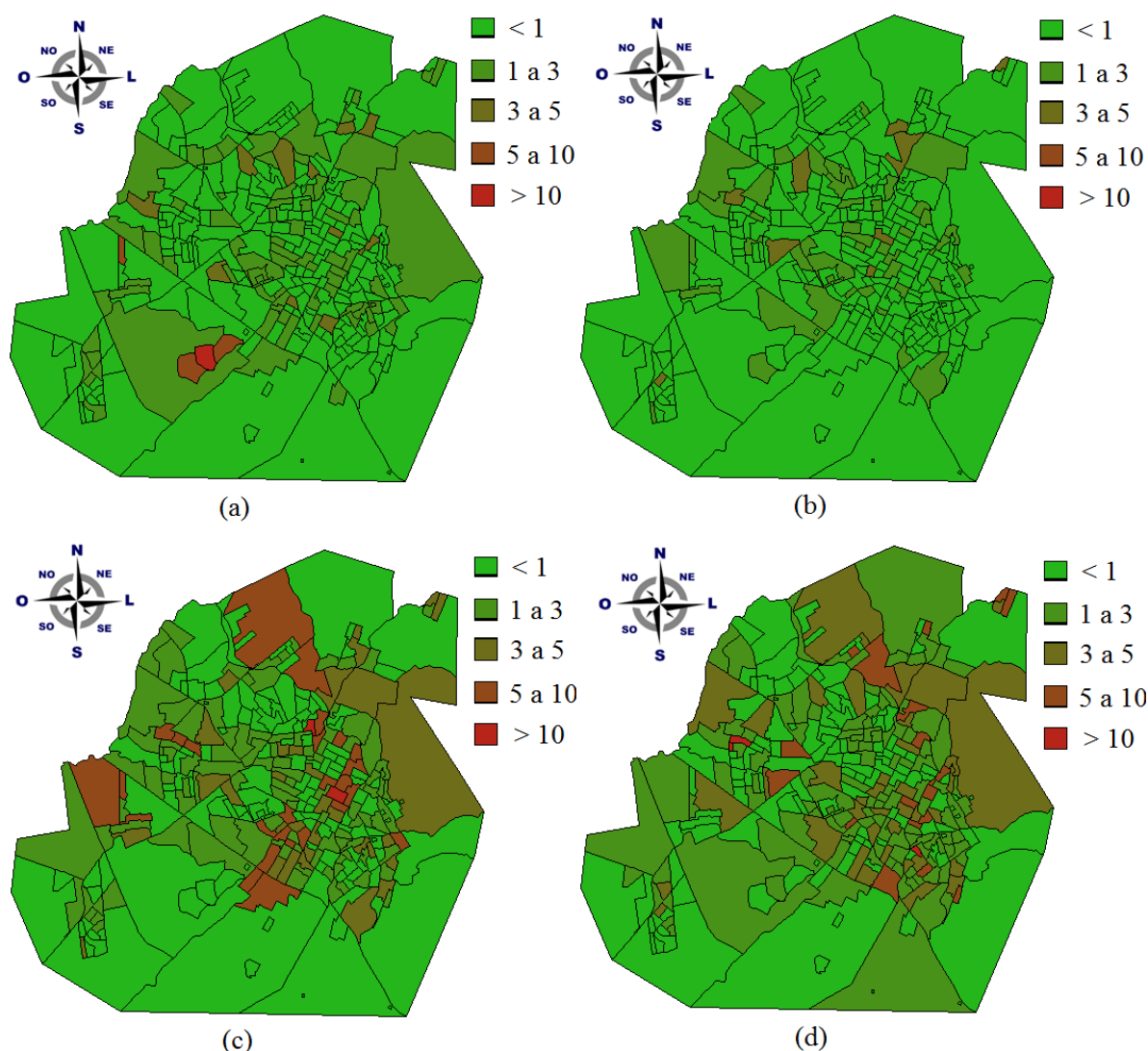


Fonte: Elaboração do próprio autor.

Na Figura 19 apresentam-se as UCs irregulares ou com PNTs, anualmente, por subáreas. Em geral, as subáreas em tom mais escuro na Figura 19, nas quais foram flagradas mais UCs irregulares, são aquelas que foram alvos de campanha de inspeções – subáreas em

tom mais escuro na Figura 18. Isso, pois, a ocorrência de fraudes está condicionada à distribuição das inspeções; quanto maior o número de inspeções em uma dada subárea, maior a probabilidade de se encontrarem UCs irregulares nessa localidade. Tal fato foi confirmado quantitativamente pela alta correlação positiva entre as variáveis *Inspeções (INSP)* e *UCs Irregulares (PER)* de 0,84 na Tabela 14.

Figura 19 – Mapas temáticos referentes ao número total de *UCs Irregulares* anualmente por subáreas. (a), (b), (c), (d) Mapas de *UCs irregulares* em 2009, 2010, 2011 e 2012, respectivamente.



Fonte: Elaboração do próprio autor.

5.4.2 Propriedades da Informação Espacial

A *informação geográfica* caracteriza-se por associar um sistema de referência baseado em coordenadas geográficas ou coordenada de projeção cartográfica a algum atributo locacional. Trata-se, portanto, de uma informação georreferenciada.

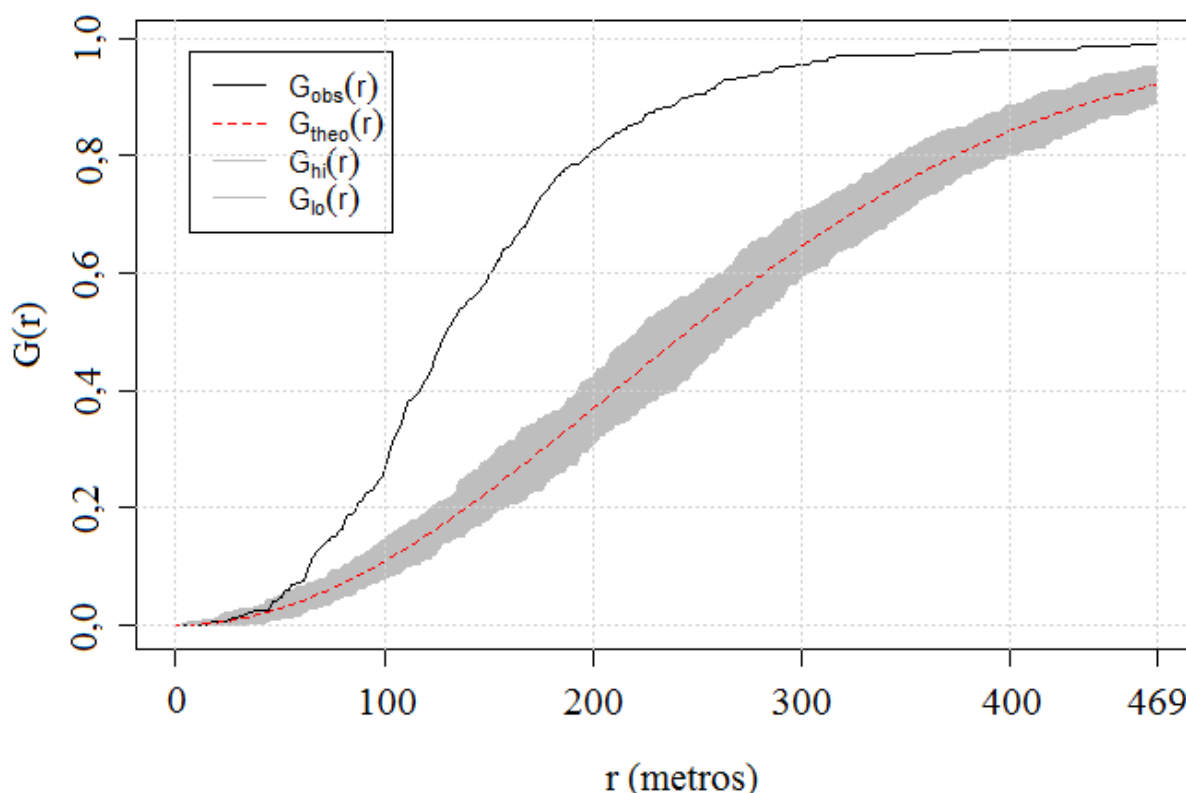
A espacialidade é uma propriedade de dependência e influência entre vizinhos. Se o valor da variável investigada em um município, por exemplo, não está relacionada ao valor da mesma em municípios vizinhos, conclui-se que a variável em questão possui independência espacial entre esses locais sendo, portanto, variável geográfica, mas não espacial. Seja por exemplo o *evento 1* a ocorrência de um caso de dengue (variável) em um *município A* (lugar) e o *evento 2* seja ocorrência de dengue em um *município B*. Se ocorre dengue em *A* e em *B* diz-se que há possibilidade de dependência espacial entre os eventos-lugares e, caso exista, tais eventos serão informações espaciais. Caso contrário, assume-se que não há dependência entre os eventos e a informação será apenas geográfica e não espacial (FERREIRA, 2014, p. 55-56).

A espacialidade é uma característica imprescindível da informação geográfica neste trabalho. A informação geográfica é composta por dados reais de inspeções em campo georreferenciados. *A priori*, conforme supracitado efetua-se uma análise exploratória para verificar se a variável geográfica *UCs Irregulares* é também uma variável espacial. É preciso constatar primeiramente se há evidência de que o espaço exerce influência na disposição espacial das PNTs no município. Caso não exista dependência espacial, o uso da análise espacial não contribui para melhor compreensão do fenômeno em estudo.

A fim de constatar a espacialidade, compara-se a distribuição dos casos de perdas com uma distribuição aleatória – comumente uma distribuição de Poisson como se mostra na seção subsequente.

5.4.3 Função de Distribuição Cumulativa

Com o intuito de medir o grau de similaridade com o padrão CSR, várias funções são utilizadas. Na Figura 20 implementa-se a *função G* estimada, denominada função $\hat{G}(r)$ empírica – representada na Figura 20 por $G_{obs}(r)$ – para todas as UCs flagradas em irregularidades de 2009 a 2012 – vide seção 3.3.3.1. Observa-se que a função $\hat{G}(r)$ empírica encontra-se acima do envelope de cor cinza para as distâncias r entre 100 a 400 metros. Esse resultado indica que as UCs com PNTs estão dispostas espacialmente em aglomerados nesse intervalo de distâncias. O envelope cinza é delimitado superiormente por $G_{hi}(r)$, inferiormente por $G_{lo}(r)$ e contém a função $G(r)$ teórica que é representada por $G_{theo}(r)$ – linha pontilhada em vermelho. Depreende-se da Figura 20 que, a partir de um raio $r > 469$ metros de qualquer transformador ao qual se conecta pelo menos uma UC com PNT, há no mínimo mais um transformador em igual situação.

Figura 20 – Função $G(r)$ ou função de distribuição cumulativa de UCs com PNTs.

Fonte: Elaboração do próprio autor.

Observa-se que a curva da *função G* empírica $G_{obs}(r)$ encontra-se fora do envelope cinza – acima dele. Isso é uma evidência de que as PNTs não estão distribuídas aleatoriamente no município, ou seja, não seguem o padrão CSR. Conforme discutido na seção anterior, além de ser informação geográfica, as PNTs são também informação espacial. Esse resultado é uma indicação de que o uso da análise espacial pode contribuir para melhor compreensão do problema das PNTs.

A partir de tal resultado, surgem as seguintes questões: Por que as PNTs estão dispostas em aglomerados no município? As características de cada subárea do município têm influência na quantidade de UCs irregulares nessas localidades? Essas questões serão abordadas no decorrer deste capítulo.

5.4.4 Mapas de Densidade Anual de Perdas Não Técnicas

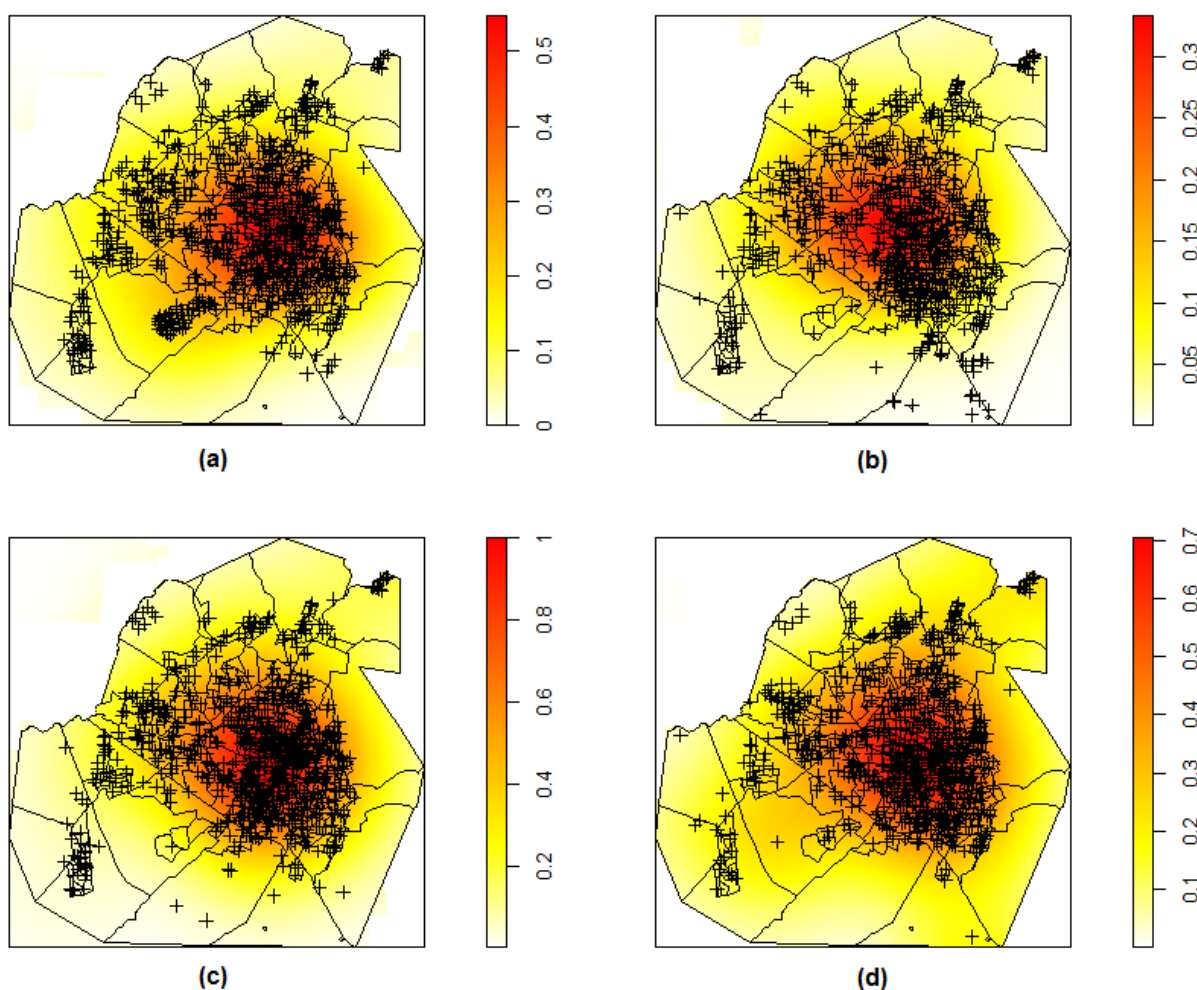
Nesta seção analisam-se os dados de inspeções por meio do estimador de intensidade não paramétrico ponderado descrito na seção 3.3.2.2 – equação (9). No estimador de intensidade não ponderado, todos os pontos interpolados têm igual ponderação. No estimador de intensidade ponderado, os pontos possuem diferentes ponderações. Conforme mencionado em seções anteriores, as UCs irregulares estão agrupadas por transformadores, pois os mes-

mos, ao contrário das UCs, estão georreferenciados. A ponderação está associada a cada transformador e é atribuída às UCs que se conectam a ele. Neste trabalho, a ponderação é a razão entre o número de UCs irregulares conectados em cada transformador pelo número total de UCs conectadas nele.

Os transformadores da rede de distribuição estão georreferenciados. Na análise de UCs irregulares, visualizam-se todos os transformadores que possuem pelo menos uma UC irregular conectada nele.

Na Figura 21, mostram-se os mapas de densidade de perdas para os anos consecutivos de 2009 a 2012. A densidade espacial de PNTs foi normalizada em relação à máxima densidade espacial atingida nos anos avaliados. As cruzes representam as inspeções realizadas.

Figura 21 (a), (b), (c) e (d) – Mapas de densidade espacial ponderada e normalizada de UCs irregulares em 2009, 2010, 2011 e 2012, respectivamente. As cruzes representam as localizações das inspeções realizadas.



Fonte: Adaptado de Faria, Melo e Padilha-Feltrin (2014a).

Da Figura 21 infere-se que, em termos globais, não há modificação significativa na localização das UCs irregulares e das inspeções realizadas ao longo dos anos. Ambas localizam-se predominantemente no centro do município – região de maior densidade de carga. Somente em 2012 é que houve maior espalhamento da densidade de perdas devido à elevada taxa de sucesso das inspeções neste ano de 19,3% conforme consta na Tabela 9.

Os mapas de densidade de perdas são úteis para visualização de áreas críticas no município. Por meio deles, é possível avaliar cada subárea individualmente e planejar campanhas de prevenção e combate às PNTs específicas para cada região. Pode-se também avaliar a taxa de sucesso das inspeções por região e melhor distribuí-las ao longo da cidade. Através desses mapas, pode-se também acompanhar a evolução da densidade de perdas ao longo do espaço e do tempo.

Observa-se que as concessionárias de energia priorizam algumas regiões da cidade em detrimento a outras em relação ao número de UCs inspecionadas. No entanto, as inspeções deveriam ser melhor distribuídas de modo a não deixar áreas descobertas que se tornam vulneráveis à expansão das PNTs.

Por fim, ressalta-se que nas extremidades do mapa, a quantidade de pontos é muito pequena, consequentemente, qualquer medida nesta área será pouco robusta (DIGGLE, 1983).

5.5 ESTUDO DE CASO 1: ESTIMAÇÃO DA DISTRIBUIÇÃO ESPACIAL DAS PERDAS NÃO TÉCNICAS VIA GAM

Nesta seção estima-se a distribuição espacial das PNTs levando em consideração as características do lugar onde elas ocorrem. Por hipótese, assume-se que tais características contribuem para explicar a disposição espacial das mesmas.

As PNTs são representadas nesta seção como eventos pontuais considerando as variáveis socioeconômicas e da rede de distribuição das subáreas onde as mesmas estão situadas. Essas são as variáveis independentes (ou explicativas) de um GAM que é associado a um estudo de *casos-controles*. Este estudo baseia-se na comparação entre dois processos espaciais pontuais marcados – vide seção 3.3.4. Um desses processos corresponde aos *casos* que são compostos pelo conjunto das UCs flagradas em irregularidades. O outro processo de pontos refere-se aos *controles* que são constituídos por UCs regulares selecionadas aleatoriamente a partir da base de clientes da concessionária.

5.5.1 Estudo de *Casos-Controles*

Uma das diversas aplicações da epidemiologia em outras áreas consiste em associar doenças contagiosas com crimes, pois a ocorrência de doenças contagiosas e de crimes seguem um padrão similar (RANSFORD; KANE; SLUTKIN, 2012).

Nessa linha, levando em conta que fraudes na rede de distribuição também são uma modalidade de crime, este trabalho utiliza um estudo denominado *casos-controles* comumente empregado na área de epidemiologia e no mapeamento de doenças (DIGGLE; ROWLINGSON, 1994; KELSALL; DIGGLE, 1995; KELSALL; DIGGLE, 1998; KIFFER et al., 2011; SHIMAKURA et al., 2001). Em epidemiologia, comparam-se pessoas doentes (os *casos*) com pessoas saudáveis (os *controles*). Considera-se que todas elas foram expostas a fatores de risco da doença no passado.

Conforme Bivand; Pebesma e Gómez-Rubio (2008, p. 173), a distribuição dos *casos* é influenciada pela distribuição heterogênea da população. Por isso, é preciso estimar a distribuição espacial da mesma e comparar aos *casos* existentes. Os *controles* representam a variação espacial da população que; neste estudo, são todas as UCs regulares – *a priori*, podem ocorrer PNTs em qualquer UC regular da cidade.

A localização das UCs com perdas (*casos*) e das UCs sem perdas ou regulares (*controles*) são os dados de entrada de um GAM. Pretende-se analisar a variação espacial das probabilidades de se encontrarem *casos* (UCs com PNTs) por subáreas ou setores censitários do município.

Neste estudo de *casos-controles*, utilizou-se um esquema de amostragem (5:1) análogo à Diggle e Rowlingson (1994); ou seja, para cada *caso* de UC com PNT existem cinco UCs regulares – *controles*. Nesse contexto, os *casos* são as 1.240 UCs flagradas em irregularidades e os *controles* são 6.200 UCs regulares selecionadas aleatoriamente entre todas as UCs regulares do município.

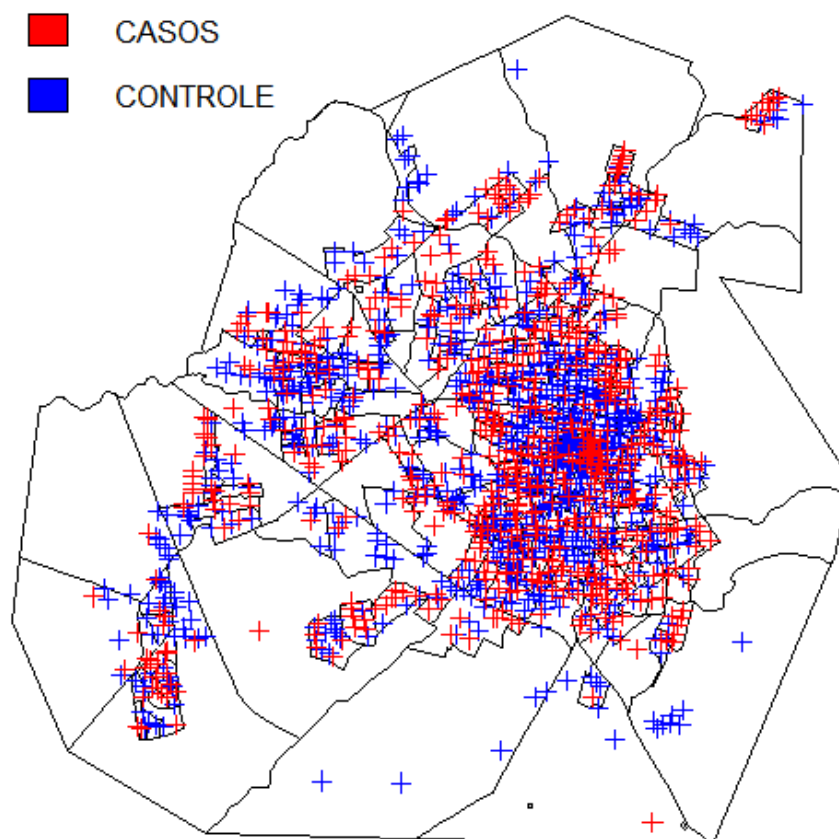
Na Figura 22 estão representadas as amostras dos eventos *casos-controles* no mapa da zona urbana do município em estudo.

5.5.2 Modelo Aditivo Generalizado (GAM) para Análise Espacial das PNTs

As probabilidades de PNTs são obtidas via simulação do GAM a partir do qual se produz um mapa de probabilidades de perdas. Entre as centenas de variáveis socioeconômicas do censo demográfico IBGE (2010) para cada setor censitário, foram selecionadas três variáveis para este primeiro estudo de caso. São elas: %UCs Alugadas, Média de Moradores e Renda Média. Além das variáveis do censo, há duas variáveis da rede de distribuição selecio-

nadas: *PERTRAFO* e *NTRAFO*. Essas são as variáveis independentes ou explicativas utilizadas no modelo de regressão espacial GAM e estão descritas na Tabela 11.

Figura 22 – Distribuição espacial dos eventos *Casos-Controles* na região urbana do município sob análise.



Fonte: Elaboração do próprio autor.

5.5.3 Mapa de Probabilidades de Perdas Não Técnicas

Neste trabalho, utiliza-se a biblioteca do *software* R[®] *mgcv* para executar o GAM (R DEVELOPMENT CORE TEAM, 2011; WOOD, 2006). Nessa biblioteca está implementado o estimador não paramétrico apresentado em Kelsall e Diggle (1998).

O GAM assume um efeito linear do total de cinco variáveis. Três variáveis são oriundas do censo demográfico IBGE (2010); as demais advêm dos dados de inspeções e dados da rede de distribuição. Três variáveis referem-se às características socioeconômicas das subáreas do município. São elas: *%UCs Alugadas*, *Média de Moradores* e *Renda Média*. *PERTRAFO* está relacionada à concentração das PNTs por transformadores e *NTRAFO* está associada à extensão da rede de distribuição associada a cada subárea do município.

O *p-valor* corresponde à menor significância para a qual a hipótese nula H_0 de não significância das variáveis pode ser rejeitada. Logo, H_0 será rejeitada se o *p-valor* for menor

ou igual a um nível de significância de referência. Neste estudo adota-se nível de significância de 0,1.

Na Tabela 15 apresentam-se as estimativas, desvios padrão e significâncias estatísticas de cada variável independente do GAM. As variáveis significantes estão assinaladas com asterisco. São elas: *PERTRAFO* (*p*-valor inferior a 2×10^{-16}), *Média de Moradores* (*p*-valor de 0,0173), *NTRAFO* (*p*-valor 0,0364) e *%UCs Alugadas* (*p*-valor 0,0861).

Tabela 15 – Estimativas, desvios padrão e significâncias estatísticas das variáveis independentes do GAM.

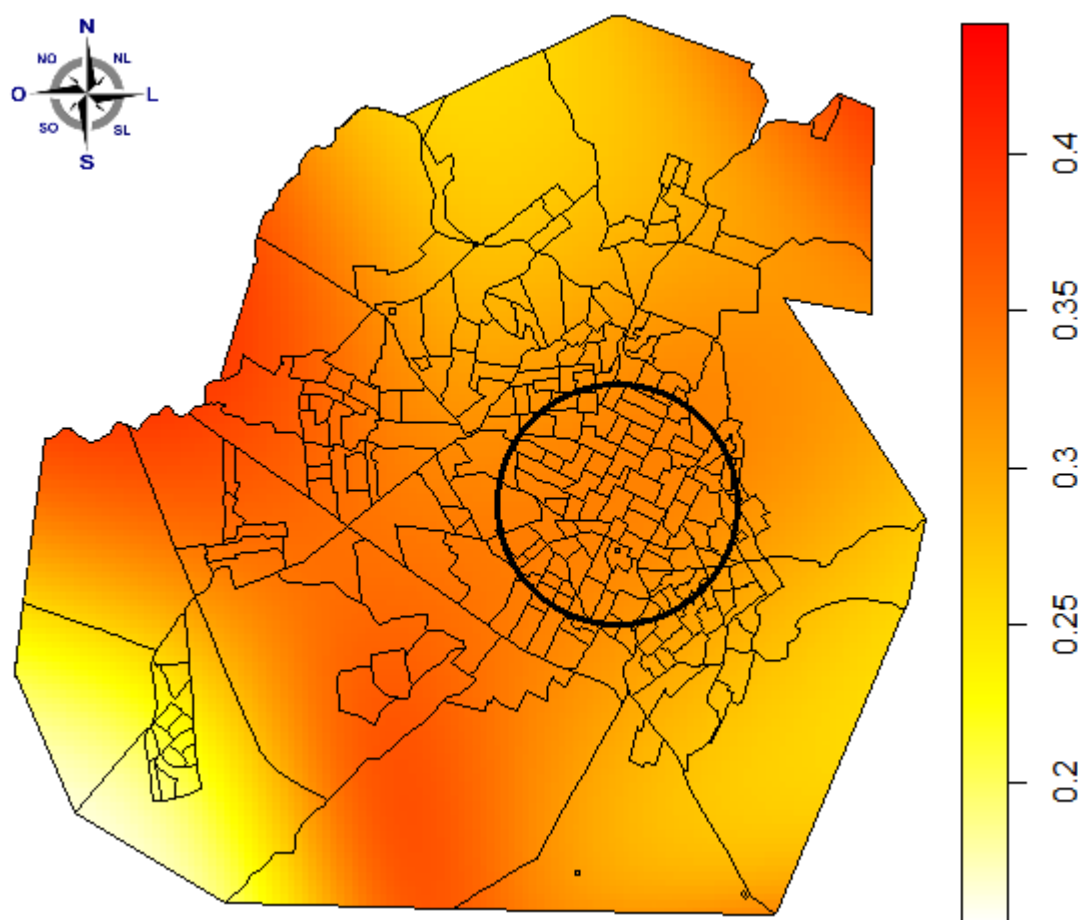
<i>Variáveis</i>	<i>Estimativas (β)</i>	<i>Desvios Padrão (σ)</i>	<i>p-valor</i>
<i>%UCs Alugadas*</i>	0,00302	0,00176	0,0861
<i>Média de Moradores*</i>	0,37790	0,15870	0,0173
<i>Renda Média</i>	−0,00001	0,00006	0,8329
<i>NTRAFO*</i>	−0,00874	0,00418	0,0364
<i>PERTRAFO*</i>	35,55900	1,26800	0,0000

Fonte: Elaboração do próprio autor.

Na Figura 23 mostra-se o principal resultado deste primeiro estudo de caso que é o mapa de probabilidades de perdas gerado a partir da execução do fluxograma da Figura 4 – vide seção 3.3.4.6. Depreende-se que a probabilidade de PNTs é alta na região central do município – região inscrita no círculo – e aumenta à medida que se desloca para a direção oeste – região de maior crescimento do município.

A aplicação do teste estatístico (vide seção 3.3.4.4) indica variação espacial significativa da probabilidade de perdas no município com *p*-valor de 0,01.

Deduzem-se da Figura 23 que as regiões centrais e oeste são aquelas com maior probabilidade de perdas. As subáreas centrais estão em uma zona comercial da cidade, onde a rede de distribuição é mais extensa e; por conseguinte, há elevado número de transformadores. Tal fato é confirmado pela significância da variável *NTRAFO*. Além disso, nessa região, há elevada densidade populacional. Muitas UCs pertencem a edifícios alugados por comerciantes. Tais constatações são ratificadas pela significância das variáveis *Média de Moradores* e *%UCs Alugadas*. Em último, a significância da variável *PERTRAFO* é um indicativo da concentração das UCs irregulares em transformadores com altas PNTs percentuais (com relação ao número total de UCs conectadas nele) que estão localizadas majoritariamente na região central e oeste do município.

Figura 23 – Mapa de probabilidades para a distribuição espacial das perdas não técnicas.

Fonte: Elaboração do próprio autor.

A variável *Renda Média* (*p*-valor 0,8329) não apresenta associação estatisticamente significativa com as PNTs no município analisado. No entanto, conforme Bailey e Gatrell (1995), embora existam variáveis que não possuam significância estatística, não é possível afirmar conclusivamente que as mesmas não influam de algum modo na presença das PNTs. A interpretação da significância estatística deve ser feita com critério no momento da análise exploratória do problema. Ademais, a significância de cada variável é modificada na medida em que se inclui (ou exclui) variáveis da análise. A significância depende também da amostragem do conjunto de *casos-controles*.

Observa-se que o termo residual da função de suavização *splines* não é significativo (*p*-valor 0,327). Isso sugere que o GAM proposto com essas variáveis é suficiente para explicar a variação espacial das PNTs para o caso particular do município em estudo.

5.6 ESTUDO DE CASO 2: ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS VIA GAM E CADEIAS DE MARKOV

Nesta seção, apresentam-se os resultados da estimação da variação espaço-temporal das PNTs via modelo de regressão espacial GAM combinada com as cadeias de Markov. A metodologia foi apresentada em detalhes na seção 4.4.2.

5.6.1 Estimação da Variação Espaço-Temporal das Perdas Não Técnicas em um Município do Interior Paulista

Após a apresentação da metodologia na seção 4.4.2, nesta seção apresentam-se os detalhes da implementação em uma cidade do interior paulista. Os dados utilizados são aqueles descritos na seção 5.2.

A partir dos dados de inspeções, constrói-se uma base de dados *casos-controles* análoga àquela construída na seção 5.5.1. Neste estudo de *casos-controles*, o conjunto dos *casos* é formado por todas as UCs irregulares de 2009 até 2011 – o ano de 2012 é reservado para validar a metodologia proposta. O conjunto dos *controles* é formado aleatoriamente por UCs regulares da cidade onde o evento perda não ocorreu.

Adota-se, neste estudo, um esquema de amostragem (9:1), isto é, para cada UC irregular pertencente ao conjunto de *casos*, existem nove UCs regulares pertencente aos *controles*. Os *casos* contêm todas as 707 UCs flagradas em irregularidades de 2009 a 2011; os *controles*, 6.363 UCs regulares selecionadas aleatoriamente a partir das mais de 80 mil UCs do município.

Ressalta-se que o esquema de amostragem (9:1) representa aproximadamente a situação real das inspeções em campo das concessionárias brasileiras, porque, em média, a cada dez inspeções realizadas, encontra-se uma UC irregular (COMETTI; VAREJÃO, 2005). Na Tabela 9, por exemplo, em média, a cada 9,2 inspeções encontra-se uma UC irregular.

5.6.1.1 Descrição das Variáveis utilizadas no Modelo Aditivo Generalizado

Foram utilizadas onze variáveis x independentes na regressão espacial GAM. Tais variáveis e suas respectivas descrições estão na Tabela 11. Elas caracterizam os aspectos socioeconômicos, os aspectos da rede de distribuição e da concentração das PNTs por transformadores e por subáreas do município.

Entre as centenas de variáveis socioeconômicas do censo demográfico IBGE (2010), foram selecionadas sete variáveis para este segundo estudo de caso. São elas: *Média*

de Moradores, Renda Média, %Domicílios Alugados, %Domicílios com Água, %Domicílios com Lixo Coletado, %Domicílios com Pavimentação e %Alfabetizados. As demais variáveis foram extraídas dos dados de inspeções e da rede de distribuição. *PERTRAFO* e *PERAREA* estão relacionadas à concentração de UCs irregulares por transformador e por subárea, respectivamente; *NTRAFO* vincula-se à extensão da rede de distribuição e *CAMPANHA* indica se houve um número elevado de UCs inspecionadas (acima de trinta); todas, por subáreas.

5.6.1.2 Validação da Metodologia

A validação da metodologia é feita através da comparação entre a previsão obtida e o número de UCs com PNTs para o ano de 2012.

Com o intuito de realizar a validação do estado de perda previsto, determina-se novamente o estado de perda para o ano de 2012 a partir do número de PNTs. Considera-se que uma subárea qualquer está no estado *regular* ou *crítico* se o número de PNTs nela for, respectivamente, inferior a três ou superior a oito. Se essa condição não for satisfeita, a subárea encontra-se no estado de *atenção*. Esses limites foram estabelecidos a partir da análise da Figura 19.

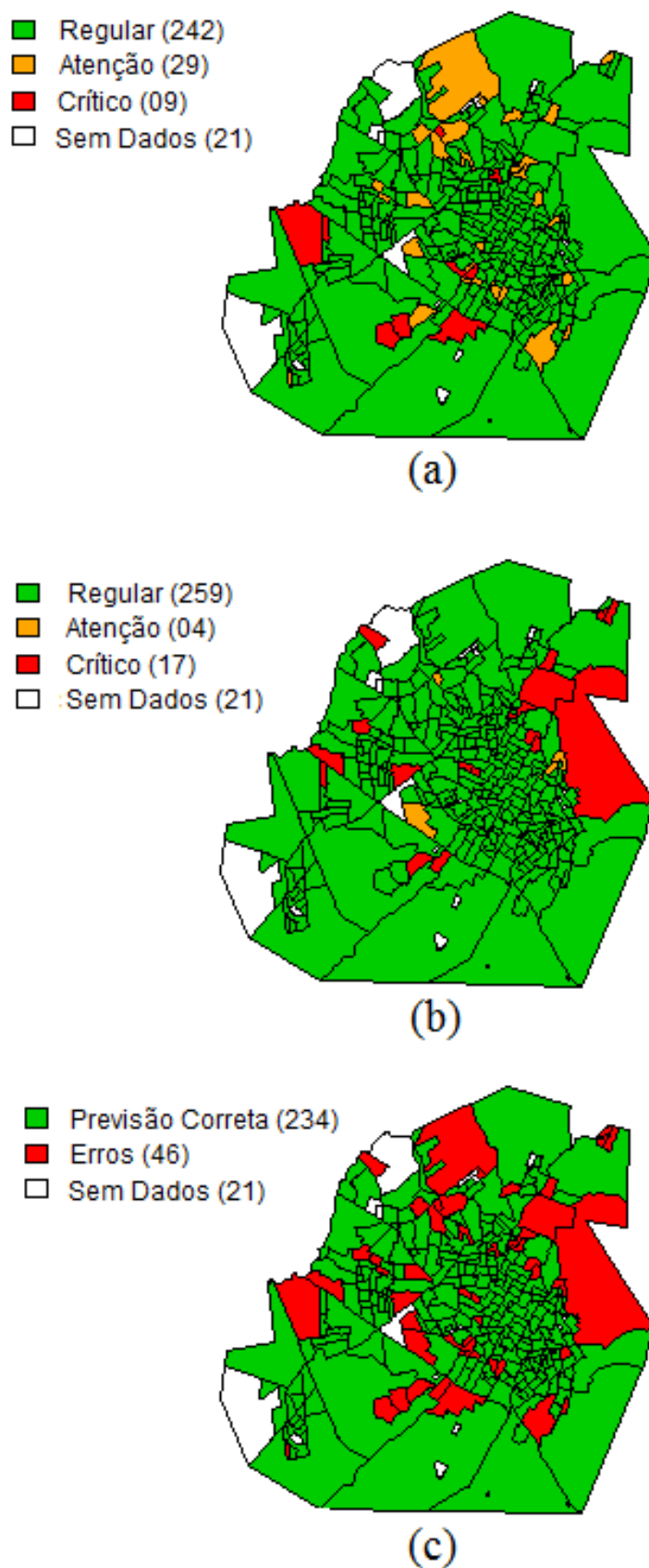
Na Figura 24 apresenta-se o resultado da aplicação da metodologia abordada nesta seção para estimar a variação espaço-temporal das PNTs. Na Figura 24 (a) e (b) mostram-se os estados de perdas por subárea no presente (via GAM) e os estados de perdas previstos (via cadeias de Markov), respectivamente. Conforme supracitado, para validar a metodologia, as subáreas são novamente categorizadas em um dos três estados de perdas conforme número de UCs com PNTs em 2012 e comparadas à previsão obtida na Figura 24 (b). Das 301 subáreas da zona urbana da cidade, foram avaliadas 280. Entre essas, houve 234 acertos (83,6%) e 46 erros (16,4%). Os estados de perdas de 21 subáreas não foram previstos por falta de dados.

5.6.1.3 Variáveis com Significância Estatística

Além de ser fundamental para construção do mapa de probabilidades, por meio da regressão espacial GAM, encontram-se as variáveis que têm relevância estatística e que, portanto, influem nas PNTs. Na Tabela 16 apresentam-se as estimativas, desvios padrão e significâncias estatísticas para as variáveis independentes x do GAM.

O *p-valor* corresponde à menor significância para a qual a hipótese nula H_0 de não significância das variáveis pode ser rejeitada. H_0 será rejeitada se o *p-valor* for menor ou igual ao nível de significância de referência. Neste trabalho, fixa-se o nível de significância α em 10% ou 0,1.

Figura 24 – (a), (b) Estados de perdas atuais (via GAM) e estados de perdas previstos (via cadeias de Markov), respectivamente, por subárea. (c) Validação da metodologia via comparação entre o estado de perda previsto em (b) e o número de PNTs, por subárea, em 2012.



Fonte: Elaboração do próprio autor.

Tabela 16 – Estimativas, desvios padrão e significâncias estatísticas da variáveis do GAM.

<i>Variáveis</i>	<i>Estimativas (β)</i>	<i>Desvio Padrão (σ)</i>	<i>p-valor</i>
<i>Média de Moradores</i>	-0,44960	0,39260	0,25219
<i>Renda Média</i>	-0,00002	0,00010	0,80550
<i>%Domicílios Alugados</i>	-1,13500	1,12600	0,31345
<i>%Domicílios com Água</i>	-1,30900	1,27800	0,30566
<i>%Domicílios com Lixo Coletado</i>	2,52800	1,95400	0,19572
<i>%Domicílios com Pavimentação</i>	-0,48210	0,81350	0,55434
<i>%Alfabetizados</i>	0,39380	2,17400	0,85628
<i>PERTRAFO*</i>	34,65000	1,55800	0,00000
<i>PERAREA</i>	2,05400	1,71300	0,23050
<i>NTRAFO*</i>	-0,01194	0,00745	0,10905
<i>CAMPANHA*</i>	0,51210	0,13350	0,00013

Fonte: Elaboração do próprio autor.

Nesse contexto, as variáveis significantes (assinaladas com um asterisco na Tabela 16), isto é, aquelas que influem nas PNTs no município analisado são: *PERTRAFO* (*p-valor* $2 \cdot 10^{-16}$) e *CAMPANHA* (*p-valor* $1,25 \cdot 10^{-4}$). *NTRAFO* (*p-valor* 0,10905) é considerada marginalmente significativa.

A significância de *PERTRAFO* indica maior concentração das perdas em transformadores com elevado percentual de UCs irregulares conectadas a ele; de *NTRAFO*, indica maior concentração em subáreas com rede de distribuição mais extensa, isto é, com mais transformadores; e de *CAMPANHA*, indica que, conforme esperado, encontram-se mais PNTs em subáreas alvos de campanhas de inspeções (subáreas em tom mais escuro na Figura 18).

Visualiza-se na Figura 18 que em todos os anos analisados (de 2009 a 2012) houve subáreas alvos de campanhas de inspeções na região central do município. Esse fato justifica a maior quantidade de UCs com PNTs encontradas nesta área como se pode ver na Figura 19. Acrescenta-se ainda que a região central, em conformidade com o resultado do GAM, possui elevada densidade populacional e extensa rede de distribuição.

O termo residual da função de suavização *splines* é significativo (*p-valor* 0,0492). Isso significa que pode existir alguma variação espacial residual não explicada pelas variáveis associadas ao GAM.

Os resultados apresentados nesta seção são também abordados em Faria; Melo; Padilha-Feltrin (2015) de forma mais resumida.

5.7 ESTUDO DE CASO 3: ESTIMAÇÃO ESPAÇO-TEMPORAL DAS PERDAS NÃO TÉCNICAS VIA GWR E CADEIAS DE MARKOV

As PNTs são representadas nesta seção via análise espacial de áreas utilizando uma abordagem a nível local de modelos geograficamente ponderados ou *geographically weighted* (GW) que levam em consideração a heterogeneidade dos dados espaciais.

Aplica-se a técnica de regressão local GWR para dados agregados por subáreas – vide seções 4.1 a 4.3. Objetiva-se produzir mapas de probabilidade de perdas no presente e no futuro, via GWR e via cadeias de Markov, respectivamente, indicando as subáreas vulneráveis às perdas, análogo à seção 5.6.

Na seção 5.6, no entanto, os mapas de probabilidade de perdas no presente foram produzidos via GAM; nesta, o serão via GWR. O GAM é um modelo de regressão adequado para modelar as perdas em nível de pontos; o GWR, em nível de áreas.

Uma diferença adicional entre a modelagem desta seção frente à seção anterior é que, conforme ilustrado na Figura 9, seção 4.4.2.1, no segundo estudo de caso, considerou-se a vizinhança tradicional baseada em subáreas fronteiriças (vide seção 4.1.1), as quais possuem frequentemente limites geográficos em comum. Nessa seção implementa-se a *vizinhança semântica*, segundo a qual, as áreas vizinhas são aquelas com atributos semelhantes – na seção 4.1.2 abordam-se as diferenças entre o estilo de vizinhança tradicional e a *vizinhança semântica*.

Emprega-se a *vizinhança semântica* para construção da matriz de ponderação W . Todos os resultados apresentados nesta seção são diretamente influenciados pela matriz de ponderação que estabelece as relações entre todas as subáreas – vide seção 3.4.3 para definição da matriz de ponderação.

5.7.1 Construção da Vizinhança Semântica

Conforme abordado na seção 4.1.2, a *vizinhança semântica* é construída via SOM (vide Apêndice A) a partir do qual se determinam as classes ou *clusters* de subáreas com atributos semelhantes.

Neste terceiro estudo de caso, os atributos avaliados por subáreas ou as variáveis explicativas do fenômeno *Taxa de Sucesso de Perdas* (variável a ser explicada ou variável dependente – equação 46) selecionadas são: *Densidade de Carga* (kVA/km²), *Renda Média* (em reais) e *%UCs Alugadas*. A primeira variável refere-se a uma característica da rede de

distribuição; as demais são oriundas do censo demográfico IBGE (2010). Todas as variáveis foram descritas na Tabela 11.

Por hipótese, tais variáveis são aquelas que possuem maior aderência, síntese e maior relação quantitativa e qualitativa com as PNTs no município em estudo. A *Densidade de Carga*, por exemplo, possui maior abrangência do que a variável *NTRAFO* que apenas contabiliza o número de transformadores por áreas e não considera a potência nominal dos mesmos. A variável *Renda Média* possui relação de proporcionalidade direta com outras variáveis socioeconômicas tais como: *%UCs com Lixo Coletado*; *%UCs com Pavimentação*, *%UCs com Água* e *%Alfabetizados*. Por fim, a variável *%UCs Alugadas*, por hipótese, representa uma condição desejável para aquele que pretende executar uma fraude, pois o mesmo possui menor risco de ser apreendido morando em um domicílio alugado do que se fosse proprietário do domicílio no qual reside.

Na Figura 25, após execução do SOM, apresentam-se os *clusters* ou classes de subáreas vizinhas e o número de subáreas ou setores censitários em cada classe. Considera-se a semelhança das mesmas com respeito às variáveis independentes ou explicativas. A saber: *Densidade de Carga*, *Renda Média* e *%UCs Alugadas*. Observa-se visualmente que, na *vizinhança semântica*, áreas vizinhas não precisam ser necessariamente áreas fronteiriças como ocorre na maior parte dos trabalhos avaliados que envolvem a vizinhança tradicional baseada em funções de *kernel*.

Após a etapa de construção dos *clusters*, determinam-se os elementos w_{ij} da matriz de ponderação W . Eles são uma medida do nível de semelhança entre os atributos de cada subárea i e j com respeito ao vetor de pesos do neurônio que representa o *cluster* ao qual essas subáreas pertencem. O processo para determinação dos elementos da matriz W foi abordado pormenorizadamente na seção 4.1.2.

Na Tabela 17 apresentam-se os principais parâmetros de calibração do SOM para obtenção da *vizinhança semântica* e dos elementos da matriz de ponderação W .

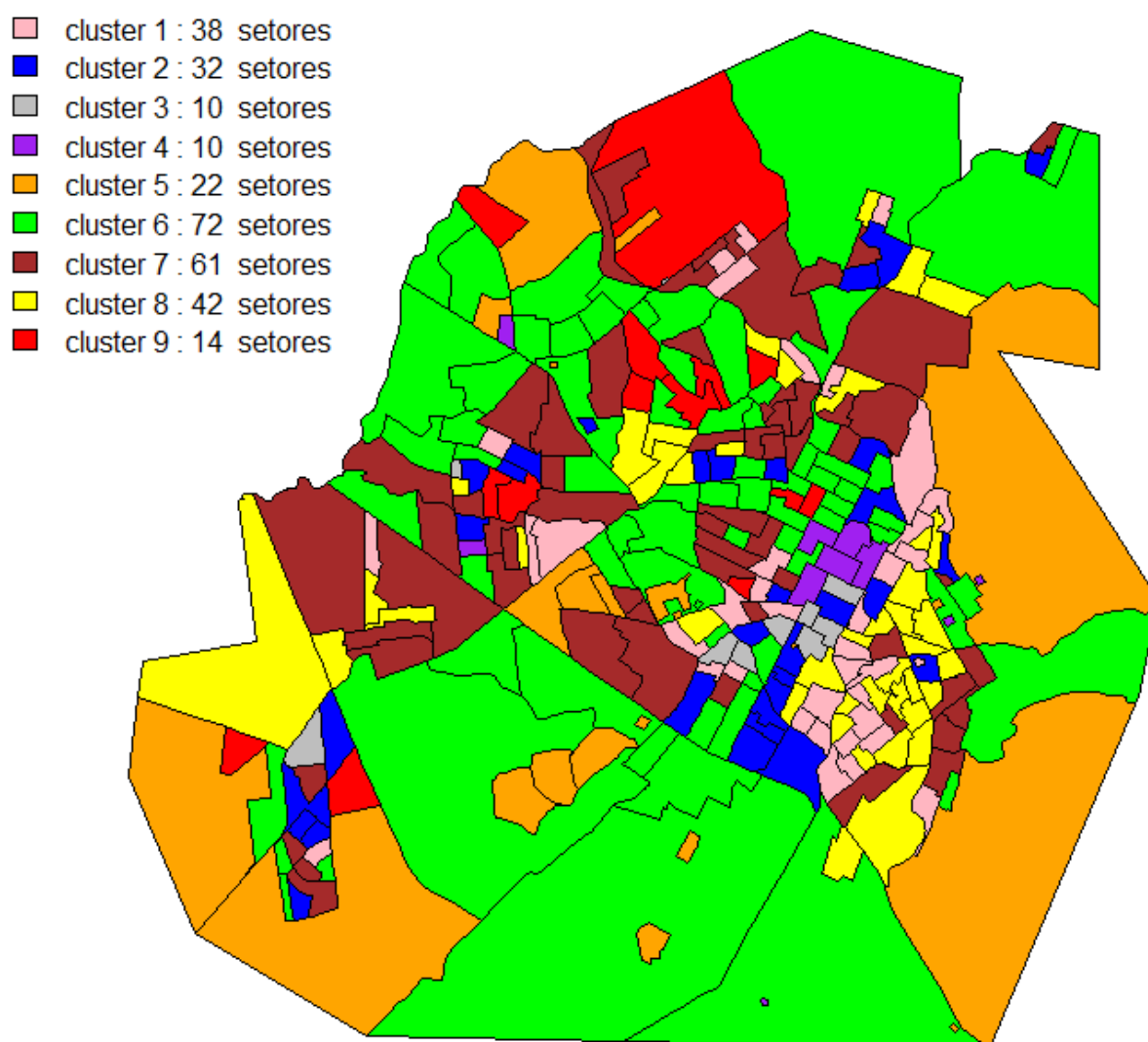
5.7.2 Resumo Estatístico Geograficamente Ponderado

Neste estudo de caso implementa-se uma regressão GW, sendo a *Taxa de Sucesso de Perdas* a variável dependente e as variáveis *Densidade de Carga*, *Renda Média* e *%UCs Alugadas*, as variáveis independentes ou explicativas. Todavia, antes de implementá-lo, conforme Gollini et al. (2015), é preciso realizar um resumo estatístico geograficamente ponderado ou GW o qual é considerado como um precursor vital que antecede a aplicação de um modelo GW.

Tabela 17 – Parâmetros para treinamento do mapa auto-organizável de Kohonem – SOM.

<i>Treinamento</i>	Aprendizado competitivo
<i>Tipo de treinamento</i>	Não-supervisionado
<i>Arquitetura da rede</i>	Reticulado hexagonal
<i>Topologia da rede</i>	Grade bidimensional 6 × 6
<i>Tipo de aprendizagem</i>	On-line
<i>Crítério de convergência</i>	Variação da magnitude dos vetores de peso da rede
<i>Função de ajuste dos vizinhos</i>	Gaussiana
<i>Topologia da vizinhança</i>	Círculo de raio unitário e fixa
<i>Taxa de treinamento</i>	Decresce exponencialmente com o acréscimo das épocas de treinamento

Fonte: Elaboração do próprio autor.

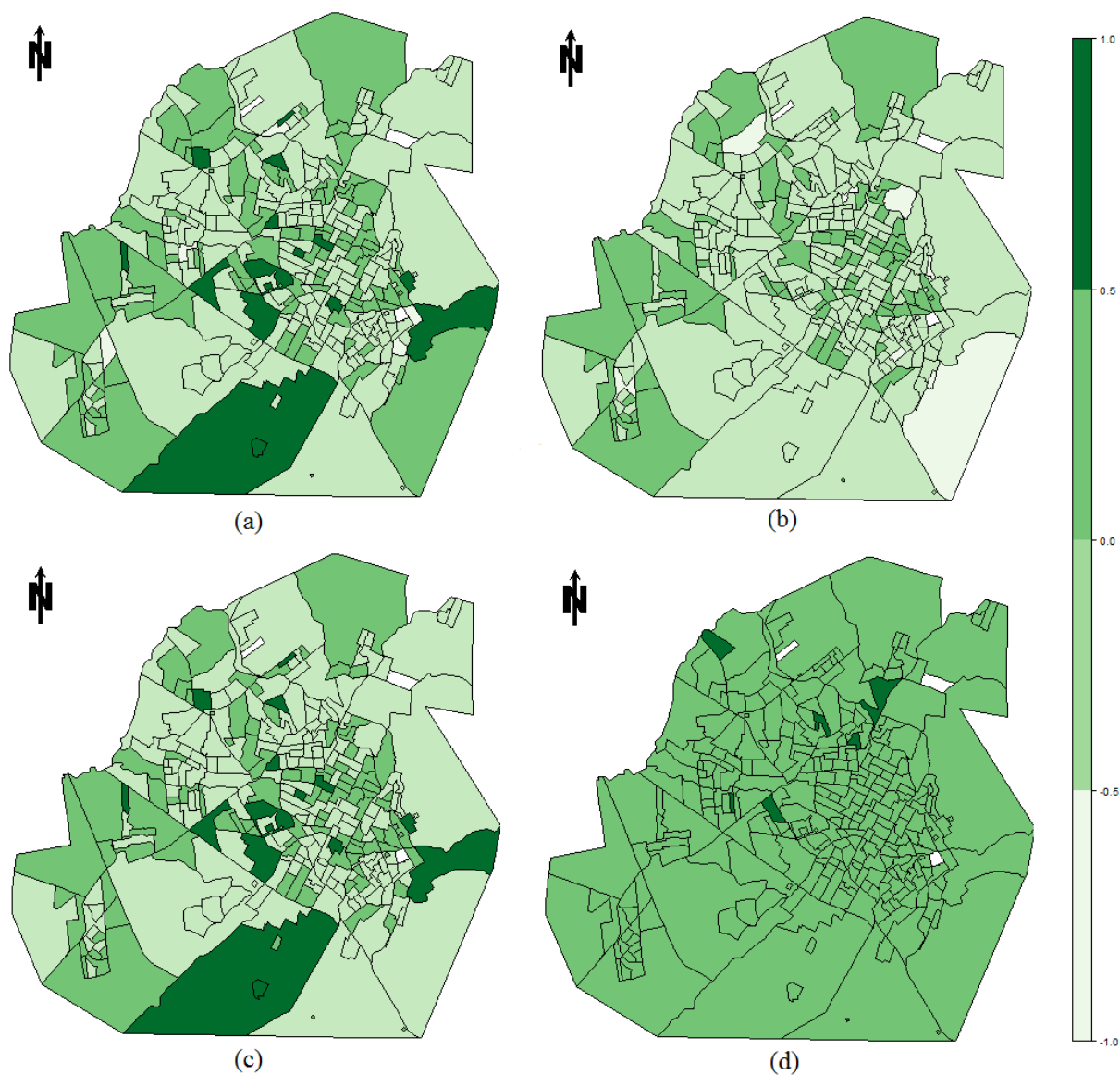
Figura 25 – Subáreas vizinhas ou semelhantes após execução do SOM.

Fonte: Elaboração do próprio autor.

A correlação GW permite, por exemplo, uma avaliação preliminar da relação não estacionária entre as variáveis dependentes e independentes de uma regressão GW. Também viabiliza uma avaliação da colinearidade local entre as variáveis independentes de uma regressão GW. Na seção 4.1.3 apresentou-se algumas métricas estatísticas GW usuais que serão empregadas nesta seção.

Na Figura 26 (a), (b) e (c) apresenta-se, respectivamente, a correlação GW entre a variável dependente *Taxa de Sucesso de Perdas* e as variáveis independentes *Densidade de Carga*, *Renda Média* e *%UCs Alugadas*. Seguem algumas inferências:

Figura 26 – Correlação GW entre a variável dependente *Taxa de Sucesso de Perdas* e as variáveis explicativas *Densidade de Carga* (a); *Renda Média* (b) e *%UCs Alugadas* (c). Em (d) a correlação GW entre as variáveis explicativas *Densidade de Carga* e *Renda Média*.



Fonte: Elaboração do próprio autor.

- ❖ há relação não estacionária entre *Taxa de Sucesso de Perdas* e a *Densidade de Carga* – Figura 26 (a) – com correlação GW moderadamente positiva nas regiões central, sul e sudoeste do município;
- ❖ há relação não estacionária entre *Taxa de Sucesso de Perdas* e a *Renda Média* – Figura 26 (b) – com correlação GW moderadamente positiva nas regiões central e sudoeste do município e moderadamente negativa nas regiões ao redor do centro do mesmo;
- ❖ e por fim, relação não estacionária entre *Taxa de Sucesso de Perdas* e a *%UCs Alugadas* – Figura 26 (c) – com correlação GW moderadamente positiva com maior intensidade nas regiões central, sul e sudoeste do município.

A multicolinearidade é um problema comum em regressões, no qual as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas. A ausência da multicolinearidade é uma das premissas para estabelecer um modelo de regressão correto. Na Figura 26 (d) apresenta-se a correlação GW entre as variáveis independentes ou explicativas *Densidade de Carga* e *Renda Média*. Observa-se que poucas subáreas apresentam forte correlação GW fortemente positiva. As demais variáveis explicativas foram avaliadas e igualmente não apresentaram forte correlação GW; portanto, o problema da multicolinearidade não é um fator relevante para este terceiro estudo de caso.

5.7.3 Regressão Geograficamente Ponderada

Neste estudo de caso, estuda-se a variabilidade local da variável *Taxa de Sucesso de Perdas* que é a variável dependente da regressão GW cujas variáveis independentes são: *Densidade de Carga*, *Renda Média* e *%UCs Alugadas*.

Na Tabela 18 apresentam-se as estimativas, desvio padrão e *p-valor* dos coeficientes de uma regressão linear global. A *Taxa de Sucesso de Perdas* está positivamente relacionada com as variáveis independentes, a exceção é a *Renda Média*. A variável *Renda Média* da Tabela 18 (assinalada com asterisco) é a única com significância estatística (*p-valor* < 0,1).

Tabela 18 – Estimativas, desvios padrão e significâncias estatísticas dos coeficientes das variáveis explicativas da regressão linear global.

<i>Variáveis</i>	<i>Estimativas (β)</i>	<i>Desvios Padrão (σ)</i>	<i>p-valor</i>
<i>Densidade de Carga</i>	0,81286	1,77184	0,6467
<i>Renda Média*</i>	-0,13912	0,07609	0,0685*
<i>%UCs Alugadas</i>	0,03317	0,09761	0,7342

Fonte: Elaboração do próprio autor.

Na Tabela 19, visualiza-se a variação dos coeficientes estimados para as variáveis independentes do GWR. Conforme Bivand, Pebesma e Gómez-Rubio (2008, p. 307), a variação local dos coeficientes do GWR demonstra a viabilidade desta técnica para o presente estudo de caso. Caso não existisse essa variação, o GWR seria desinteressante para esta base de dados e um modelo mais simplificado poderia ser utilizado em seu lugar; um modelo de regressão global, por exemplo.

Tabela 19 – Resumo dos coeficientes estimados das variáveis independentes do GWR.

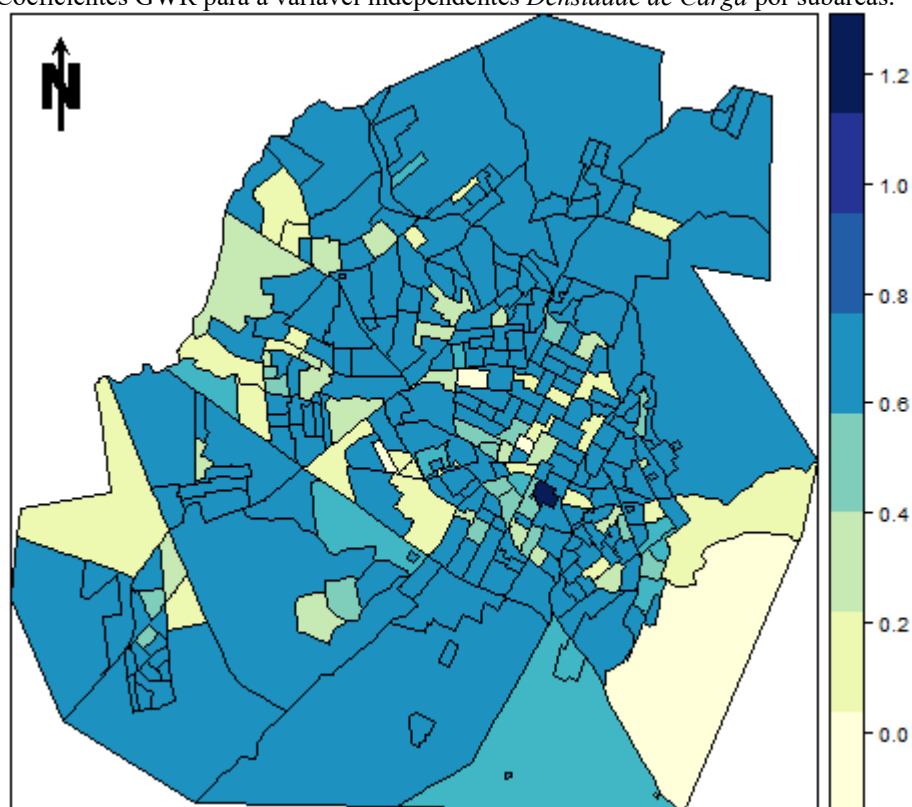
<i>Variáveis</i>	<i>Mínimo</i>	<i>1º Quartil</i>	<i>Mediana</i>	<i>3º Quartil</i>	<i>Máximo</i>
<i>Densidade de Carga</i>	−0,05617	0,62150	0,81290	0,82050	1,21900
<i>Renda Média</i>	−0,18260	−0,14600	−0,13910	−0,13890	−0,1112
<i>%UCs Alugadas</i>	−0,00044	0,02834	0,03318	0,03343	0,0359

Fonte: Elaboração do próprio autor.

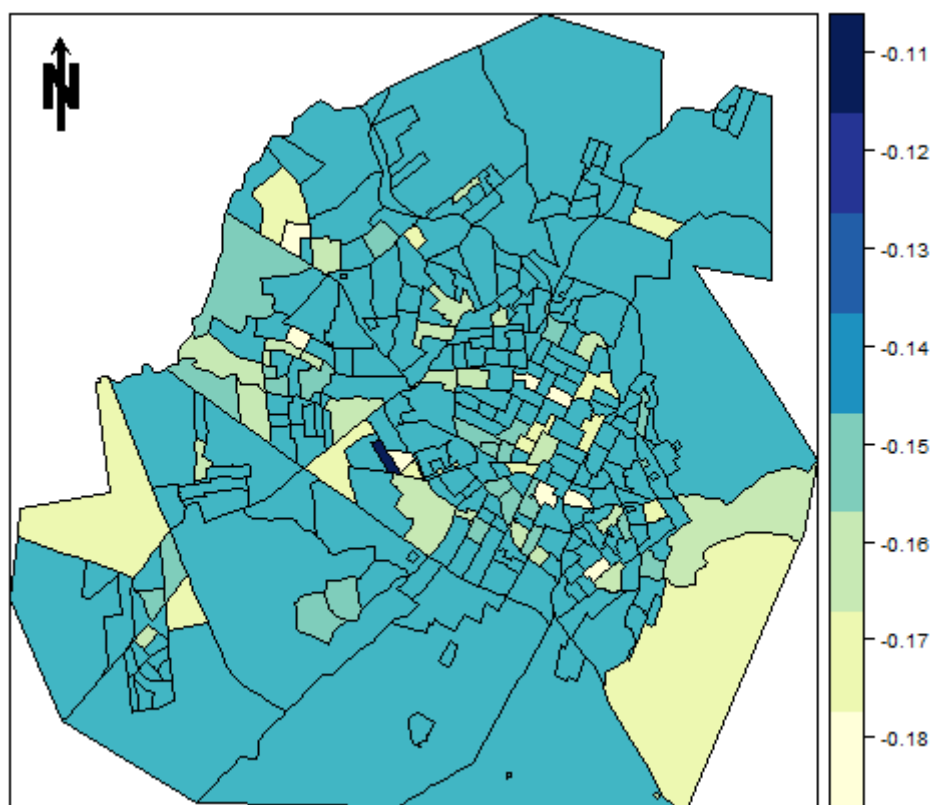
Comparam-se os resultados das Tabelas 18 e 19 para a regressão linear global e o GWR, respectivamente. Os coeficientes estimados para a *Densidade de Carga*, por exemplo, têm uma variação geográfica ou amplitude de −0,05617 até 1,219, conforme Tabela 19. O coeficiente para a regressão global para esta mesma variável é de 0,81286 conforme Tabela 18. O coeficiente da regressão global está inserido na faixa dos coeficientes da regressão espacial GWR.

Nas Figuras 27 a 29 mapeiam-se as estimativas dos coeficientes da regressão espacial GWR para as variáveis independentes *Densidade de Carga*, *Renda Média* e *%UCs Alugadas* da Tabela 19.

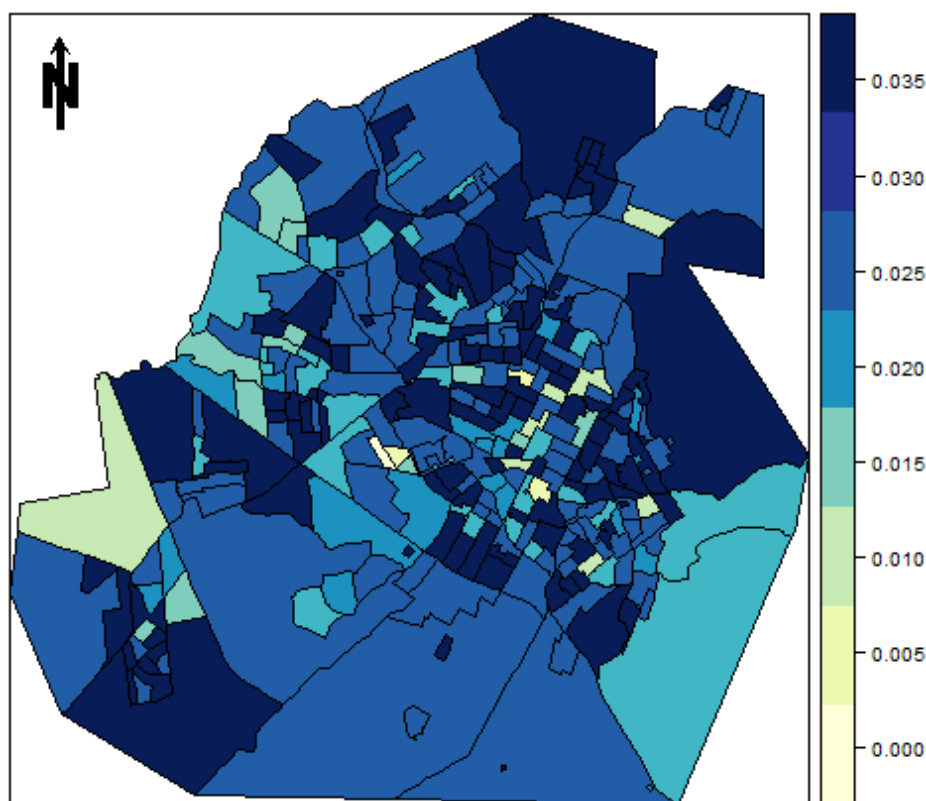
Nesta seção obtêm-se os mapas de probabilidade de perdas no presente obtidos de forma análoga ao segundo estudo de caso. Os dados de entrada para o GWR são dados de inspeções (número de inspeções e o número de UCs irregulares por subáreas) pertencentes aos anos de 2009 a 2011 – o ano de 2012 é reservado para validar a proposta – e os dados socioeconômicos e da rede de distribuição – variáveis *Densidade de Carga*; *Renda Média* e *%UCs Alugadas*. Empregam-se as cadeias de Markov conforme abordado na seção 4.4.2 para levar as probabilidades de perdas do presente (obtidas via GWR) para o futuro. Dessa forma, estima-se o mapa de probabilidades de perdas no futuro (em 2012) o qual é confrontado com as UCs irregulares reais encontradas em campo nesse mesmo ano.

Figura 27 – Coeficientes GWR para a variável independentes *Densidade de Carga* por subáreas.

Fonte: Elaboração do próprio autor.

Figura 28 – Coeficientes GWR para a variável independente *Renda Média* por subáreas.

Fonte: Elaboração do próprio autor.

Figura 29 – Coeficientes GWR para a variável independente %UCs Alugadas por subáreas.

Fonte: Elaboração do próprio autor.

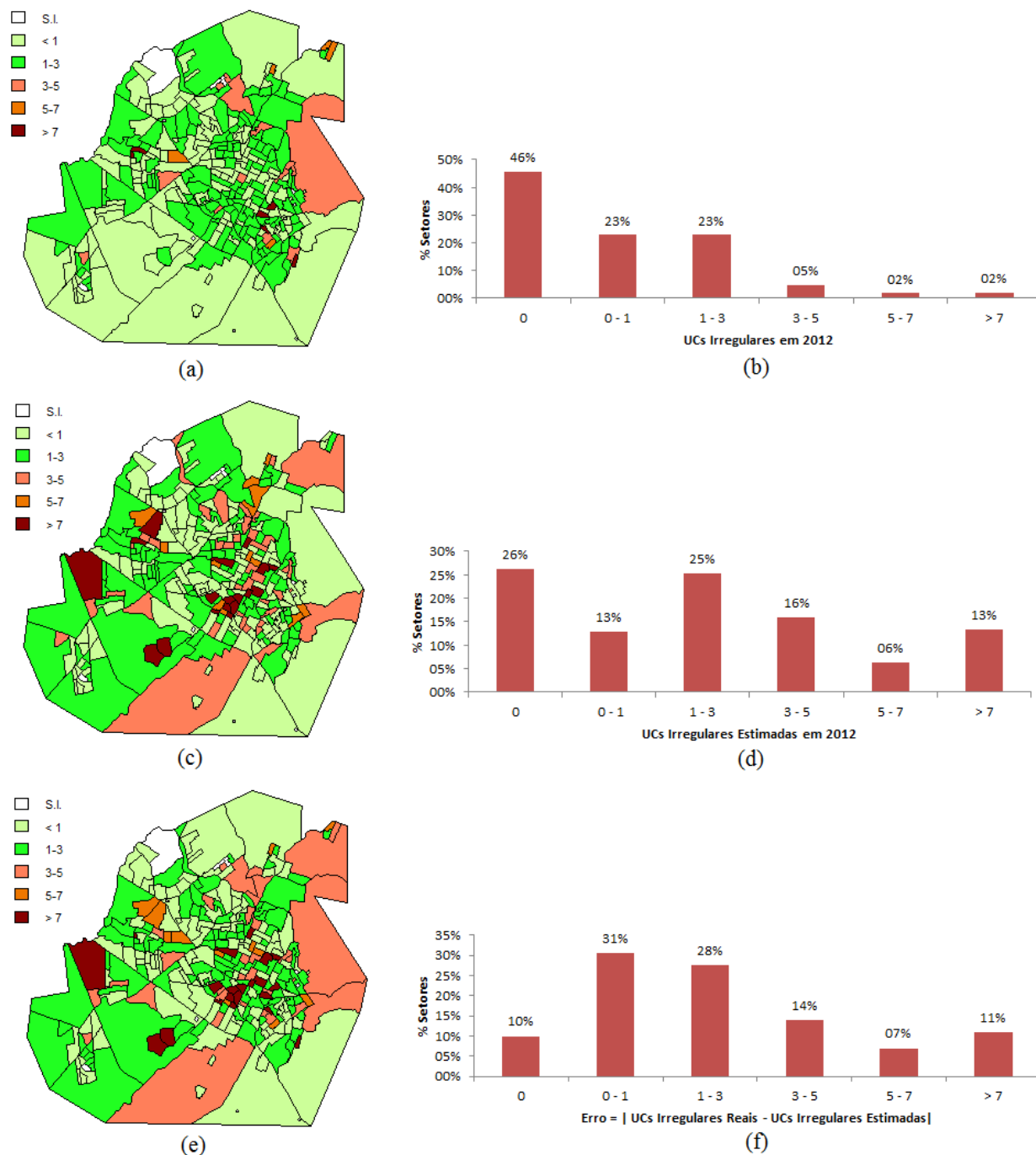
Na Figura 30 apresenta-se o principal resultado deste terceiro estudo de caso. Estimam-se as UCs irregulares em 2012 via modelo de regressão espacial por áreas GWR em conjunto com as cadeias de Markov – Figuras 30 (c) e (d).

As UCs irregulares previstas ou estimadas nas Figuras 30 (c) e (d) são comparadas com as UCs irregulares obtidas em campo em 2012 – Figuras 30 (a) e (b). Nas Figuras 30 (e) e (f) apresenta-se o erro da comparação, isto é, o valor absoluto da diferença entre as UCs irregulares reais e as UCs previstas.

Observa-se que o número de UCs irregulares previstas foi superior ao número de UCs reais flagradas em irregularidades em 2012 na maior parte das subáreas da Figura 30. Esse resultado é coerente e sugere que provavelmente existem mais UCs irregulares no município do que aquelas que foram encontradas em 2012 pelas equipes de inspeção em campo. Não é viável financeiramente inspecionar todas as UCs do município. Portanto, certamente, sempre existirão UCs irregulares que permanecerão ocultas, como se fossem regulares. Sabe-se também que o número de inspeções na maior parte das concessionárias brasileiras está muito aquém do desejável (COMETTI; VAREJÃO, 2005). Adicionalmente, conforme a Ta-

bela 9, o número de inspeções realizadas em 2012 foi inferior em comparação com os anos de 2009 a 2011; logo, a tendência é de que se encontrem menos UCs irregulares em 2012.

Figura 30 – Mapas de probabilidade de perdas no futuro via GWR e via cadeias de Markov. (a), (b) Distribuição espacial e histograma de frequência relativa; ambos, com respeito às UCs flagradas em irregularidades em 2012, respectivamente. (c), (d) Distribuição espacial e histograma de frequência relativa; ambos, com respeito às UCs irregulares estimadas ou previstas em 2012, respectivamente. (e), (f) Distribuição espacial e histograma de frequência relativa; ambos, com respeito ao erro absoluto entre as UCs irregulares reais e as UCs irregulares estimadas em 2012, respectivamente.



Fonte: Elaboração do próprio autor.

Os mapas de probabilidade de perdas obtidos neste terceiro estudo de caso – Figura 30 – são uma ferramenta importante para auxiliar no combate e prevenção às PNTs. Por meio deles, visualizam-se, quais são as subáreas com maior (e com menor) probabilidade de perdas. Logo, de posse desses mapas, pode-se, por exemplo, otimizar a distribuição das inspeções na cidade; mais inspeções em subáreas críticas e menos inspeções em subáreas não críticas.

5.8 CONSIDERAÇÕES FINAIS

Neste capítulo estimou-se a variação espaço-temporal das PNTs via mapas de probabilidade de perdas. Foram efetuados três estudos. Na primeira parte do capítulo realizou-se uma explanação a respeito dos dados disponíveis e das características do município sob análise. Na sequência, tem-se a análise exploratória a partir da qual se visualiza, por meio de gráficos em cores, o padrão espacial das PNTs. Constata-se a viabilidade do uso da análise espacial de dados geográficos para o estudo das PNTs.

Na segunda parte, no primeiro estudo de caso, apresenta-se uma análise para estimar um mapa de probabilidades de perdas. Utilizam-se as localizações pontuais dos *casos* e *controles* e as características socioeconômicas das subáreas do município. As subáreas das regiões central e oeste do município foram consideradas mais vulneráveis às PNTs, pois possuem maior probabilidade de perdas. As variáveis estatisticamente significantes foram: *Média de Moradores*, *%UCs Alugadas*, *NTRAFO* e *PERTRAFO*.

Na parte final utilizou-se uma combinação entre o modelo de regressão aditivo generalizado (GAM) e modelo geograficamente ponderado (GWR); ambos, combinados individualmente com as cadeias de Markov para produção de mapas de probabilidades de perdas no presente e no futuro. Nesse estudo foram consideradas variáveis socioeconômicas e da rede de distribuição de energia elétrica.

Na Tabela 20 há um comparativo entre as regressões espaciais GAM e GWR empregadas neste trabalho para construção do mapa de probabilidades de perdas. O GAM é um modelo de regressão espacial adequado para modelar as perdas como pontos no espaço geográfico – padrões de pontos. Uma desvantagem em utilizá-lo diz respeito a não disponibilização das variáveis socioeconômicas do IBGE por pontos (por uma questão de confidencialidade); elas estão agregadas somente por áreas. Logo, todos os pontos no interior de uma subárea têm o mesmo valor de *Renda Média*, *%UCs Alugadas*, *%UCs com Água*, etc. Em contrapartida, o modelo de regressão espacial local por áreas GWR é mais adequado para incorporar as

variáveis socioeconômicas do IBGE dispostas por subáreas, pois ele é um modelo concebido para lidar com dados agregados por áreas ou polígonos no espaço geográfico. Portanto, o GAM apresenta limitações aos dados censitários que não são apresentadas pelo GWR.

Tabela 20 – Comparativo entre os modelos de regressões espaciais GAM e GWR.

<i>Atributos</i>	<i>Regressões Espaciais</i>	
	<i>GAM</i>	<i>GWR</i>
<i>Técnica</i>	Por Pontos	Por Áreas
<i>Vizinhança</i>	Convencional baseada na distância Euclidiana	Semântica baseada na semelhança dos atributos entre as áreas
<i>Limitação aos Dados Censitários (disponíveis por subáreas)</i>	Todos os pontos no interior de cada subárea têm igual valor	Não existe essa limitação

Fonte: Elaboração do próprio autor.

Com respeito à modelagem da estrutura de vizinhos, no modelo com o GAM e com o GWR utilizaram-se, respectivamente, a vizinhança convencional (baseada na distância Euclidiana entre subáreas) e a vizinhança semântica (baseada na semelhança entre os atributos das subáreas). Em suma, a regressão espacial GWR com a estrutura de *vizinhança semântica* mostrou-se mais apropriada ao problema das PNTs e aos dados censitários disponíveis por subáreas do que a modelagem via GAM com a estrutura de vizinhança convencional.

Observa-se, em último, que o desempenho desses modelos de regressões depende fundamentalmente da qualidade dos dados de entrada e do grupo de variáveis dependentes e independentes selecionado para efetuar a regressão espacial (BAILEY; GATRELL, 1995; BIVAND; PEBESMA; GÓMEZ-RUBIO, 2008).

Na Tabela 21 apresenta-se uma lista das principais funções do *software* R[®] utilizadas neste estudo com suas respectivas bibliotecas e descrições.

Tabela 21 – Algumas das funções R[®] utilizadas neste estudo.

<i>Funções R[®]</i>	<i>Biblioteca</i>	<i>Descrição</i>
<i>mse2d()</i>	<i>splancks</i>	Estimação da largura de banda do estimador de <i>kernel</i> não paramétrico
<i>density()</i>	<i>spatstat</i>	Estimação de <i>kernel</i> não paramétrico
<i>kernels()</i>	<i>spatstat</i>	Estimação de <i>kernel</i> não paramétrico
<i>envelope()</i>	<i>spatstat</i>	Simulação das funções <i>G</i> , <i>F</i> , <i>K</i> , empíricas e teóricas
<i>lambdahat()</i>	<i>spatialkernel</i>	Estimação de <i>kernel</i> não paramétrico conforme Kelsall e Diggle (1995)
<i>spkernel2d()</i>	<i>splancks</i>	Estimação de <i>kernel</i> não paramétrico
<i>gam()</i>	<i>mgcv</i>	Simula um modelo aditivo generalizado
<i>gwss()</i>	<i>GWmodel</i>	Resumo estatístico GW

Fonte: Elaboração do próprio autor.

6 CONCLUSÕES

Neste trabalho incorporou-se, de maneira inédita, o espaço geográfico ao problema das PNTs. O lugar onde elas ocorrem foi caracterizado a partir dos atributos socioeconômicos do censo demográfico e da rede de distribuição de energia elétrica. Determinou-se a evolução das perdas no espaço geográfico do município em estudo ao longo do tempo via mapas de probabilidade de perdas. Esses mapas indicam, em termos probabilísticos, quais são as subáreas vulneráveis às PNTs no presente e em um futuro próximo (um ano à frente).

O presente trabalho foi composto por três estudos principais. No primeiro, constatou-se que as UCs com PNTs estão dispostas espacialmente em aglomerados em algumas regiões do município em análise. No segundo estudo, as perdas foram representadas como pontos no espaço geográfico. Foi desenvolvido um modelo de regressão espacial GAM que levou em consideração as variáveis independentes (ou explicativas) socioeconômicas e da rede elétrica que caracterizam as regiões do município onde estão as perdas. A partir dos resultados dessa metodologia, foram estimadas as subáreas do município com maior risco às perdas possibilitando o planejamento de ações mais específicas de combate e de prevenção às mesmas. No terceiro estudo, apresentou-se uma metodologia para estimar a dinâmica das PNTs. Estimou-se a vulnerabilidade das subáreas às perdas no presente via GAM (dados agregados por pontos) e via GWR (dados agregados por áreas). Essas regressões espaciais foram combinadas individualmente com as cadeias de Markov a fim de prever as subáreas vulneráveis às perdas no futuro. O comportamento regional dinâmico das PNTs foi caracterizado pelo *estado de perda* em cada subárea possibilitando a construção de um mapa de probabilidades de perdas para o município em estudo.

Neste estudo constatou-se também que as variáveis socioeconômicas e da rede elétrica não são as únicas explicações para as perdas; no entanto, elas influem diretamente na disposição espacial das mesmas. Como cada região possui atributos específicos; logo, a incorporação do espaço à análise foi uma alternativa viável no sentido de esclarecer os fatores espaciais (ou locais) relevantes, ou seja, aqueles que possuem significância estatística.

O intuito deste trabalho foi obter a resposta para três questões fundamentais ao problema das PNTs. A saber:

- ❖ Onde estão as perdas?
- ❖ Por que as perdas estão concentradas em determinadas subáreas do município?
- ❖ Onde estarão localizadas as perdas no futuro?

Em geral, os trabalhos avaliados em perdas respondem somente a primeira pergunta. Eles são focados na detecção pontual das mesmas no presente. Neste trabalho ampliou-se a discussão. Ao longo do presente trabalho, as três questões supracitadas foram respondidas de maneira a complementar os trabalhos prévios. Utilizaram-se as regressões espaciais GAM (análise espacial de padrões de pontos) e GWR (análise espacial de dados agregados por áreas) em conjunto com as cadeias de Markov para produção de mapas de probabilidade de perdas. Por meio desses mapas probabilísticos, estimaram-se as subáreas vulneráveis às perdas no presente e no futuro. Dessa forma, com a localização das perdas por subáreas no presente e no futuro, a primeira e a última questão supracitadas foram respondidas. Ademais, via testes de significância estatística, estimaram-se quais os atributos locais (referentes ao local onde se encontram as perdas) que mais influem na disposição espacial das mesmas – respondendo a segunda questão supracitada.

Em último, observa-se que este estudo foi executado a partir de dados provenientes de um município do interior paulista por restrição de dados disponíveis. No entanto, as metodologias desenvolvidas podem ser executadas em programas computacionais de análise estatística ou SIGs e aplicadas a outras localidades a partir da disponibilidade dos dados do lugar que se deseja estudar. A saber: dados cadastrais dos clientes, histórico de inspeções e dados públicos do censo demográfico.

6.1 PRINCIPAIS CONTRIBUIÇÕES DO TRABALHO

Este estudo representa um novo viés para a área de perdas não técnicas com contribuições relevantes aos trabalhos prévios. Em primeiro, destaca-se a incorporação do espaço geográfico ao problema das PNTs. Os trabalhos em perdas comumente não avaliam as características do lugar onde ocorrem as perdas; em geral, preocupam-se somente em localizar pontualmente onde estão as UCs irregulares. Os poucos trabalhos encontrados que consideram a dimensão espacial das perdas, em geral, fazem-no de forma preliminar ou rudimentar. Neste trabalho buscou-se complementar os trabalhos existentes com a adição da dimensão espacial ao problema das PNTs via técnicas de análise espacial de dados geográficos. Essas são técnicas consolidadas pertencentes à estatística espacial (ou estatística bidimensional) e são utilizadas comumente nas áreas de epidemiologia, agronomia, geologia, etc.

Outra contribuição importante é o uso das regressões espaciais GAM e GWR. As variáveis socioeconômicas do censo demográfico e as variáveis da rede elétrica são as variáveis independentes ou explicativas dessas regressões espaciais. Essas variáveis foram empre-

gadas para auxiliar na compreensão do porquê o padrão espacial das perdas encontra-se em aglomerado no espaço geográfico do município. Ademais, por meio de testes de significância estatística, estimaram-se quais dessas variáveis exercem influência na disposição espacial das PNTs.

Nas regressões espaciais supracitadas, GAM e GWR, utilizaram-se algumas variáveis independentes não usuais. Os poucos trabalhos em perdas que incluem a dimensão espacial, utilizam comumente variáveis relacionadas à renda. Dessa forma, estabelecem a relação que aparece em muitos estudos (e em muitas cidades) que há entre as PNTs e a pobreza. Neste estudo, no entanto, empregaram-se novas variáveis independentes para as regressões espaciais. A saber: a densidade de carga (em kVA/km²) e o percentual de UCs alugadas por subáreas.

Destaca-se também o resultado da metodologia que são os mapas de probabilidades de perdas. As regressões espaciais supracitadas são combinadas com as cadeias de Markov de tempo discreto para estimar a evolução das PNTs no espaço geográfico do município ao longo do tempo. Esses mapas probabilísticos indicam as subáreas do município com maior (e com menor) vulnerabilidade às perdas. Tais mapas são uma ferramenta gráfica, de fácil interpretação e que podem ser empregados para antecipar uma série de ações de prevenção e combate às PNTs.

A introdução da “vizinhança semântica” é outra contribuição relevante e inédita deste trabalho. As técnicas de análise espacial de dados agregados utilizam uma estrutura de vizinhança baseada na distância Euclidiana entre as subáreas – áreas vizinhas são aquelas mais próximas entre si. Essa estratégia é adequada para as áreas de pesquisa em que a análise espacial é comumente empregada tais como: epidemiologia, geologia, etc. No entanto, essa estratégia de vizinhança não é a mais recomendável para o problema das PNTs. Devido a isso, criou-se a “vizinhança semântica”. Nessa nova estrutura, a vizinhança é definida não mais via distância Euclidiana, mas por meio da semelhança entre os atributos das subáreas – áreas vizinhas são aquelas cujos atributos têm maior semelhança entre si, independente da distância entre as mesmas. A “vizinhança semântica” expande a noção de vizinhança tradicional e possibilita que subáreas distantes exerçam maior influência do que subáreas próximas e vice-versa. Ademais, esse novo padrão de vizinhança possibilita o estudo de subáreas desconhecidas, que não receberam inspeções, via comparação dos atributos dessas subáreas com os atributos de outras que foram inspecionadas.

Os estudos apresentados nesta tese são compactos e de fácil implementação em programas computacionais de análise estatística ou SIGs. As metodologias propostas reque-

rem poucos dados de entrada os quais estão disponíveis no departamento de perdas das empresas distribuidoras de energia. A saber: dados cadastrais dos clientes, dados de inspeções, base de transformadores georreferenciados¹¹ e dados públicos do censo demográfico.

6.2 TRABALHOS FUTUROS

Seguem algumas sugestões para trabalhos futuros:

- ❖ incrementar as regressões espaciais GAM e o GWR com outras variáveis socioeconômicas não implementadas e investigar essas variáveis com outros métodos exploratórios;
- ❖ utilizar modelos da área de sistemas inteligentes para prever a dinâmica das perdas;
- ❖ realizar a previsão global do montante de perdas e depois distribuí-las para as localidades.

¹¹ As UCs da base de dados disponível para este trabalho, ao contrário dos transformadores, não estão georreferenciadas. Por isso, as mesmas foram associadas aos transformadores aos quais se conectam a fim de viabilizar o uso das técnicas de análise espacial de dados geográficos. Se as UCs da base de dados estivessem georreferenciadas, a base de transformadores georreferenciados da rede elétrica seria dispensável.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL. **Metodologia de tratamento regulatório para perdas não-técnicas de energia elétrica**. Brasília, DF: SRE/ANEEL, 2010. (Nota Técnica, n. 271/2010-SER/ANEEL). Disponível em: <http://www.aneel.gov.br/aplicacoes/audiencia/arquivo/2010/040/documento/nota_tecnica_n%C2%BA_271_perdas_ao_tecnicas.pdf>. Acesso em: 5 ago. 2014.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL. **Metodologia de tratamento regulatório para perdas não-técnicas de energia elétrica**. Brasília, DF: SRE/ANEEL, 2011. (Nota Técnica, n. 0031/2011-SER/ANEEL). Disponível em: <http://www.aneel.gov.br/aplicacoes/audiencia/arquivo/2010/040/documento/nota_tecnica_n%C2%B0_0312011-sreaneel-perdas_ao_tecnicas-2_fase.pdf>. Acesso em: 2 jul. 2014.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL. **Procedimentos de distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST: módulo 7: cálculo de perdas na distribuição**. Brasília, DF: ANEEL, 2013. Disponível em: <http://www.aneel.gov.br/arquivos/PDF/Modulo7_Revisao_3-Final.pdf>. Acesso em: 15 nov. 2014.

ANGELOS, E. W. S.; SAAVEDRA, O. R.; CORTÉS, O. A. C.; SOUZA, A. N. Detection and identification of abnormalities in customer consumptions in power distribution systems. **IEEE Trans. on Power Delivery**, Piscataway, v. 26, n. 4, p. 2436-2442, 2011.

BAILEY, C.; GATRELL, A. C. **Interactive spatial data analysis**. Essex: Longman Scientific & Technical, 1995. 411 p.

BASTOS, P. R. F. M. **Diagnóstico de perdas comerciais de energia elétrica na distribuição usando rede Bayesiana**. 2011. 171f. Tese (Doutorado em Engenharia Elétrica) – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2011.

BERMAN, M.; DIGGLE, P. J. Estimating Weighted Integrals of the Second-Order Intensity of a Spatial Point Process. **Journal of the Royal Statistical Society**, London, v. 51, n. 1, p. 81-92, 1989.

BIVAND, R. S.; PEBESMA, J.; GÓMEZ-RUBIO, V. **Applied spatial data analysis with R**. New York: Springer, 2008. 378 p.

BRUNSDON, C.; FOTHERINGHAM, A. S.; CHARLTON, M. E. Geographically Weighted Regression: a method for exploring spatial nonstationarity. **Geographical Analysis**, v. 28, n. 4, p. 281-298, 1996.

BURROUGH, P. **Dynamic modelling and geocomputation: a primer**. London: John Wiley & Sons, 1998.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 6. ed. São Paulo: Saraiva, 2010. 540 p.

- CABRAL, J. E.; PINTO, J. O.; GONTIJO, E. M.; FILHO, J. R. Fraud detection in electrical energy consumers using rough sets. In: IEEE International Conference on Systems, Man and Cybernetics, 2004, The Hague. **Proceedings...** Piscataway: IEEE, 2004. p. 3625-3629.
- CAMARGO, C. C. B. **Confiabilidade aplicada a sistemas de potência elétrica**. Rio de Janeiro: LTC, 1981. 206 p.
- CARREÑO, E. M. **Previsão espacial de demanda em sistemas de distribuição com uma base reduzida de dados**. 2008. 106 f. Tese (Doutorado em Engenharia Elétrica) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2008.
- COMETTI, E. S.; VAREJÃO, F. M. **Melhoramentos da identificação de perdas comerciais através da análise computacional inteligente do perfil de consumo e dos dados cadastrais dos consumidores**. Vitória: UFES/NINFA, 2005. 11 p. Disponível em: <http://ninfa.inf.ufes.br/public_files/mip/ArtigoRelatorioFinal.pdf>. Acesso em: 15 fev. 2011.
- COUCLELIS, H. **Space, time, geography. geographical information systems**. New York: John Wiley, 1997.
- CRESSIE, N. **Statistics for spatial data**. Chichester: John Wiley, 1991.
- DANTAS, P. R. P. **Avaliação de perdas de energia elétrica não-técnicas: metodologia aplicada no município de Salvador/BA**. 2006. 98f. Dissertação (Mestrado em Engenharia Elétrica) – Departamento de Engenharia e Arquitetura, Universidade Salvador, Salvador, 2006.
- DIGGLE, P. J. **Statistical analysis of spatial point patterns**. 2. ed. London: Academic Press, 1983.
- DIGGLE, P. J.; RIBEIRO, P. J. **Model-based geostatistics**. London: Springer, 2007.
- DIGGLE, P. J.; ROWLINGSON, B. A conditional approach to point process modelling of elevated risk. **Journal of the Royal Statistical society**. Oxford, v. 157, n. 3, p. 433-440, 1994.
- DRUCK, S.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A. V. M. (Ed.). **Análise espacial de dados geográficos**. Brasília, DF: Ed. EMBRAPA, 2004.
- FARIA, L. T. **Sistema inteligente híbrido intercomunicativo para detecção de perdas comerciais**. 2012. 112 f. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2012.
- FARIA, L. T.; MELO, J. D.; PADILHA-FELTRIN, A. Análise espacial de pontos para mapeamento de perdas comerciais. In: SIMPÓSIO BRASILEIRO DE SISTEMAS ELÉTRICOS, 5., 2014, Foz do Iguaçu. **Anais...** Foz do Iguaçu: SBSE, 2014a. p. 1-8.
- FARIA, L. T.; MELO, J. D.; PADILHA-FELTRIN, A. Estimção da Distribuição Espacial das Perdas Comerciais através de um Modelo Aditivo Generalizado. In: CONGRESSO BRASILEIRO DE AUTOMÁTICA, 20., 2014, Belo Horizonte. **Anais...** Belo Horizonte: CBA, 2014b. p. 2873-2880.

FARIA, L. T.; MELO, J. D.; PADILHA-FELTRIN, A. Spatial-temporal estimation for non-technical losses. **IEEE Trans. Power Delivery**, Piscataway, v. 31, n. 1, p. 362-369, 2015.

FARIA, L. T.; PADILHA-FELTRIN, A.; MINUSSI, C. R. Sistema Inteligente Híbrido Inter-comunicativo para Detecção de Perdas Comerciais em Sistemas de Distribuição de Energia Elétrica. In: SIMPÓSIO BRASILEIRO DE SISTEMAS ELÉTRICOS, 4., 2012. **Anais...** Goiânia: SBSE, 2012. p. 1-8.

FERREIRA, H. M. **Uso de ferramentas de aprendizado de máquina para prospecção de perdas comerciais em distribuição de energia elétrica**. 2008. 89f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, São Paulo, 2008.

FERREIRA, M. C. **Iniciação à análise geoespacial: teoria, técnicas e exemplos para geoprocessamento**. São Paulo: Ed. Unesp, 2014. 343 p.

FINE, P.; VICTORA, C. G.; ROTHMAN, K. J.; MOORE, P. S.; CHANG, Y.; CURTIS, V.; HEYMANN, D. L.; SLUTKIN, G.; MAY, R. M.; PATEL, V.; ROBERTS, I.; WORTLEY, R.; TORGERSON, C.; DEATON, A. John Snow's legacy: epidemiology without borders. **Lancet**, London, v. 13, n. 381, p. 1302-1311, 2013.

FOTHERINGHAM; BRUNSDON, C.; CHARLTON, M. **Geographically weighted regression: the analysis of spatially varying relationships**. Chichester: John Wiley & Sons, 2002.

FRANCISCO, E. R.; FAGUNDES, E. B.; PONCHIO, M. C.; ZAMBALDI, F. Development of an indicator of propensity to energy commercial losses using geospatial statistical techniques and sócio-economic data: the case of AES Eletropaulo. **Revista de Administração Machenzie**, v. 11, n. 4, art. 93, p. 178-197, 2010.

GUERREIRO, J. I.; LEÓN, C.; BISCARRI, F. Increasing the efficiency in non-technical losses detection in utility companies. In: IEEE MELECON, 2010, Valletta. **Proceedings...** Piscataway: IEEE, 2010. p. 136-141.

GOLLINI, I.; BINBIM L.; CHARLTON M.; BRUNSDON, C.; HARRIS, P. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. **Journal of Statistical Software**, v. 63, p. 1-50, 2015.

HASTINGS, N. A. J. **Dynamic programming: with management applications**. London: The Butterworths Group, 1973.

HAYKIN, S. **Neural networks: a comprehensive foundation**. 2. ed. New Jersey: Prentice Hall, 1999. 823 p.

HEPPENSTALL, A. J., CROOKS, A. T., SEE, L. M., BATTY, M. (Eds.). **Agent-based models of geographical systems**. New York: Springer, 2012. 746 p.

HUANG, S. C.; LO, Y. L.; LU, C. N. Non-Technical Loss Detection Using State Estimation and Analysis of Variance. **IEEE Trans. on Power Systems**, Piscataway, v. 28, n. 3, p. 2959-2966, 2013.

HUFF, D. **How to lie with statistics**. London: Penguin Book, 1973. 129 p.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Censo demográfico 2010**: Agregado por setores censitários. 2. edição. Brasília, DF, 2010. Disponível em: <<http://www.ibge.gov.br>>. Acesso em: 05 nov. 2013.

JIANG, R. J. Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In: IEEE/PES TRANSMISSION AND DISTRIBUTION CONFERENCE AND EXHIBITION, 2002, Dallas. **Proceedings...** Piscataway: IEEE, 2002. p. 6-10.

KELSALL, J. E.; DIGGLE, P. J. Kernel estimation of relative risk. **Bernoulli**, The Hague, v. 1, n. 1-2, p. 3-16, 1995a.

KELSALL, J. E.; DIGGLE, P. J. Non-parametric estimation of spatial variation in relative risk. **Statistics in Medicine**, Chichester, v. 14, n. 21-22, p. 2335-2342, 1995b.

KELSALL, J. E.; DIGGLE, P. J. Spatial variation in risk of disease: a nonparametric binary regression approach. **Applied Statistics**, Abingdon, v. 47, n. 4, p. 559-573, 1998.

KIFFER, C. R. V.; CAMARGO, E. C. G.; SHIMAKURA, S. E.; RIBEIRO, P. J.; BAILEY, T. C.; PIGNATARI, A. C. C.; MONTEIRO, A. M. V. 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. **International Journal of Health Geographics**, London, v. 10, n. 1, p. 17, 2011.

KOHONEN, T. **Self-Organizing Maps**. 3. ed. Berlin: Springer, 2001, 502 p.

LAMBIN, E. F. **Modeling deforestation processes**: a review, trees series b: research report., Luxemburgo: European Commission, 1994.

LARSON, R.; FARBER, B. **Estatística Aplicada**. 4. ed. São Paulo: Pearson Prentice Hall, 2010. 640 p.

LEÓN, C.; BISCARRI, F.; MONEDERO, I.; GUERRERO, J. I.; BISCARRI, J.; MILLÁN, R. Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies. **IEEE Trans. Power Systems**, Piscataway, v. 26, n. 4, p. 1798-1807, 2011.

LIMA, T. F. M.; FARIA, S. D.; FILHO, B. S. S.; CARNEIRO, T. G. S. Modelagem de sistemas baseada em agentes: alguns conceitos e ferramentas. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal. **Anais...** Natal: [s. n.], 2009. p. 5279-5286.

LU, B.; HARRIS, P.; CHARLTON, M.; BRUNSDON, C. The GWmodel R package: further topics for exploring Spatial Heterogeneity using Geographically Weighted Models. **Geospatial Information Science**, Singapore, v. 17, n. 2, p. 85-101, 2014.

MAGNABOSCO, A. **CPFL investirá R\$ 700 milhões em medição inteligente**. São Paulo: Exame.com, 2015. Disponível em: <<http://exame.abril.com.br/negocios/noticias/cpfl-investira-r-700-milhoes-em-medicao-inteligente>>. Acesso em: 30 mar. 2015.

MELO, J. D. **Aplicações de sistemas multiagentes na previsão espacial de demanda elétrica em sistemas de distribuição**. 2010. 152 f. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2010.

MELO, J. D.; CARREÑO, E. M.; PADILHA-FELTRIN, A. Análise espacial de eventos pontuais na alocação de transformadores da rede elétrica secundária de distribuição. In: CONGRESSO DE AUTOMÁTICA, 2012, Campina Grande. **Anais...** Campina Grande: CBA, 2012. p. 1-8.

MIOTO, R. **Quando Nicolas Cage faz filmes, mortes são evitadas**: veja outras correlações estatísticas que mentem. São Paulo: Folha de São Paulo, 2015. Disponível em: <<http://www1.folha.uol.com.br/ciencia/2015/05/1626815-quando-nicolas-cage-faz-filmes-mortes-sao-evitadas-veja-outras-correlacoes-estatisticas-que-mentem.shtml>>. Acesso em: 11 maio 2015.

NAGI, J., YAP, K. S., TIONG, S. K., AHMED, S. K., MOHAMAD, M. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. **IEEE Trans. Power Delivery**, Piscataway, v. 25, n. 2, p. 1162-1171, 2010.

NAGI, J., YAP, K. S., TIONG, S. K., AHMED, S. K., NAGI, F. Improving SVM-Based Non-technical loss detection in power utility using the fuzzy inference system. **IEEE Trans. Power Delivery**, Piscataway, v. 26, n. 2, p. 1284-1285, 2011.

NIZAR, A. H., DONG, Z. Y., WANG, Y. Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method. **IEEE Trans. Power Systems**, Piscataway, v. 23, n. 3, p. 946-955, 2008.

NIZAR, A. H., DONG, Z. Y., ZHANG, P. Detection rules for non technical losses analysis in power utilities. In: IEEE POWER AND ENERGY SOCIETY GENERAL MEETING, 21., 2008, Pittsburgh. **Proceedings...** New York: IEEE, 2008. p. 1-8.

OLIVEIRA, J. **Maconha consome um terço da energia furtada na Inglaterra**. São Paulo: Estadão, 2013. Disponível em: <<http://economia.estadao.com.br/noticias/geral,maconha-consome-um-terco-da-energia-furtada-na-inglaterra,158547e>>. Acesso em: 05 jul. 2013.

OLIVEIRA, M. E. **Avaliação de metodologias de cálculo de perdas técnicas em sistemas de distribuição de energia elétrica**. 2009. 137f. Tese (Doutorado em Engenharia Elétrica) - Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2009.

PENIN, C. A. S. **Combate, prevenção e otimização das perdas comerciais de energia elétrica**. 2008. 227f. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.

PERIM, G. T.; DIAS, H. B. P.; COMETTI, E. S. Uma abordagem baseada em conhecimento para identificação de perdas elétricas. In: Congresso Brasileiro de Eficiência Energética, 2007. Vitória. **Anais...** Vitória: Associação Brasileira de Eficiência Energética, 2007. 6 p.

PETERNELLI, L. A.; MELLO, M. P. **Conhecendo o R: uma visão estatística**. Viçosa, MG: Ed. UFV, 2011. 185 p.

PORRAS, J. A.; RIVERA, H. O.; GIRALDO, F. D.; CORREA, B. S. A. Identification of Non-Technical Electricity Losses in Power Distribution Systems by Applying Techniques of Information Analysis and Visualization. **IEEE Latin America Trans.**, Piscataway, v. 13, p. 659-664, 2015.

R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2013. Disponível em: <<http://www.R-project.org/>>. Acesso em: 02 de abr. 2013.

RAMOS, C. C. O. **Caracterização de perdas comerciais em sistemas de energia através de técnicas inteligentes**. 2014. 144f. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2014.

RAMOS, C. C. O., SOUSA, A. N., PAPA, J. P., FALCÃO, A. X. A New approach for non-technical losses detection based on optimum-path forest. **IEEE Trans. Power Systems**, Piscataway, v. 26, n. 1, p. 181-189, 2011.

RAMOS, C. C. O., SOUZA, A. N., FALCÃO, A. X., PAPA, J. P. New insights on nontechnical losses characterization through evolutionary-based feature selection. **IEEE Trans. on Power Delivery**, Piscataway, v. 27, n. 1, p. 140-146, 2012.

RANSFORD, C. L., KANE, C., SLUTKIN, G. CeaseFire: a disease control approach to reduce violence and change behavior. In: WALTERMAUER, A. T. W. (Ed). **Epidemiological criminology**. Lodon: Routeledge, 2012. p. 232-242.

RAUBER, T. W.; DRAGO, I.; VAREJÃO, F. M.; QUEIROGA, R. M. Extração e seleção de características na identificação de perdas comerciais na distribuição de energia elétrica. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25., São Leopoldo. **Anais...** São Leopoldo: Sociedade Brasileira de Computação, 2005. p. 1-4.

RIPLEY, B. D. The second order analysis of stationary point processes. **Journal of Applied Probability**, Sheffield, v. 13, n. 2, p. 255-266, 1976.

RODRIGUES, D; RAMOS, C. C. O.; PAPA, J. P. Black hole algorithm for non-technical losses characterization. IEEE Latin American Symposium, 6., 2015, Montevideo. **Proceedings...** Montevideo: [s. n.], 2015. p. 1-4.

RODRIGUES, T. C. V. **Regressão binomial negativa geograficamente ponderada: modelando superdispersão espacial**. 2012. 120f. Dissertação (Mestrado em Estatística) – Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2012.

SCHABENBERGER, O.; GOTWAY, C.A. **Statistical methods for spatial data analysis**. London: Chapman & Hall, 2004.

SHIMAKURA, S. E., CARVALHO, M.S., AERTS, D. R. G. C., FLORES, R. Distribuição espacial do risco: modelagem de mortalidade infantil em Porto Alegre, Rio Grande do Sul, Brasil. **Caderno de Saúde Pública**, v. 17, n. 5, p. 1251-1261, 2001.

- SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais: para engenharias e ciências aplicadas**. São Paulo: Artliber, 2010. 398 p.
- SIMPSON, P. K. **Artificial neural systems: foundations, paradigms, applications and implementations**. New York: Pergamon, 1989.
- SISTEMA ESTADUAL DE ANÁLISE DE DADOS – SEADE. **Perfil municipal de Presidente Prudente**. São Paulo: SEADE, 2010. Disponível em: <<http://www.seade.gov.br/produtos/perfil/perfil.php>>. Acesso em: 08 abr. 2014.
- SMITH, T. B. Electricity theft: a comparative analysis. **Energy Policy**, Oxford, v. 32, p. 2067-2076, 2004.
- SPERANDIO, M.; COELHO, J.; QUEIROZ, H. L. Identificação de agrupamentos de consumidores de energia elétrica através de Mapas Auto-Organizáveis. In: SEMINÁRIO BRASILEIRO SOBRE QUALIDADE DA ENERGIA ELÉTRICA, 2003, Aracaju. **Anais...** Aracaju: Núcleo de Estudos e Pesquisas do Nordeste, 2003.
- TERRAVIEW 4.2.0. **Projeto TerraView**. São José dos Campos: INPE, 2010. Disponível em: <www.dpi.inpe.br/terraview>. Acesso em: 01 nov. 2013.
- TOBLER, W. R. Cellular geography. In: GALE, S.; OLSSSEN, G. (Ed.) **Philosophy in geography**. Dordrecht: Reidel, 1979. p. 379-386.
- TREVIZAN, R. D.; ROSSONI, A.; BRETAS, A. S.; GAZZANA, D. S.; MARTIN, R. P.; BRETAS, N. G.; BETTIOL, A. L.; CARNIATO, A.; PASSOS, L. F. N. Non-technical losses identification using optimum-path forest and state estimation. In: POWERTECH, 2015, Eindhoven. **Proceedings...** New York: IEEE, 2015. p. 1-6.
- VENABLES, W. N. et al. **An Introduction to R: a programming environment for data analysis and graphics**. [S. l.: s. n.], 2012. Disponível em: <<http://cran.r-project.org/doc/manuals/R-intro.pdf>>. Acesso em: 05 out. 2012.
- WOLFRAM, S. **Cellular automata and complexity: collected papers**. Readings: Addison – Wesley, 1994, 608 p.
- WOOD, S. N. **Generalized additive models: an introduction with R**. Boca Raton: CRC, 2006.
- WOOLDRIDGE, M. **Multiagent systems**. Chichester: John Wiley and Sons, 2002. 348 p.
- ZAPATA, C. J. **Análisis probabilístico y simulación**. Pereira: Universidad Tecnológica de Pereira, 2010.

APÊNDICE A – MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN

Os mapas auto-organizáveis de Kohonen ou *self organization maps* (SOM) são um tipo de rede neural artificial com arquitetura reticulada, aprendizado não supervisionado competitivo e que permite a identificação de padrões em vetores de dados multivariados. Tal rede é comumente construída em torno de uma grade unidimensional ou bidimensional de neurônios para capturar similaridades, regularidades e correlações importantes no espaço n -dimensional das n variáveis de entrada; agrupando-as em classes ou em *clusters*. Trata-se de uma ferramenta largamente utilizada para reconhecimento de padrões nas mais diversas áreas do conhecimento, para realizar agrupamentos e outros usos como para facilitar a visualização de relações entre as variáveis de entrada (KOHONEN, 2001; SILVA; SPATTI; FLAUZINO, 2010).

O SOM é um mapeamento que fornece a representação estrutural dos dados de entrada através dos vetores de pesos dos neurônios. Caracteriza-se pela formação de um mapa topográfico dos padrões de entrada onde as localizações dos neurônios mapeiam as características estatísticas (em geral não-lineares e dinâmicas) intrínsecas dos padrões de entrada, conforme disposto na Figura 31 (HAYKIN, 1999). No exemplo, o padrão de entrada $x \in \mathbb{R}^n$ é representado (ou mapeado) no espaço de saída bidimensional pelo vetor de pesos $w^{(nv)}$ do neurônio vencedor $nv(x)$ cuja localização é função de x .

A.1 APRENDIZADO COMPETITIVO

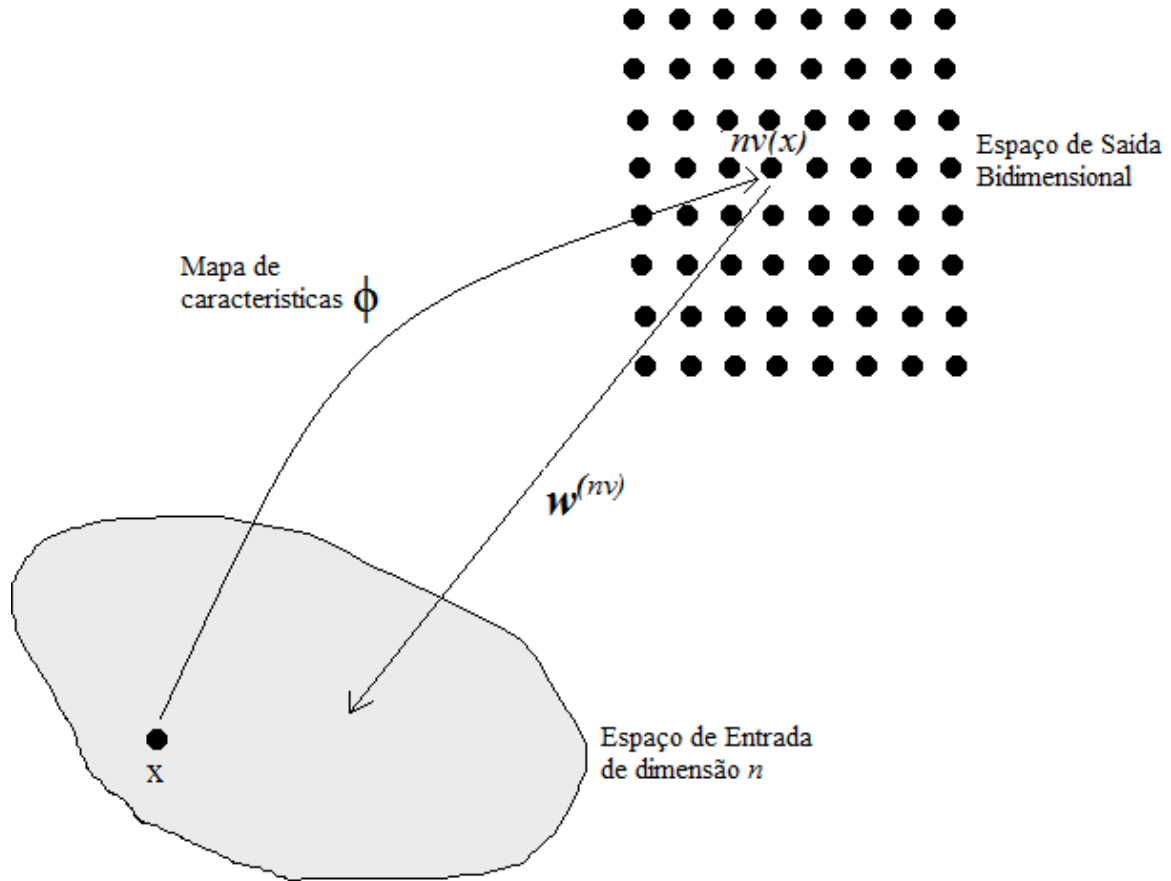
O aprendizado de uma rede neural consiste basicamente da calibração gradativa dos vetores de pesos dos neurônios durante a fase de treinamento à medida que as amostras de entrada são apresentadas sequencialmente à rede. O aprendizado competitivo baseia-se na concorrência entre os neurônios que competem entre si e há um único neurônio vencedor para cada amostra apresentada à rede neural.

A estratégia para determinar qual será o neurônio vencedor – nv – consiste em determinar a proximidade entre o vetor de pesos de cada neurônio com respeito ao vetor de entrada da k -ésima amostra, conforme equação (47):

$$dist_j^{(k)} = \sqrt{\sum_{i=1}^n \left(x_i^{(k)} - w_i^{(j)}\right)^2}, \quad j = \{1, \dots, n_1\}, \quad (47)$$

sendo $dist_j^{(k)}$ a norma Euclidiana da diferença entre a k -ésima amostra de entrada $\mathbf{x}^{(k)}$ e o vetor de pesos $\mathbf{w}^{(j)}$ do j -ésimo neurônio da rede neural. Em (47) há n amostras de entrada e n_1 neurônios na grade neural.

Figura 31 – Mapeamento das amostras de entrada do espaço n -dimensional para o espaço bidimensional da grade neural – $\mathbb{R}^n \Rightarrow \mathbb{R}^2$.



Fonte: (HAYKIN, 1999, p. 477).

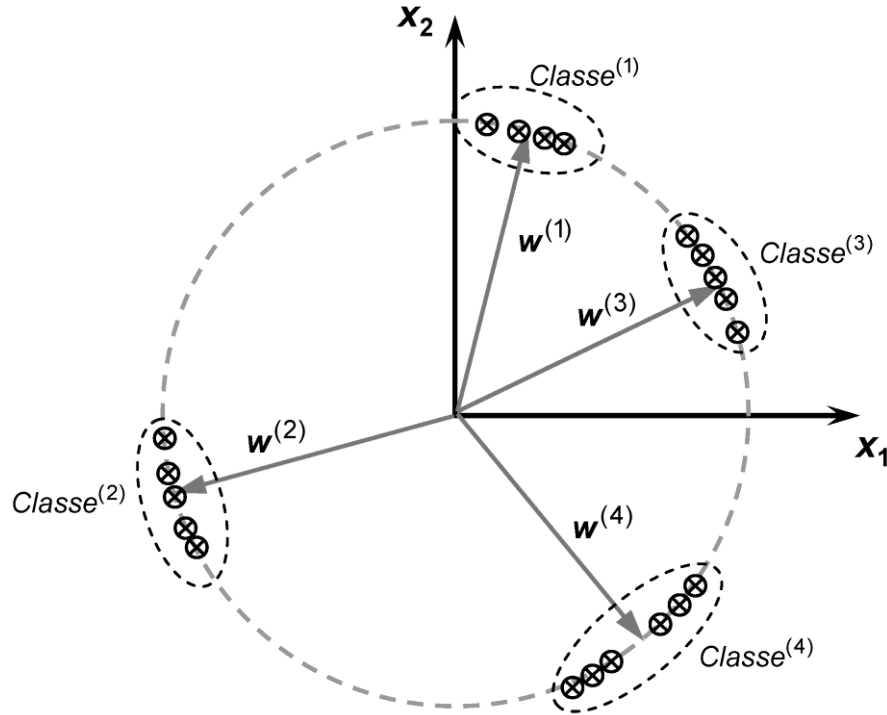
O neurônio j cujo vetor de peso apresenta a menor distância Euclidiana $dist_j^{(k)}$ é nomeado “neurônio vencedor” da competição sendo considerado o neurônio mais ativo em relação à amostra k . O vetor de pesos do neurônio vencedor $\mathbf{w}^{(nv)}$ é ajustado via equação (48) de modo a aproximá-lo ainda mais da amostra previamente apresentada. $\eta(n_e)$ é a taxa de aprendizagem e decresce à medida que o número de épocas (ou iterações) n_e aumenta.

$$\mathbf{w}_{atual}^{(nv)} = \mathbf{w}_{anterior}^{(nv)} + \eta(n_e)(\mathbf{x}^{(k)} - \mathbf{w}_{anterior}^{(nv)}) \quad (48)$$

Na Figura 32 apresenta-se, no espaço bidimensional das variáveis de entrada x_1 e x_2 , a disposição dos quatro vetores de pesos da rede neural associados aos quatro neurônios

do SOM após convergência do algoritmo de treinamento competitivo. Ressalta-se que cada um dos vetores de pesos aponta para o centro de cada um dos quatro aglomerados ou *clusters*. Nesse exemplo, o número de classes é vinculado ao número de neurônios da estrutura neural.

Figura 32 – Disposição geométrica dos vetores de pesos dos neurônios do SOM no espaço bidimensional após treinamento da rede neural.



Fonte: (SILVA; SPATTI; FLAUZINO, 2010, p. 226).

Após a fase de treinamento do SOM, quando uma nova amostra for apresentada à rede, basta verificar qual o neurônio vencedor correspondente que estará associado a uma classe de amostras semelhantes entre si.

A.2 VIZINHANÇA INTERNEURÔNIOS

Os mapas topológicos bidimensionais podem ter um arranjo retangular (como na Figura 33), hexagonal, etc. Dentre os inúmeros critérios de vizinhança interneurônios, um dos mais comuns consiste em adotar um raio R em torno do neurônio cuja vizinhança deseja-se definir. Seja a distância entre os neurônios da grade neural da Figura 33 unitária – $R = 1$. Define-se $\Omega_j^{(R)}$ como sendo o conjunto dos neurônios vizinhos do neurônio j inscritos em um círculo de raio R em torno de j . Em (49) são apresentados alguns exemplos de vizinhanças interneurônios para a grade neural da Figura 33.

Um neurônio j que venceu a competição após a apresentação da amostra $\mathbf{x}^{(k)}$ tem seu vetor de pesos ajustado bem como de seus vizinhos através das expressões (48) e (50), respectivamente. Na equação (50), $\mathbf{w}^{viz}$ corresponde aos pesos dos “neurônios vizinhos” (viz) do neurônio vencedor (nv).

$$\begin{aligned}\Omega_1^{(1)} &= \{2,5\} \\ \Omega_2^{(1)} &= \{1,3,6\} \\ (\dots) \\ \Omega_{16}^{(1)} &= \{12,15\}\end{aligned}\tag{49}$$

Visando maior eficiência no treinamento da rede, a taxa de aprendizagem η_e reduz gradativamente no transcorrer no processo iterativo de treinamento. Em Haykin (1999), apresenta-se a equação (51) onde a taxa de treinamento é função exponencial decrescente do número de épocas n_e . Segundo Haykin (1999), o treinamento subdivide-se em duas fases: a fase de ordenação e de refinamento com τ_2 e $500n_1$ épocas, respectivamente. Tal referência adota 1.000 épocas ou iterações para a fase de ordenação, portanto, $\tau_2 = 1000$.

$$\mathbf{w}_{atual}^{(viz)} = \mathbf{w}_{anterior}^{(viz)} + \eta(n_e)\alpha^{(\Omega)}(n_e)(\mathbf{x}^{(k)} - \mathbf{w}_{anterior}^{(viz)})\tag{50}$$

$$\eta(n_e) = \eta_0 e^{\left(-\frac{n_e}{\tau_2}\right)}\tag{51}$$

$$n_e \cong \tau_2 + 500n_1\tag{52}$$

O lado direito da equação (50) – para ajuste do vetor de pesos dos neurônios vizinhos – é multiplicado pela função Gaussiana $\alpha^{(\Omega)}(n_e)$ definida em (53). Quanto maior a norma Euclidiana da diferença entre os vetores de pesos de j e de seus vizinhos Ω_j , menores serão os ajustes dos neurônios vizinhos. A função $\alpha^{(\Omega)}$ em (53) é função de $\sigma(n_e)$, definido em (54), o qual é função exponencial decrescente do número de épocas. Em (55) define-se o parâmetro τ_1 , que aparece em (54), e que depende do valor de outras duas constantes: τ_2 e σ_0 . Segundo Haykin (1999, p. 475), σ_0 é o raio da grade neural.

$$\alpha^{(\Omega)}(n_e) = e^{-\frac{\|\mathbf{w}^{(venc)} - \mathbf{w}^{(\Omega)}\|^2}{2\sigma(n_e)^2}}\tag{53}$$

$$\sigma(n_e) = \sigma_0 e^{-\frac{n_e}{\tau_1}} \quad (54)$$

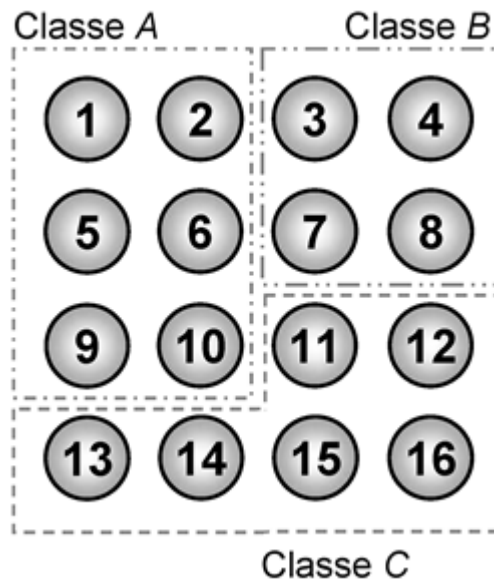
$$\tau_1 = \frac{\tau_2}{\log(\sigma_0)} \quad (55)$$

A.3 MAPAS DE CONTEXTO

Após treinamento do SOM, agrupam-se os neurônios em classes maiores e mais representativas do problema para construção do mapa de contexto. Existem inúmeras estratégias para construção desses mapas como técnicas estatísticas e conhecimento especialista.

Neste estudo adota-se o seguinte procedimento para obtenção do mapa de contexto: sejam os neurônios vizinhos do neurônio 1; $\Omega_1 = \{2,5\}$ na Figura 33. Seleciona-se aquele que possui a menor norma Euclidiana da diferença entre os vetores de pesos do neurônio 1 e de seus respectivos vizinhos. Supõe-se que o neurônio 2 foi selecionado. Neste momento, a *Classe A* = $\{1,2\}$. O processo repete-se para o último neurônio ingressante na classe. Encerra-se o processo quando o neurônio selecionado já está contido na classe em formação. Na Figura 33 apresenta-se o mapa de contexto completo após treinamento do SOM e posterior análise. O mapa foi particionado em três classes *A*, *B* e *C* cujos neurônios estão indicados na Figura 33 por polígonos tracejados.

Figura 33 – Mapa de contexto 4×4 para uma topologia bidimensional após treinamento do SOM.



Fonte: (SILVA; SPATTI; FLAUZINO, 2010, p. 235).

Uma abordagem com mais detalhes a respeito do SOM foge ao escopo do presente trabalho. Ao leitor interessado recomenda-se Haykin (1999); Silva, Spatti e Flauzino (2010) e Simpson (1989).