UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Leandro Brioschi Mineti

DESENVOLVIMENTO E APLICAÇÃO DE MODELOS DE REGIONALIZAÇÃO BAYESIANOS PARA A ANÁLISE DOS ÍNDICES DE EFICIÊNCIA OPERACIONAL E DOS INDICADORES DE DURAÇÃO EQUIVALENTE DE INTERRUPÇÃO POR UNIDADE CONSUMIDORA DAS EMPRESAS BRASILEIRAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Belo Horizonte
2020

Leandro Brioschi Mineti

DESENVOLVIMENTO E APLICAÇÃO DE MODELOS DE REGIONALIZAÇÃO BAYESIANOS PARA A ANÁLISE DOS ÍNDICES DE EFICIÊNCIA OPERACIONAL E DOS INDICADORES DE DURAÇÃO EQUIVALENTE DE INTERRUPÇÃO POR UNIDADE CONSUMIDORA DAS EMPRESAS BRASILEIRAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção da UFMG como parte dos pré-requisitos para obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte
2020

Mineti, Leandro Brioschi.
M664d Desenvolvimento e aplicação de modelos de regionalização bayesianos para análise dos índices de eficiência operacional e dos indicadores de duração equivalente de interrupção por unidade consumidora das empresas Brasileiras de distribuição de energia elétrica [recurso eletrônico] / Leandro Brioschi Mineti. - 2020.
1 recurso online (68 f. : il., color.) : pdf.

Orientador: Marcelo Azevedo Costa.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndices: f.66-68.
Bibliografia: f.61-65.

1. Engenharia de produção - Teses. 2. Aglomeração – Teses. 3. Teoria bayesiana de decisão estatística – Teses. I. Costa, Marcelo Azevedo. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 658.5(043)

Ficha catalográfica elaborada pela Bibliotecária Letícia Alves Vieira - CRB-6/2337
Biblioteca Prof. Mário Werneck - Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

UFMG

FOLHA DE APROVAÇÃO

Desenvolvimento e aplicação de modelos de regionalização bayesianos para a análise dos índices de eficiência operacional e dos indicadores de duração equivalente de interrupção por unidade consumidora das empresas Brasileiras de distribuição de energia elétrica

LEANDRO BRIOSCHI MINETI

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO, como requisito para obtenção do grau de Mestre em ENGENHARIA DE PRODUÇÃO, área de concentração PESQUISA OPERACIONAL E INTERVENÇÃO EM SISTEMAS SOCIOTÉCNICOS, linha de pesquisa Modelagem Estocástica e Simulação.

Aprovada em 01 de fevereiro de 2021, pela banca constituída pelos membros:

Marcelo Azevedo Costa

Prof(a). Marcelo Azevedo Costa - Orientador
ufmg

Prof(a). Marcos Oliveira Prates
DEST/UFMG

Mayrink

Prof(a). Vinicius Diniz Mayrink
DEST/UFMG

Belo Horizonte, 1 de fevereiro de 2021.

Este trabalho é dedicado a minha família, pelo apoio incondicional durante essa jornada. Agradeço especialmente ao meu orientador, Professor Marcelo Azevedo Costa, pelos ensinamentos.

"Pensar es olvidar diferencias, es generalizar, abstraer."
Jorge Luis Borges em *Funes el memorioso*

Resumo

Nos segmentos da economia que operam em condição de monopólio natural, a ausência da competição entre fornecedores pode gerar resultados indesejados para os consumidores, como serviços de baixa qualidade e alto custo. Para evitar esse cenário, é usual a presença do agente regulador, que deve determinar parâmetros de operação e avaliar os resultados das empresas envolvidas. No contexto da distribuição de energia elétrica no Brasil, a agente regulador é a Agência Nacional de Energia Elétrica (ANEEL). Em geral, os parâmetros de operação e índices de desempenho das empresas Brasileiras do setor de energia são afetados por variáveis gerenciáveis e não-gerenciáveis. Essas últimas variáveis relacionadas ao seu ambiente de atuação. Os efeitos ambientais são particularmente importantes no Brasil, dada a amplitude e heterogeneidade do seu território. A primeira parte deste trabalho explora o efeito da heterogeneidade das condições ambientais e seu impacto nos seus índices de eficiência das empresas distribuidoras de energia elétrica. Para contornar o problema, é implementado o método Bayesiano de regionalização para análise espacial de índices de eficiência, permitindo a criação de regiões contíguas que apresentem condições ambientais mais homogêneas. Essa metodologia foi proposta originalmente para problemas de epidemiologia e foi adaptada para o problema proposto. A segunda parte do trabalho estende esta metodologia incluindo um modelo de regressão espacial onde, além do número e das posições dos conglomerados espaciais, ou clusters, pode-se estimar o impacto de cada covariável em cada um deles. A metodologia atualizada é utilizada para a análise da Duração Equivalente de Interrupção por Unidade Consumidora (DEC), um importante indicador de performance no setor de distribuição. Os resultados demonstram a possibilidade de estimar o número de clusters, suas posições e os coeficientes de regressão associados às variáveis que impactam o indicador.

Palavras-chaves: Setor elétrico, *Clustering*, Regressão Espacial, Estatística Bayesiana.

Abstract

In the economy segments that operate under a natural monopoly condition, the absence of competition among suppliers can generate unwanted results for consumers, such as low-quality services at a high cost. To avoid this scenario, the regulatory agency's presence, which must determine the process parameters and evaluate these companies' results, is usual. In the context of electricity distribution in Brazil, this role is exercised by the National Electric Energy Agency (ANEEL). Companies' operating parameters and performance indices are affected by manageable and non-manageable variables related to their operating environment. These environmental effects are important in Brazil, given the breadth and heterogeneity of its territory. The first part of this work explores the effect of environmental conditions' heterogeneity and its impact on electric energy distribution companies' efficiency scores. To circumvent the problem, the Bayesian method of regionalization is implemented for spatial analysis of efficiency indexes, allowing the creation of contiguous regions that present more homogeneous environmental conditions. This methodology was proposed for the epidemiology problem and was adapted to the proposed problem. The second part of the work extends this methodology to include a spatial regression model where, in addition to the number and positions of the clusters, each covariate's impact on each one can be estimated. The updated methodology is used to analyze the *Duração Equivalente de Interrupção por Unidade Consumidora* (DEC) index, an important performance indicator in the distribution sector. The results demonstrate the possibility of estimating the number of clusters, their positions, and the regression coefficients associated with the variables that impact the indicator.

Keywords: Electrical sector, *Clustering*, Spatial Regression, Bayesian Statistics.

Sumário

1 INTRODUÇÃO 8

1.1 Estimando conglomerados espaciais utilizando os índices de eficiência operacional das empresas brasileiras de distribuição de energia elétrica 8

1.2 Desenvolvimento de um modelo de regressão espacial Bayesiano aplicado à análise da duração equivalente de interrupção de energia elétrica por unidade consumidora 9

2 BAYESIAN DETECTION OF CLUSTERS IN EFFICIENCY SCORE MAPS: AN APPLICATION TO BRAZILIAN ENERGY REGULATION 11

2.1 Introduction 11

2.2 Related work 13

2.3 Material and methods 14

2.3.1 The Bayesian model 14

2.3.2 Reversible jump MCMC algorithm 16

2.3.3 The 2015 Brazilian Data Envelopment Analysis model 19

2.3.4 The database 19

2.3.5 Simulation study 20

2.3.6 Estimating the posterior location of the clusters 21

2.4 Results 22

2.5 Conclusions 29

2.6 Acknowledgements 29

3 A NOVEL CLUSTERING-BASED SPATIAL REGRESSION MODEL APPLIED TO CONSUMER POWER OUTAGE INDICATOR 30

3.1 Introduction 30

3.2 Material and methods 32

3.2.1 Literature review 32

3.2.2 The DEC indicator 34

3.2.3 Multiple Linear Regression Model 35

3.2.4 Bayesian Regression Model 36

3.2.5 Bayesian Regression Model with Spatial Clusters 37

3.2.5.1 Prior distribution for the number of clusters 38

3.2.5.2 Prior distribution for the coefficients and variance 38

3.2.6 Reversible Jump Markov Chain Monte Carlo 38

3.2.6.1 Birth Step 39

3.2.6.2 Death Step 40

3.2.6.3 Update Step 40

3.2.7 Estimating the location of the clusters 40

3.2.8 The database 41

3.2.9 Simulation study 44
3.3 Results 46
3.3.1 Simulation results 46
3.3.2 Analysis of the consumer power outage indicator 46
3.4 Discussion 58
3.5 Conclusion 59

REFERÊNCIAS 61

A APÊNDICE DO PRIMEIRO ARTIGO 66
A.1 Details of the RJMCMC sampler 66

B APÊNDICE DO SEGUNDO ARTIGO 68
B.1 Details of the RJMCMC sampler 68

1 Introdução

O mercado de distribuição de energia elétrica no Brasil, como na maioria dos países, opera em monopólio natural. Consequentemente, os consumidores finais não podem escolher a distribuidora de energia com tarifas baixas e alta qualidade. Sem uma regulamentação adequada, a falta de concorrência permite que as distribuidoras de energia cobrem preços abusivos sem melhorar a qualidade do serviço.

No Brasil, a Agência Nacional de Energia Elétrica (ANEEL - Agência Nacional de Energia Elétrica), criada em 1996, é o regulador do setor de energia responsável pelos cálculos de tarifas, avaliação da qualidade dos serviços, entre outras atividades relacionadas à geração, transmissão, distribuição e comercialização de energia (ANEEL, 1996). O serviço de distribuição compreende a entrega de energia elétrica a consumidores residenciais e a pequenos consumidores comerciais e industriais.

1.1 Estimando conglomerados espaciais utilizando os índices de eficiência operacional das empresas brasileiras de distribuição de energia elétrica

Com o intuito de estimular a competitividade entre as empresas distribuidoras de energia elétrica, a Agência Nacional de Energia Elétrica (ANEEL) utiliza técnicas de *Benchmarking* para determinar seus escores de eficiência relativa. Este procedimento permite encontrar as empresas que estão na fronteira de eficiência - *Benchmark* do setor - e comparar as demais empresas ante esse limiar. Desde 2011, parte do modelo regulatório implementado pela ANEEL utiliza o *Data Envelopment Analysis* (DEA).

Uma das premissas do DEA é que as empresas estejam submetidas às mesmas condições ambientais. No contexto da distribuição de energia no Brasil, essa premissa é problemática, pois as empresas distribuidoras atuam em condições ambientais muito distintas uma das outras. Uma forma de lidar com essas diferenças ambientais é corrigir o índice de eficiência gerado pelo modelo DEA com uma análise que leve em consideração os fatores ambientais, o chamado ajuste em segundo estágio. Na primeira parte da pesquisa, Capítulo 2, propõe-se uma forma alternativa de tratar este problema. Aplicando o método Bayesiano de regionalização, pretende-se agrupar as empresas em regiões contíguas homogêneas, ajustando o problema ao requisito de homogeneidade dos modelos de *Benchmarking*.

A Estatística Espacial dispõe de diversas metodologias para lidar com dados que estão associados a algum tipo de componente espacial. Usualmente, nos problemas de regionalização, o número de regiões em que a área de estudo deve ser dividida é um dado pré-estabelecido. Neste trabalho, assume-se que esta informação não está disponível e precisa ser também estimadas. De forma a lidar com esta dificuldade, a parte inicial do trabalho busca adaptar a metodologia de regionalização proposta por Knorr-Held (KNORR-HELD; RASSER, 2000) para análise de dados numéricos geo-referenciados das empresas Brasileiras de distribuição de energia elétrica.

Uma metodologia não-paramétrica Bayesiana de regionalização utilizando o método *Markov Chain Monte Carlo* com Saltos Reversíveis, Green (GREEN, 1995) é proposto. Dado um atributo de interesse, a metodologia particiona a área em clusters que sejam homogêneos em relação à variável de interesse analisada. Este trabalho foi publicado na revista *Applied Mathematical Modelling* no ano de 2019 com o título *Bayesian detection of clusters in efficiency score maps: An application to Brazilian energy regulation.*

1.2 Desenvolvimento de um modelo de regressão espacial Bayesiano aplicado à análise da duração equivalente de interrupção de energia elétrica por unidade consumidora

A atividade de uma distribuidora de energia elétrica, doravante denominada Operador do Serviço de Distribuição (OSD), envolve processos diversos e complexos como manutenção dos ativos de energia elétrica, atendimento ao cliente, entrega de energia, entre outros. Assim, o regulador brasileiro propôs indicadores-chave de desempenho para monitorar a qualidade do serviço de distribuição de eletricidade. Entre os principais indicadores de desempenho propostos, o regulador avalia a falta de energia fornecida, ou um indicador de falta de energia ao consumidor, denominado DEC (*Duração Equivalente de Interrupção por Unidade Consumidora*) (ANEEL, 2016). O indicador DEC mede o tempo médio de interrupção do serviço de entrega de energia ao consumidor. De fato, o indicador DEC é calculado como a média do indicador de falta de energia ao consumidor entre as áreas geográficas de eletricidade de uma determinada empresa. Cada área geográfica, doravante denominada conjunto elétrico, é definida pelo regulador a partir da quantidade de ativos elétricos e da quantidade de consumidores em cada área, antes do cálculo do indicador DEC global.

Além disso, para cada conjunto elétrico, o regulador estima um limite superior para o indicador DEC. Se o indicador DEC observado ultrapassar esse limite regulatório, serão cobradas multas. Além disso, o OSD também deve compensar os consumidores urbanos se a queda de energia for superior a 2 horas e os consumidores rurais se a queda de energia for superior a 5 horas. No ano de 2019, as indenizações por falta de energia no Brasil foram estimadas em R$ 617.718.741,81 (ANEEL, 2019).

Conforme demonstrado, o indicador DEC tem impactos financeiros expressivos. No entanto, o indicador DEC tem prós e contras. Uma das principais vantagens é a simplicidade, o que facilita para o OSD manter o controle de qualidade e intervir, se necessário. Em contraste, o indicador DEC resume uma complexa atividade de distribuição. Assim, não é trivial avaliar os impactos financeiros das decisões individuais de gestão sobre o indicador de falta de energia ao consumidor e, consequentemente, sobre as compensações pagas referentes à falta de energia.

Com base em evidências técnicas, suspeita-se que variáveis ambientais, como a precipitação, afetem o indicador DEC, assim como o tamanho das equipes de manutenção. Dados os recursos limitados, é de extrema importância que os OSDs avaliem, quantitativamente, os principais direcionadores operacionais e ambientais e seus efeitos potenciais no indicador DEC.

O uso de informações geográficas na análise do desempenho dos OSDs brasileiros foi introduzido pela primeira vez por (GIL et al., 2017). Resumidamente, o Brasil possui grande diversidade ambiental e socioeconômica, principalmente devido à sua dimensão continental. Portanto, é improvável que apenas fatores gerenciais afetem o desempenho dos OSDs. No entanto, uma alternativa simples para a modelagem de fatores ambientais e socioeconômicos, quanto pertinentes, consiste em segregar a região estudada em áreas geográficas menores nas quais os OSDs localizados na mesma área são semelhantes com relação aos fatores ambientais e socioeconômicos. A mesma abordagem pode ser aplicada à análise dos OSDs brasileiros. Por exemplo, alguns OSDs brasileiros têm áreas de concessão maiores do que países europeus. Assim, a área de concessão pode ser dividida geograficamente para ajustar a sua heterogeneidade ambiental e socioeconômica.

O trabalho descrito no Capítulo 3 propõe um modelo Bayesiano de regressão espacial baseado aplicado à análise do indicador DEC. O modelo de regressão inclui variáveis operacionais, financeiras e climáticas como variáveis independentes. A regressão espacial permite estimar coeficientes que variam geograficamente, permitindo aprimorar a capacidade preditiva do modelo. Por exemplo, os resultados indicam que há três conglomerados espaciais associados aos impacto da variável ambiental no indicador DEC. Para cada conglomerado, é estimado um coeficiente de regressão diferenciado. Em regiões de baixa precipitação os impactos no indicador DEC são mais atenuados do que considerando a componente ambiental em conglomerados de intensa precipitação. Comportamentos regionalizados para as demais variáveis preditoras foram estimados e são condizentes com o conhecimento prévio de especialistas do setor. Os resultados também mostram que o modelo proposto atinge um coeficiente de determinação preditivo de $R^2_{pred} = 67{,}6\%$, o que constitui um modelo razoavelmente preciso. A partir do modelo ajustado, a empresa de distribuição pode direcionar futuras decisões gerenciais para reduzir as interrupções de energia do consumidor e, consequentemente, as compensações pagas aos consumidor. Esta é a primeira proposta de um modelo de regressão espacial baseado em conglomerados aplicado à análise de indicadores de falta de energia.

O artigo referent ao Capítulo 3 foi recentemente submetido à revista *Applied Mathematical Modelling* e se encontra em processo de revisão.

2 Bayesian detection of clusters in efficiency score maps: an application to Brazilian energy regulation

2.1 Introduction

In 2011, the Brazilian regulator (ANEEL) first applied benchmarking models to estimate the efficiency costs, i.e., efficient operational costs, of the electricity energy distribution utilities, hereafter named DSOs (distribution service operators). Efficient costs are the upper bound cost estimated by the regulator and which DSOs can charge consumers in the electricity distribution tariff, in the following years.

Electricity distribution is a classic case of natural monopoly, due to the technological and economic features of this service which allow a single provider, in general, to meet the overall demand at a lower cost. Consequently, competition does not thrive under these conditions. Eventually, all firms but one will either exit the market or fail (LAZAR, 2011). If no competition exists, then energy tariffs can be overpriced and energy quality can be compromised. The use of benchmarking methodologies aims at creating an artificially competitive market in which the regulator imposes constraints on the tariff prices and the modes of production. These constraints, therefore, avoid overpricing, retain the cost advantage that monopolists may take, but also allow companies to receive fair economic revenues (BOGETOFT; OTTO, 2011) Thus, efficient costs must guarantee a proper balance among DSOs' revenues, quality of service and fair tariffs to final consumers. Furthermore, efficient costs are estimated using models such as data envelopment analysis (DEA) (CHARNES; COOPER; RHODES, 1978; FARRELL, 1957; BANKER; CHARNES; COOPER, 1984), stochastic frontier analysis (SFA) (AIGNER; LOVELL; SCHMIDT, 1977), among others (WINSTEN, 1957; KUOSMANEN, 2012). The ratio between observed cost and efficient cost is named as the efficiency score and lies within the range 0-1.

ANEEL has applied DEA using the total of 61 Brazilian DSOs. DEA comprises a weighted linear model which is estimated using linear programming techniques. The current 2015 DEA model has mean operational costs as the input variable and underground network, overhead network, high voltage network, total number of consumers, weighted energy market, non-technical losses and consumer-hour of interrupted energy as output variables. Input and output variables comprises average values using yearly data from 2011 to 2013. In addition, the data set comprises 13 non-discretionary or environmental variables. These latter variables are not related to the electricity distribution process, but to the environment in which the DSOs are located. The current model produces extremely lower efficiency scores for some companies, which may reach 32.39%. Consequently, some DSOs may bankrupt if the current cost efficiencies are applied. These results have been criticized by Bogetoft (BOGETOFT, 2014), Bogetoft and Lopes (BOGETOFT; LOPES, 2015) and Lopes et al. (LOPES et al., 2016). To overcome these limitations, ANEEL proposed a discretionary solution of a maximum cost reduction rate of 5% per year.

Benchmarking, or *best-practice* modeling, requires a data set of comparable DSOs. DSOs must produce the same outputs using the same inputs and are subject to similar environmental conditions. In practice, there is heterogeneity with respect to production as well as environment. Therefore, one may claim that DSOs with similar operating and environmental conditions must be clustered prior to benchmark modeling. Thus, DSOs heterogeneity can be minimized. Nevertheless, a major point in clustering analysis is the estimate of the optimal number of clusters, or groups and their most likely locations. Most statistical clustering methods assume that the number of groups, say k, is known in advance. Then, clustering analysis aims at estimating the elements, or DSOs in each k-th group based on similarity statistics. These similarity statistics can use production variables, i.e., input and output variables, environmental variables or a mix of production and environmental variables.

Gil et al. (GIL et al., 2017) first found statistical evidence of spatial similarities for the 2015 Brazilian estimated efficiencies. That is, the spatial distribution of the estimated Brazilian DSOs cost efficiencies were not randomly scattered within the Brazilian territory. Spatial statistical analysis showed that DSOs with lower estimated efficiencies were, on average, geographically closer. Similarly, DSOs with larger estimated efficiencies were, on average, also geographically closer. Similar statistical results were found using environmental information. For example, DSOs with high precipitation are geographically closer, as well as DSOs with low precipitation. Recently, Silva et al. (SILVA et al., 2018) found substantial changes in DSO operational costs if environmental information is included in the current Brazilian benchmarking model. It can be argued that the efficiencies are geographically clustered because of geographically similar production conditions, or similar environmental conditions, or a mixture of similar production and environmental conditions.

Given prior evidence of spatial clustering of Brazilian DSOs, this work proposes a spatial clustering method for cost efficiencies. The main motivation is the Brazilian energy distribution regulation, which comprises utilities with different sizes and scattered in a large geographical territory. A Bayesian framework is proposed in which a prior distribution for the number of clusters is chosen and then the posterior distribution of the number of clusters, given the data, is estimated using a reversible jump Markov Chain Monte Carlo (RJMCMC) simulation. This posterior distribution provides detailed statistical information about the number of clusters in the data and their locations. Point estimates for the number of clusters can be calculated using the mean, median or mode of the posterior distribution. In addition, a high probability density (HPD) interval for the number of clusters is estimated. Briefly, a model which applies the reversible-jump Markov Chain Monte Carlo algorithm, to identify the number and the location of spatial clusters assuming a Gaussian distribution, is proposed.

Results using simulated data with different numbers of clusters show that the proposed method is able to estimate the true number of clusters. Using the 2015 Brazilian data, the posterior mode indicated two clusters of DSOs: one cluster with DSOs having large cost efficiencies and the second cluster with DSOs having lower cost efficiencies. To the best of our knowledge, this is the first work to present a spatial statistical procedure that estimates the number of groups in energy regulation.

This paper is organized as follows. Section 2 presents the literature review about clustering techniques applied to benchmarking models. Section 3 presents the proposed Bayesian clustering model, the simulation study and the case study. Section 4 presents the results and section 5 presents the conclusion.

2.2 Related work

Clustering of DSOs into homogeneous groups for benchmarking analysis has been proposed by Llorca et al. (LLORCA; OREA; POLLITT, 2014) for the electricity transmission industry. Llorca et al. propose the Latent Class Model (LCM) in which, given a cost function, each firm i can be allocated to a group j ($j \in 1, .., J$) by means of a group membership probability $p(j|i)$. The cost function estimates are different in each group j. Estimates for all parameters, including group membership probability are achieved using maximum likelihood. The total number of groups J is known in advance. Dai and Kuosmanen (DAI; KUOSMANEN, 2014) apply hierarchical clustering algorithms, partitioning methods and model based clustering methods to identify similar groups of DSOs in energy regulation. Output-input ratios are used as the clustering variables. The final model includes Normal Mixture Model (NMM) (MCLACHLAN; BASFORD, 1988) for group clustering, and StoNED (KUOSMANEN, 2012) for efficiency analysis. Using NMM, the number of clusters of DSOs are estimated. Similarly, Samoilenko and Osei-Bryson (SAMOILENKO; OSEI-BRYSON, 2008) apply cluster analysis, DEA and decision trees (RAZI; ATHAPPILLY, 2005) to estimate and evaluate relative efficiencies of DSOs. Agrell et al. (AGRELL et al., 2014) apply the LCM to cluster DSOs into homogeneous groups. In sequence, DEA, Stochastic Frontier Analysis (AIGNER; LOVELL; SCHMIDT, 1977) and Modified OLS (AFRIAT, 1972) are used to estimate cost efficiencies. The number of groups is fixed as four. It is worth mentioning that most of the aforementioned clustering methods require the number of clusters be known in advance. Furthermore, none of the methods accounts for any geographical information.

The Bayesian approach was originally applied to SFA models by Broeck et al. (BROECK et al., 1994). The Bayesian paradigm states that, given the prior distribution about a parameter of interest Θ, say $p(\Theta)$, a vector of data (D) and the likelihood function $p(\mathbf{D}|\Theta)$ the prior information can be updated using the Bayes Theorem $p(\Theta|D) \propto p(\mathbf{D}|\Theta) \times p(\Theta)$, where $p(\Theta|D)$ is known as the posterior distribution. If the prior distribution is flat, $p(\Theta) \propto 1$, then the posterior distribution is proportional to the likelihood function. Flat prior distributions are known as weakly informative distributions. Informative distributions represent expert information which can be combined to data using the Bayes Theorem.

Gil et al. (GIL et al., 2017) first applied spatial statistical methods to Brazilian energy distribution utilities. Spatial statistical methods were applied to test the spatial correlation of cost efficiencies. A Bayesian second stage model, accounting for spatial latent structure, was proposed to estimate corrected efficiencies.

Using the Bayesian approach, sophisticated models can be estimated using informative, weakly informative or non-informative prior distributions. Although mathematical estimates are nontrivial to very complex models, samples of the posterior distributions can be obtained using Markov Chain Monte Carlo (MCMC) methods (GELMAN, 2014). Furthermore, reversible jump Markov Chain Monte Carlo (GREEN, 1995) computation generates samples of posterior distributions for vectors of parameters with varying dimensions. Using the posterior distribution, Bayesian inference and High Probability Density intervals (CHEN; SHAO, 1999) for the parameters of interest are achieved.

Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000) first applied RJMCMC to estimate geographical variations in disease rates. The proposed Bayesian model assumes a Poisson likelihood function (CLAYTON; BERNARDINELLI, 1992). The outcome is the number of cases in each area. The RJMCMC generates samples of the posterior distribution of the vector of parameters, including the number of clusters. In general, implementing a RJMCM algorithm is not a trivial task. Brooks et al. (BROOKS; GIUDICI; ROBERTS, 2003) provide detailed information about efficient construction of RJMCMC.

The adapted RJMCMC methodology resembles a second stage analysis (RAY, 1988), in which a function of the estimated efficiencies are used as the dependent variables and environmental variables are used as covariates in statistical regression models. This approach was first introduced by Banker and Morey (BANKER; MOREY, 1986). The adapted RJMCMC method assumes that DSO efficiencies are the random variables of interest. The statistical model assumes that the mean parameter of the efficiencies may change over the space, thus affecting local groups of DSOs. Nonetheless, the number of means, i.e., the number of groups, their locations and the means are unknown and must be estimated.

2.3 Material and methods

2.3.1 The Bayesian model

Adapted from Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000), suppose the data comprises random variables Y_i measured in a set of n DSO regions, $i = 1,...,n$. The main idea is that the mean parameter is the same for a set of one or more contiguous regions, $Y_i \sim Normal(\mu_j,\sigma^2)$. Therefore, cluster $C_j \subset \{1,...,n\}$ is defined as a set of adjacent regions with mean μ_j. The definition of cluster implies that the clusters $C_1,...,C_k$ cover all the studied area and there is no overlap among them: $C_1 \bigcup ... \bigcup C_k = 1,...,n$. In the limiting case: $k = 1$, the mean is the same for all regions, whereas for $k = n$, each region has its own mean parameter. It is assumed that the response variables Y_i, $i = 1,...,n$, are conditionally independent, given the mean vector $M_k = (\mu_1,...,\mu_k)$. The likelihood function of the response vector $\mathbf{y} = (y_1,..,y_n)$ is defined as:

$$L(\mathbf{y} \mid M_k, \sigma^2) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{1}{\sigma} \phi \left(\frac{y_i - \mu_j}{\sigma} \right) \tag{2.1}$$

where $\phi(.)$ is the density of the standard normal distribution.

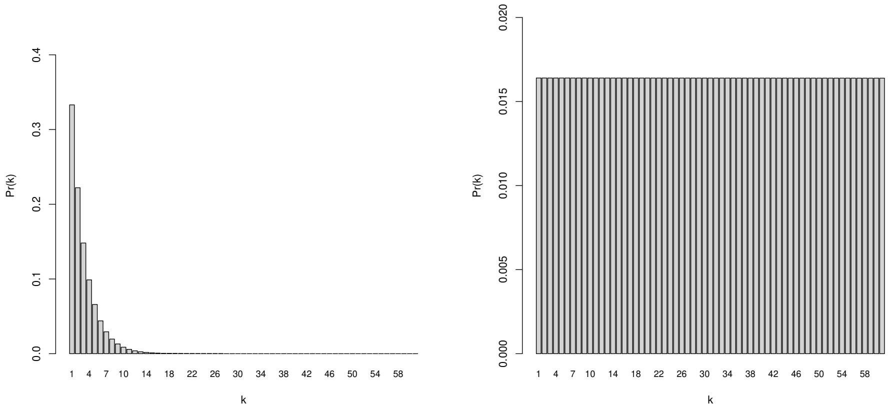
The clusters model

As the first step in the definition of clusters with size k, k regions $g_1,...,g_k$ are selected as centers. Each center $g_j \in 1,...,n$ defines a cluster C_j with $g_j \in C_j$. The vector of centers $G_k = (g_1,...,g_k)$ defines a clustering configuration, i.e., every region belongs to a cluster. Furthermore, let $d(i_1, i_2)$ be the measure of distance between regions i_1 and i_2, defined as the minimum number of geographical boundaries that have to be crossed to move from i_1 to i_2. This measure can be calculated using the adjacency matrix (CRESSIE, 2015). The distance $d(i_1, i_2)$ is used to assign each area to one of the clusters. Each region i is assigned to the nearest cluster center. Nonetheless, the order of the cluster centers in vector G_k creates priority for selecting areas. For example, center g_1, which occupies the first position in vector G_k, has priority to select the nearest areas. In sequence, center g_2 has preference in selecting the remaining nearest areas, and so on. Therefore, a cluster configuration defined by vector $G_2 = (1,2)$ is, in general, different from the cluster configuration $G_2^* = (2,1)$. Consequently, the space of possible cluster configurations is large which improves the mixing property of the RJMCMC algorithm.

Prior distribution for the number of clusters

Based on Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000), the prior distribution for the number of clusters, $Pr(k)$, $k = 1,...,n$ is proportional to $(1 - c)^k$, where $c \in [0,1)$ is a parameter defined by the user. Small values of c represent a weakly informative prior distribution. Figure 2.1 shows two different prior distributions for k. Using $c = 0.333$ the prior mean is 3, whereas using $c = 0.001$ the prior distribution is similar to a discrete uniform distribution with a prior mean of 31 ($n = 61$).

$$Pr(k) \propto (1 - c)^k. \tag{2.2}$$



(a) Prior distribution for k using $c = 0.333$. (b) Prior distribution for k using $c = 0.001$.

Figura 2.1 – Prior distributions for the number of clusters using different values for c.

Prior distribution for the vector of means

It is assumed that the vector of means $M_k = (\mu_1,\ldots,\mu_k)$ comprises independent and identically distributed components from a normal distribution with mean μ_0 and variance σ_0^2. Therefore, the prior distribution for vector M_k is:

$$Pr\left(M_k \mid k,\mu_0,\sigma_0^2\right) = \left(\frac{1}{\sqrt{2\pi}\sigma_0}\right)^k \prod_{j=1}^{k} \phi\left(\frac{\mu_j - \mu_0}{\sigma_0}\right). \tag{2.3}$$

In other words, $\mu_j \sim Normal(\mu_0,\sigma_0^2)$. In practice, using standardized values of y_i, $y_i^* = (y_i - \bar{y})/sd(y)$, where $\bar{y}$ is the sample mean and $sd(y)$ is the sample standard deviation, the natural choices are $\mu_0 = 0$ and $\sigma_0^2 = 1$.

2.3.2 Reversible jump MCMC algorithm

Samples from the posterior distribution for the number of clusters are generated using a reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm. Each sample comprises values of σ^2, M_k and G_k in a given step of the RJMCMC algorithm, described below. Initially, a start configuration for the number of clusters (k and G_k), mean vector (M_k) and variance parameter (σ^2) is created. Then, given the value of the c parameter, the RJMCMC algorithm randomly chooses one of the following five moves: *birth*, *death*, *shift*, *switch* and *update* moves with probabilities of π_B, π_D, π_{Sh}, π_{Sw} and π_{Up}, respectively. These probabilities are pre-specified by the user. Figure 2.2 shows the flowchart of the RJMCMC algorithm and each move is described next.

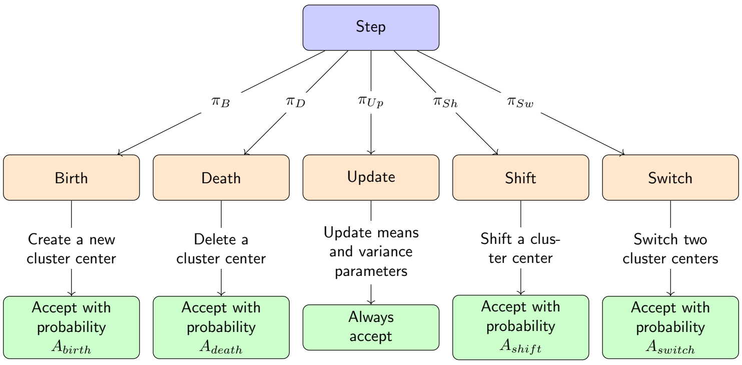

Step
π_B
π_D
π_{Up}
π_{Sh}
π_{Sw}
Birth
Death
Update
Shift
Switch
Create a new cluster center
Delete a cluster center
Update means and variance parameters
Shift a cluster center
Switch two cluster centers
Accept with probability A_{birth}
Accept with probability A_{death}
Always accept
Accept with probability A_{shift}
Accept with probability A_{switch}

Figura 2.2 – RJMCM algorithm flowchart.

Birth move

If the *birth* move is selected, a new cluster configuration is created by randomly selecting a new cluster center among the areas which are not cluster centers. This new area is randomly imputed in the vector of centers, creating the new vector G_{k+1}. Given the r-th position of the new cluster center, $r \in \{1,\ldots,k+1\}$, a new vector of means, M_{k+1}, is created. The mean parameter μ_r of the new cluster center is randomly selected using a normal distribution with a mean of:

$$\mu_* = \frac{\sigma^2}{n_r\sigma_0^2 + \sigma^2} \cdot \mu_0 + \frac{n_r\sigma_0^2}{n_r\sigma_0^2 + \sigma^2} \cdot \bar{y}_r, \tag{2.4}$$

and variance of:

$$\sigma_*^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n_r}{\sigma^2}}, \tag{2.5}$$

where n_r is the number of areas in the new cluster, $\bar{y}_r$ is the observed mean in the new cluster, μ_0 and σ_0^2 are the parameters of the prior distribution. This proposal distribution, $\varphi(\mu_r) = \phi\left(\frac{\mu_r-\mu_*}{\sigma_*}\right)$ is the full conditional distribution, i.e., the prior distribution for μ_r times the relevant likelihood, times a normalizing constant. The new cluster configuration with dimension $k+1$ is accepted with probability given by Eq. (2.6).

$$A_{Birth} = \frac{L(\mathbf{y} \mid M_{k+1}, G_{k+1}, \sigma^2)}{L(\mathbf{y} \mid M_k, G_k, \sigma^2)} \cdot (1-c) \cdot \frac{\phi\left(\frac{\mu_r-\mu_0}{\sigma_0}\right)}{\phi\left(\frac{\mu_r-\mu_*}{\sigma_*}\right)}, \tag{2.6}$$

where $(1-c) = \frac{Pr(k+1)}{Pr(k)}$ is the prior ratio of the number of clusters, which penalizes jumps from k to $k+1$.

If accepted, the new cluster configuration (G_{k+1} and M_{k+1}) replaces the previous configuration (G_k and M_k). Therefore, the state of the RJMCMC is now of dimension $k+1$.

Death move

If the *death* move is selected, a new cluster configuration is created by randomly deleting one of the current cluster centers. Thus, a cluster center g_r ($r \in 1,...,k$) in vector M_k is randomly deleted, creating the new vector G_{k-1}. The associated mean parameter μ_r is also deleted from vector M_k, creating the new vector M_{k-1}. The new cluster configuration with dimension $k-1$ is accepted with probability given by Eq. (2.7).

$$A_{Death} = \frac{L(\mathbf{y} \mid M_{k-1}, G_{k-1}, \sigma^2)}{L(\mathbf{y} \mid M_k, G_k, \sigma^2)} \cdot \frac{1}{(1-c)} \cdot \frac{\phi\left(\frac{\mu_r-\mu_*}{\sigma_*}\right)}{\phi\left(\frac{\mu_r-\mu_0}{\sigma_0}\right)}, \tag{2.7}$$

where $\frac{1}{(1-c)} = \frac{Pr(k-1)}{Pr(k)}$. If accepted, the new cluster configuration (G_{k-1} and M_{k-1}) replaces the previous configuration (G_k and M_k). Therefore, the state of the RJMCMC is now of dimension $k-1$.

Shift move

If the *shift* move is selected, a new cluster configuration is created by shifting a randomly selected cluster center in vector G_k, say g_r, to a randomly selected area, also defined in cluster g_r. Briefly, a cluster center is shifted towards one of the areas in the selected cluster, which is not the current center. Thus, creating a new vector G_k^*. The dimension of the new cluster configuration (k) and the vector of means (M_k) are not changed. The new cluster configuration with dimension k is accepted with probability given by Eq. (2.8).

$$A_{Shift} = \frac{L(\mathbf{y} \mid M_k, G_k^*, \sigma^2)}{L(\mathbf{y} \mid M_k, G_k, \sigma^2)} \cdot \frac{n(G_k)}{n(G_k^*)} \cdot \frac{m(g_r)}{m(g_r^*)} \tag{2.8}$$

where $m(g_r)$ is the number of free neighbors in cluster g_r, i.e., the number of areas in cluster g_r which are not the cluster center, and $n(G_k)$ is the number of cluster centers with non-zero free neighbors.

Switch move

If the *switch* move is selected, then two cluster centers in vector G_k are switched. Initially, two clusters indexes, say i and j ($i \neq j$, $i,j \in 1,...,k$) are randomly selected. Then, the cluster centers g_i and g_j are switched in vector G_k, creating the new vector G_k^*. In addition, the mean parameters μ_i and μ_j are accordingly switched in vector M_k, creating the new vector M_k^*. The new cluster configuration with dimension k is accepted with probability given by Eq. (2.9).

$$A_{Switch} = \frac{L(\mathbf{y} \mid M_k^*, G_k^*, \sigma^2)}{L(\mathbf{y} \mid M_k, G_k, \sigma^2)}. \tag{2.9}$$

If accepted, the new cluster configuration (G_k^* and M_k^*) replaces the previous configuration (G_k and M_k). It is worth noticing that the new state of the RJMCMC is also of dimension k.

Update move

If the *update* move is selected, then the mean parameters μ_j, $j = 1,...,k$, of vector M_k and the variance parameter σ^2 are changed. The mean parameter of each cluster, μ_j, is changed using a normal distribution with a mean of:

$$\mu_* = \frac{\sigma^2}{n_j\sigma_0^2 + \sigma^2} \cdot \mu_0 + \frac{n_j\sigma_0^2}{n_j\sigma_0^2 + \sigma^2} \cdot \bar{y}_j \tag{2.10}$$

and variance of:

$$\sigma_*^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n_j}{\sigma^2}}, \tag{2.11}$$

where n_j is the number of areas in cluster j, $\bar{y}_j$ is the observed mean in cluster j, μ_0 and σ_0^2 are the parameters of the prior distribution. Similar to Eq. (2.4) and (2.5), Eq. (2.10) and (2.11) represent a normal full conditional distribution, i.e., the prior distribution for μ_j times the relevant likelihood.

In sequence, the σ^2 parameter is updated assuming a weakly informative prior distribution. The variance parameter σ^2 is changed using an inverse-chi-squared distribution with $n-k$ degrees of freedom and scaling parameter of s^2:

$$\sigma^2 | \mathbf{y}, M_k \sim \text{Inv-}\chi^2_{(n-k, s^2)}, \tag{2.12}$$

where $s^2 = \frac{1}{n-k}\sum_i \left(y_i - \mu_{j_{(i)}}\right)^2$, k is the current number of clusters and n is the number of areas (sample size).

The proposed RJMCMC algorithm, the spectral clustering method and the database are available in the R package *gbdcd* (MINETI; COSTA, 2018).

2.3.3 The 2015 Brazilian Data Envelopment Analysis model

The DSOs cost efficiencies were estimated by the Brazilian regulator using a DEA model (COOPER; SEIFORD; ZHU, 2004) and average yearly data from 2011 to 2013. The current model is presented in Technical Note 66/2015 (ANEEL, 2015) and reproduced below. The NDRS (non-decreasing returns to scale) input-oriented efficiency of the reference DSO (Distribution Service Operator), is calculated using the linear programming model shown in Eq. (2.13).

$$\begin{aligned}
&\theta_0 = \max \sum_{j=1}^{s} \nu_j y_j^0 + \varphi \\
&\text{subject to:} \\
&\sum_{i=1}^{m} u_i x_i^0 \leq 1. \\
&\sum_{j=1}^{s} \nu_j y_j^k - \sum_{i=1}^{m} u_i x_i^k + \varphi \leq 0, \;\; k = 1, \dots, n. \\
&u_i, \nu_j, \varphi \geq 0.
\end{aligned} \tag{2.13}$$

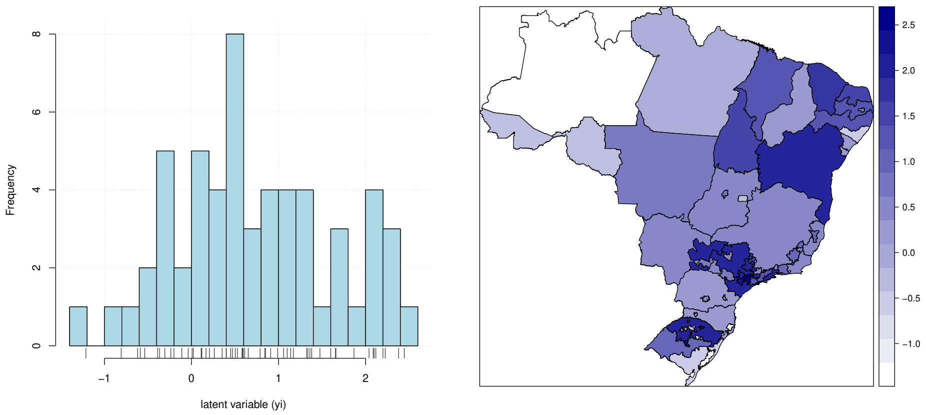
where y_j are the outputs $(j = 1, \dots, s)$, x_i are the inputs $(i = 1, \dots, m)$, ν_j are the output weight parameters, u_i are the input weight parameters, φ is the parameter associated with the non-increasing returns to scale property and θ_0 is the input efficiency estimated for the reference DSO. The DEA model uses operating costs (OPEX) as the input; and, as output variables, number of consumers, weighted power consumption, high level network extension, low level network extension, underground network extension, non-technical losses and duration of interruption of energy. Non-technical losses and duration of interrupted energy are included as negative outputs, which are alternative representations for non-desired inputs in the DEA model (COOK; ZHU, 2013). In addition, weight restrictions are also included in the linear programming model. Further details about the cost efficiency estimates are found in Gil et al. (GIL et al., 2017) and Lopes et al. (LOPES et al., 2016).

2.3.4 The database

The database comprises cost efficient indexes estimated for 61 energy distribution utilities located in Brazil. The estimated cost efficiencies are within the range $0 - 1$ and some of the efficiencies are saturated at 1. The proposed Bayesian cluster model assume that the random variables are normally distributed, which is not the case of the original cost efficiencies. To overcome this limitation, unbiased cost efficiencies were calculated using the original DEA model and the bootstrap procedure proposed by Simar and Wilson (SIMAR; WILSON, 1998). In sequence, the inverse of a logistic function was applied to the unbiased estimated cost efficiencies, generating a new latent variable y_i for each company. The observed y_i values were used in the proposed Bayesian cluster model. The y_i values were calculated using Eq. (2.14),

$$y_i = \log\left(\frac{\tilde{\theta}_i}{1-\tilde{\theta}_i}\right) \tag{2.14}$$

where $\tilde{\theta}_i$ is the unbiased cost efficiency (SIMAR; WILSON, 1998). Figure 2.3 shows the histogram of the estimated latent variables (a) and their spatial distribution in the Brazilian territory (b).



(a) Histogram of the latent variables y_i. (b) Brazilian map of the latent variables y_i.

Figura 2.3 – Graphical analysis of the latent variables y_i.

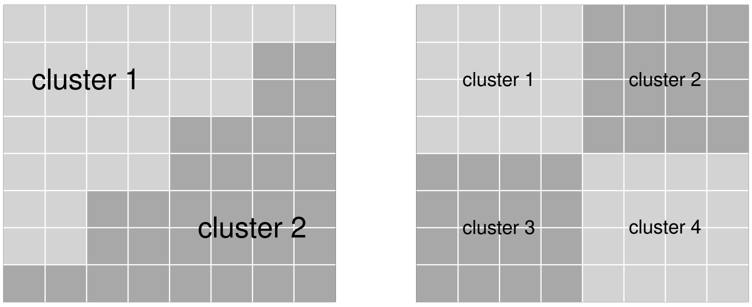
2.3.5 Simulation study

A simulation study was created to evaluate the proposed Bayesian cluster method. A regular 8 × 8 connected grid, with a total of 64 areas, was used as the study region. Three different scenarios were evaluated. The first scenario, named scenario A, has only one cluster. Observed values were created using a standard normal distribution with a mean of zero and variance equal to one, $Y_i \sim Normal\,(0,1)$. The second scenario, named scenario B, has two clusters. Each cluster has 32 areas. Figure 2.4 (a) shows the cluster configurations for scenario B. Observed values were created using a normal distribution with variance equal to one and different means in each cluster. The difference between means, δ, was calculated using Eq. (2.15), which is the minimum distance between two groups of independently and normally distributed variables with similar variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$, which rejects the null hypothesis of equal means $(H_0 : \mu_1 = \mu_2)$.

$$\delta = 2 \cdot Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}, \tag{2.15}$$

where Z is the score statistic, n_1 and n_2 are the sample sizes of groups 1 and 2, respectively, and α is the error type I. If $\alpha = 0.05$, then $Z_{\frac{0.05}{2}} \approx 1.96$.

The third scenario, named scenario C has four clusters. Figure 2.4 (b) shows the cluster configurations for scenario C. The clusters shown in the diagonal share the same mean, say μ_1.



(a) Simulated scenario with 2 clusters. (b) Simulated scenario with 4 clusters.

Figura 2.4 – Simulated scenarios with 2 and 4 clusters.

The clusters shown in the off diagonal also share the same mean μ_2. Consequently, the difference between means was also calculated using Eq. (2.15). Both scenarios B and C use $\mu_1 = 0$ and $\mu_2 = \delta$, or $\mu_2 = \mu_1 + \delta$.

For each scenario, 1,000 simulations were evaluated. The c parameter of the prior cluster distribution was set at 0.3, which represents a prior mean close to 3. Therefore, the prior informative distribution assumes that the number of clusters in the data is small. A start cluster configuration is created by randomly selecting 3 cluster centers, i.e., the initial number of clusters is the prior mean. Initial mean parameters, μ_j, are created using a standard normal distribution. The RJMCM algorithm was executed for 1,000,000 iterations in order to guarantee convergence of the chain, this is known as the *burn-in*. Then, 1,000,000 new iterations were run to get samples of the posterior distribution. Therefore, 1,000,000 cluster configurations, i.e., $[G_k^s, M_k^s]$ for $s \in 1, ..., 1{,}000{,}000$, were sampled. The RJMCMC probabilities of the moves were set as $\pi_B = \pi_D = 0.35$ and $\pi_{Sh} = \pi_{Sw} = \pi_{Up} = 0.10$.

Furthermore, for each simulation, the point estimate of the number of clusters was calculated as the mode of the posterior distribution, i.e., the most frequent cluster size found in the RJMCMC samples. In addition, 95% highest probability density (HPD) intervals (CHEN; SHAO, 1999) of the number of clusters were estimated for each simulation.

2.3.6 Estimating the posterior location of the clusters

Using the sampled cluster configurations, the posterior distribution of the number of clusters can be obtained using the observed sizes of the sampled cluster centers G_k^s. Likewise, the posterior distribution for each area mean $\mu_i|\mathbf{y}$ can be obtained using the observed sampled cluster vector of means M_k^s.

Following Feng et al. (FENG et al., 2016), the posterior estimates of the clusters locations were calculated using the marginal frequency of pairs of areas sharing geographical boundaries and located in the same cluster. Let i and j be two areas sharing geographical boundaries, hereafter represented as $i \sim j$. Let $I_s(i \sim j)$ be the indicator function which is equal to one if the pair $i \sim j$ belongs to one of the clusters in the s sampled RJMCMC iteration, and zero otherwise.

Let $f_{i\sim j} = \sum_s [I_s(i \sim j)]$ be the absolute frequency in which the pair $i \sim j$ was observed in the same cluster in all RJMCMC sampled interations. Using the $f_{i\sim j}$ values, the similarity matrix $S_{n\times n}$ is defined as:

$$S_{i,j} = \frac{f_{i\sim j}}{M}. \tag{2.16}$$

where M is total number of RJMCMC iterations. $S_{i,j}$ represents the empirical probability that areas i and j are grouped in the same cluster based on the RJMCMC samples. Using the Ng-Jordan-Weiss spectral clustering algorithm (NG; JORDAN; WEISS, 2002) an approximate posterior estimate of the cluster configuration with $\hat{k}$ clusters is obtained using matrix S. Feng et al. (FENG et al., 2016) claims that *spectral clustering method generates the posterior mean estimate of the clustering structure using the pairwise cluster membership linkage.*

Briefly, the spectral clustering algorithm works as follows. Let D be a diagonal matrix whose ith diagonal is the sum of S's ith row. Calculate $L = D^{-1/2}SD^{-1/2}$. Find the k largest eigenvalues of matrix L and the associated eigenvectors. Apply a simple K-means cluster algorithm using the rows of the selected eigenvectors as points in $\mathcal{R}^k$. The cluster assignment of each row indicates the clustering membership. Further details about spectral clustering algorithms are found in Elavarasi et al. (ELAVARASI; AKILANDESWARI; SATHIYABHAMA, 2011).

2.4 Results

Table 2.1 shows the simulation results using weakly informative prior cluster size distribution ($c = 0.001$), informative prior cluster size distribution ($c = 0.333$) for scenarios A, B and C. The informative prior cluster size distribution has a prior mean of 3 clusters as shown in Figure 2.1 (a). Table 2.1 shows the average point estimate of the cluster size, the average range of the HPD intervals and the proportion of simulations in which the true cluster sizes are within the HPD interval. It is worth mentioning that the method is very robust regarding changes in the probabilities of the moves. In general, different probabilities affect the speed of convergence but do not affect the posterior samples.

Results show that the average HPD size is larger using the weakly informative prior distribution as compared to the informative prior distribution. This is because the weakly informative prior distribution has a large variance as compared to the informative prior distribution. Thus, the more informative the prior distribution, the smaller the HPD size. Nonetheless, it is worth noticing that the proportion of simulations, in which the true cluster size is within the HPD interval, is close to 1 (100%), using both informative and weakly informative prior distributions.

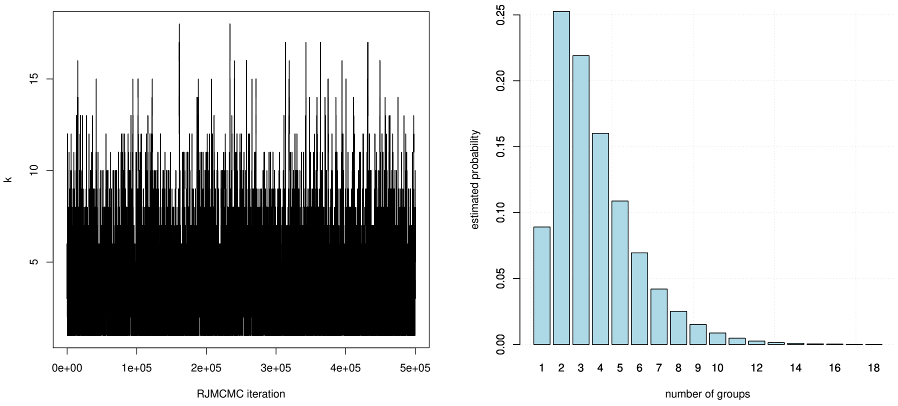
Regarding the point estimates of the number of clusters, the average posterior number of clusters ($\bar{k}|\mathbf{y}$) was close to the true number of clusters in scenarios A and B, using both weakly informative and informative prior distributions. It is worth noticing that using informative prior distribution the point estimates were much closer to the true cluster size, as compared to using weakly informative prior distribution. In scenario C, the weakly informative prior distribution achieved an average posterior number of clusters close to the true number of clusters, as compared to the informative prior distribution. This is because the informative prior distribution has a prior mean smaller than the true number of clusters, i.e., the prior distribution was specified

Tabela 2.1 – Simulated results using scenarios with different cluster sizes, informative and weakly informative prior distributions.

number of clusters	prior c parameter	$\bar{k}\vert\mathbf{y}$	average HPD size	HPD proportion
1	0.001	1.40	10.5	0.99
	0.333	1.19	4.2	1.00
2	0.001	2.33	11.2	1.00
	0.333	2.05	5.1	1.00
4	0.001	3.80	14.8	1.00
	0.333	2.26	6.8	1.00

inaccurately. Nonetheless, although the informative prior distribution was inaccurate in scenario C, the true number of clusters was inside the HPD intervals in all simulations.

Figure 2.5 shows the RJMCMC samples and the posterior distribution of the number of clusters using the Brazilian cost efficiencies database. Results show that the posterior point estimate of the number of clusters is 2 with the 95% HPD interval of 1 to 9 clusters.

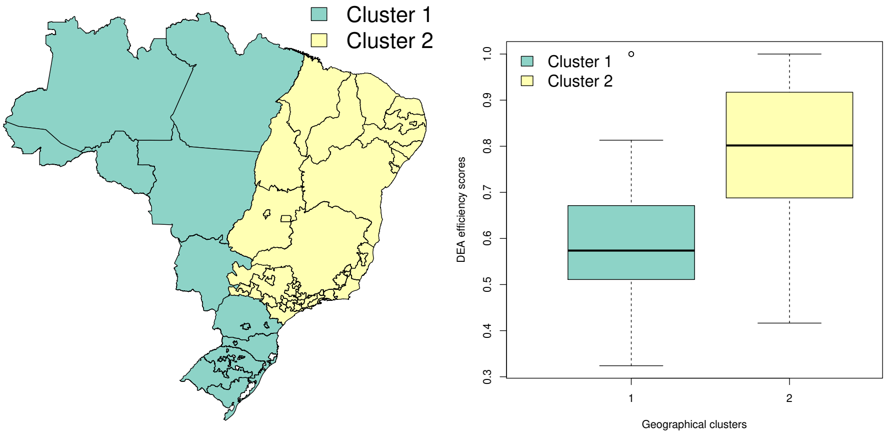


(a) Posterior RJMCMC cluster sizes (k).

(b) Posterior distribution of k.

Figura 2.5 – Posterior distribution of the cluster size using RJMCMC.

Figure 2.6 shows the location of the estimated two clusters in the Brazilian territory using the spectral clustering algorithm. Cluster 1, with higher efficiencies, comprises the DSOs located in the northeast, southeast, one DSO located in the north and one DSO located in central west regions. Cluster 2, with lower efficiencies, comprises DSOs located in north, central-west and south regions. It is worth noticing from Figure 2.6 (b) that one DSO with efficiency of 100% is located in the group of lower efficiencies, and one DSO with efficiency close to 40% is located in the group of higher efficiencies. These DSOs represent extreme cases in each group.



(a) Geographical location of the estimated two clusters.

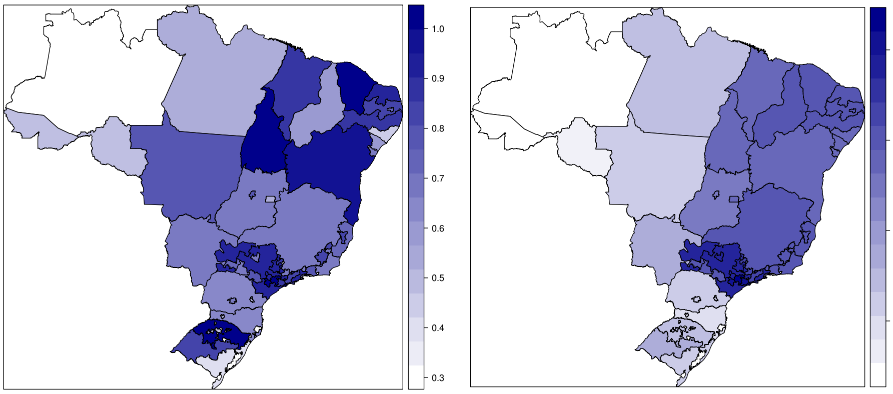
(b) Boxplots of DEA efficiencies in the estimated two clusters.

Figura 2.6 – Estimated results using the posterior point estimate, i.e., the posterior mode of the number of clusters equal to 2.

Using the RJMCMC samples of the vector of means, M_k^s, a smooth map of the DSOs efficiencies can be estimated by applying the logistic function to the point estimates of the mean parameters, $\mu_{j_{(i)}}|\mathbf{y}$, for each DSO i. Figure 2.7 (b) shows the smooth map of the efficiencies. Results indicate a cluster of DSOs with mean efficiencies of 0.75 located in the state of São Paulo. Southeast and northeast regions share a mean efficiency of 0.67. North and south regions share a mean efficiency of 0.63; and two DSOs located in the north have the lowest mean efficiency. Original efficiencies are shown in 2.7 (a) for comparison.

Figure 2.7 shows significant differences for some regions, i.e., some DSOs. These results represent important information for future efficiency improvements. As shown in figure 2.6 (b), there is one DSO with an efficiency of 100% located in the cluster of low efficiencies. Figure 2.7 (b) shows the map of the smoothing efficiencies estimated using the proposed RJMCMC approach. The smoothing map is generated using geographical information of the DSOs, i.e., the smoothing estimate of the efficiency in one DSO represents an average value of geographically closer DSOs. The 100% DSO is located in the south of Brazil (bottom) as can be seen in Figure 2.7 (a). This particular DSO can be named as a benchmark to its peers because, although located in a cluster of lower efficiencies, it managed to achieve a large efficiency. On the contrary, Figure 2.6 (b) also shows one DSO with a lower efficiency located in the cluster of larger efficient DSOs. Figure 2.7 (a) shows that this particular DSO is located in the northeast of Brazil (top right), closer to the coast. This low efficient DSO is geographically surrounded by large efficient DSOs. Therefore, there is evidence that, despite being located in a favorable geographical location, this DSO faces serious management limitations. Furthermore, its neighboring DSOs can be used as benchmarks, i.e., best management practices, for future efficient improvement.

In general, results show statistical evidence of geographical clusters of Brazilian DSOs with respect to their efficiencies. As mentioned, spatial clusters of DSOs indicate similar production conditions, similar environmental conditions or a mixture of both. In Brazil, private and public DSOs are geographically closer. Furthermore, regions are subject to local environmental conditions.



(a) Original DEA efficiencies.

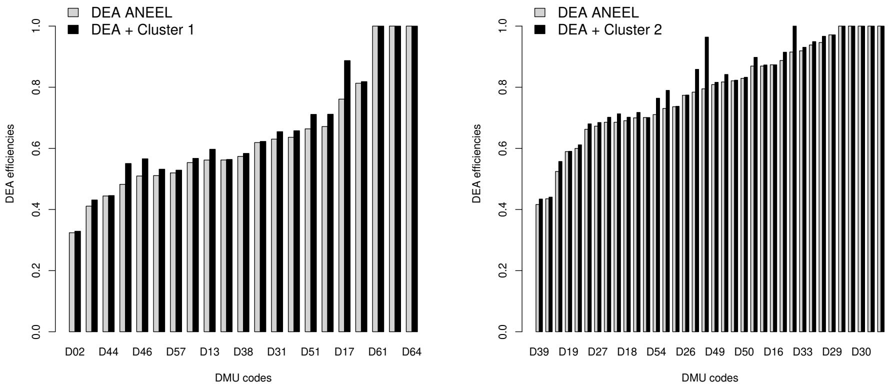
(b) Posterior mean estimates of the cost efficiencies (μ_i).

Figura 2.7 – Comparison between original efficiencies and posterior mean estimates.

The detected cluster with low efficiencies comprises, on average, public DSOs subject to harsh rain conditions. The detected cluster with high efficiencies comprises, on average, private companies subject to mild environmental conditions. It is worth mentioning that the method was built under the assumption that heterogeneity can be detected using the location. Thus, if the available information is random with respect to location, then the method will detect only one cluster.

Figure 2.8 shows the efficiency changes if the ANEEL DEA model (GIL et al., 2017) is applied separately to each cluster. In general, the cluster efficiencies are closer to the original efficiencies. Some significant changes are found for a few DSOs. By splitting the DSOs into different groups, the frontier of each group is changed. By including a geographical component in the clustering process, it can be claimed that the DSOs in the same group are homogeneous with respect to environmental conditions, thereby adjusting the estimated cost efficiencies to environmental conditions. Table A.1 shows the operational costs, the ANEEL DEA efficiencies, the DSOs cluster efficiencies, the ANEEL DEA efficiency costs, the efficiency costs estimated for each cluster and differences between the clusters and original efficiency costs. Results are shown in decreasing order of operational costs differences. DSOs with major differences between original and clusters efficiencies are highlighted. Major changes in operational costs are related to large operational costs and large changes in efficiencies. For example, using the proposed cluster analysis, efficient costs for LIGHT could be increased by R$ 53,871,122.75 or U$ 20,328,725.57 (considering an average exchange rate of R$ 2.65 per U$ 1.00 in year 2014). Similar results are found for CEMAT which is located in a different cluster. Overall, using the proposed cluster analysis, efficient costs could be increased by U$ 140,177,549.31.

A standard hierarchical clustering approach (JOHNSON; WICHERN; others, 2002) was applied in order to illustrate main differences between the proposed spatial clustering approach and a regular statistical approach. Figure 2.9 shows a dendrogram using the original cost efficiencies, i.e., not using any geographical information. The euclidean distance and the Ward hierarchical clustering approach were applied (KAUFMAN; ROUSSEEUW, 2009). Using the dendrogram information, DSOs were divided into two groups as shown in Figure 2.10 (a).



(a) Efficiency changes in cluster 1.

(b) Efficiency changes in cluster 2.

Figura 2.8 – Efficiency changes using the original DEA ANEEL benchmarking model applied separately to each cluster.

Results show that, in general, DSOs with larger efficiencies are not geographically contiguous and are located in the north, northeast, southeast and south regions. It is worth mentioning that there is no DSO with a large efficiency located in the north of Brazil. Figure 2.10 (b) compares the distribution of the estimated efficiencies in both groups of larger efficiencies and smaller efficiencies. Furthermore, some low efficient DSOs are geographically closer to many large efficient DSOs. As previously mentioned, for cost efficient estimation, DSOs with similar prodution and environmental conditions must be clustered prior to benchmark modeling. In addition, Heaton et al. (HEATON; CHRISTENSEN; TERRES, 2017) points that *the choice of dissimilarity metric will change the resulting clusters*. Therefore, using a standard hierarchical clustering method may generate inconsistent results.

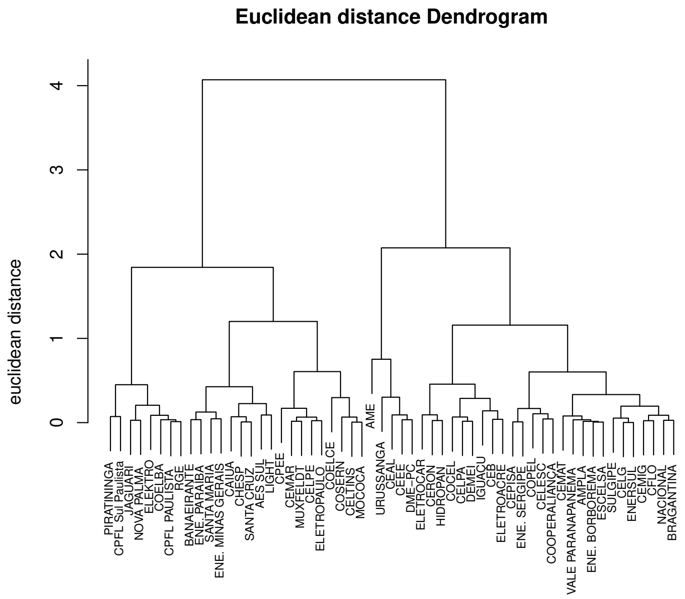
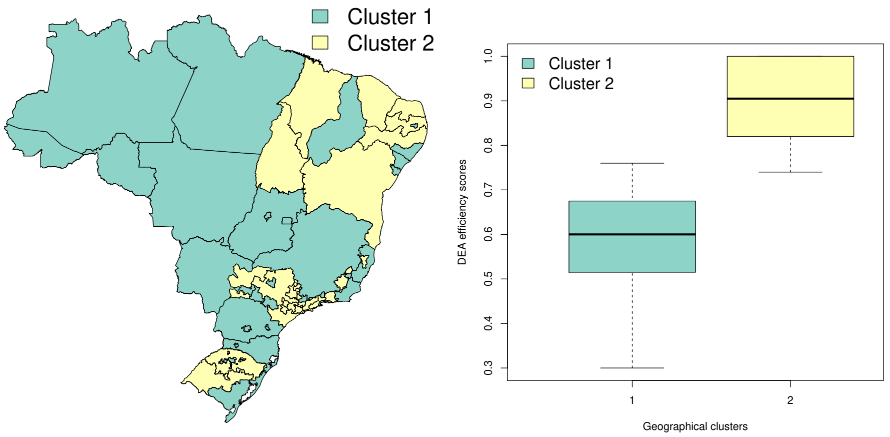


Figura 2.9 – Dendrogram using euclidean distance and hierarchical clustering

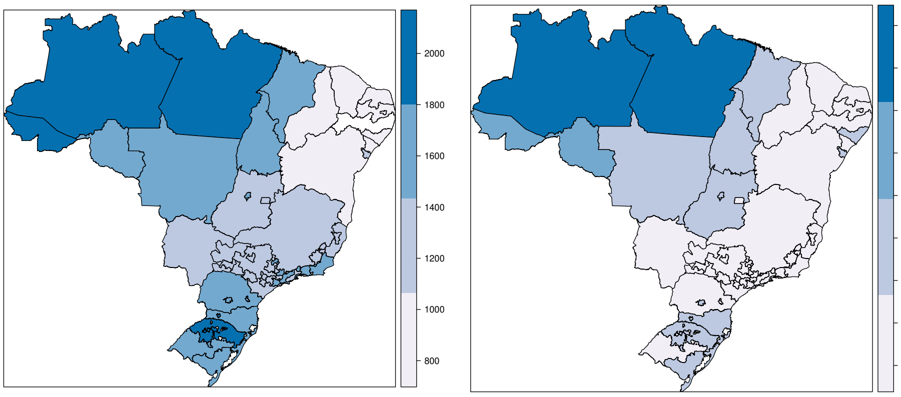


(a) Geographical location of the estimated two clusters.

(b) Boxplots of DEA efficiencies in the estimated two clusters.

Figura 2.10 – Estimated results using the k-means cluster analysis (k=2).

Figure 2.11 shows the spatial distribution of precipitation and environmental index in the Brazilian territory. Both variables were indicated in Gil at al. (GIL et al., 2017) as strongly correlated to DSO cost efficiencies. Values were grouped based on the quantile distribution. It is worth noticing the similarities among groups presented in Figure 2.11 (a) and Figure 2.11 (b) and the estimated groups using our proposed approach shown in Figure 2.6. Briefly, our proposed Bayesian model was able to aggregate DSOs with large precipitation and large environmental index. It is worth mentioning that the proposed Bayesian model do not rely on any environmental information except the geographical adjacency of DSOs. Therefore, there is evidence that the geographical adjacency is a proxy for environmental information for the Brazilian DSO database. Furthermore, the proposed clustering estimation process combines both the performance ratemaking scheme (DEA) and geographical information. This is because the proposed method uses the original estimated efficiencies, i.e., a prior performance ratemaking scheme, as the input to a secondary geographical clustering approach. Consequently, the estimated groups are homogeneous with respect to the performance ratemaking scheme and geographical location.



(a) Precipitation (mm)

(b) Environmental Index

Figura 2.11 – Maps of precipitation and environmental index grouped based on the observed quantiles.

Finally, the original reversible jump implementation proposed by Green (GREEN, 1995) includes only three moves: *birth*, *death* and *update*. The *birth* and *death* moves are related to the dimensionality changes, whereas the *update* move comprises a standard MCMC step in order to update the parameters of the model, given a fixed dimension. Although the probability of the *birth* and *death* moves can be changed arbitrarily, the same effect can be obtained by changing the prior distribution of the number of clusters. Therefore, in general, equal probabilities for *birth* and *death* moves are applied.

It is worth mentioning that the RJMCMC algorithm requires the user to set the Birth and Death *a priori* probabilities. These *a priori* probabilities are used as the frequencies in which the algorithm will propose a new configuration with one additional spatial cluster (birth) or a new configuration with a deleted spatial cluster (death). However, given the new cluster configuration, the RJMCMC will accept the new configuration with probability estimated using the *a priori* distributions and the likelihood. Therefore, the final birth and death rates may be different from the *a priori* birth and death probabilities.

Furthermore, according to Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000) the *shift* and *switch* moves were proposed in order to improve mixing performance. The authors also argue that they seem to be unnecessary. Therefore, we tested the proposed methodology assuming only three moves: *birth*, *death* and *update* with equal probabilities. We evaluated the *birth* and *death* acceptance rate, i.e., after randomly choosing a move step we evaluated the proportion of steps in which the move was definitely accepted. Further details of the RJMCMC algorithm are found in the appendix. For the proposed RJMCMC with five moves the *birth* and *death* acceptance rates are 27.27% and 33.27%, respectively. Using the RJMCMC with only three moves (*birth*, *death* and *update*), the *birth* and *death* acceptance rates are 27.53% and 33.72%, respectively. Identical results were found for the posterior distribution of the cluster size (k) and the posterior local means (μ_j). Therefore, results show that the *shift* and *switch* moves are unnecessary, as discussed by Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000).

2.5 Conclusions

Benchmarking models such as DEA and SFA aim at estimating efficiencies of DSOs using inputs, outputs and, eventually, environmental data. Proper efficient estimates require a data set of comparable DSOs, i.e., homogeneous DSOs. Heterogeneity among DSOs can be controlled using clustering analysis. Clusters of DSOs can be estimated using production information or environmental information. However, most clustering methods require the total number of clusters in advance. The larger the number of clusters the smaller the data information in each cluster.

Spatial information provides an alternative to heterogeneity analysis. It only requires the spatial location of the DSOs, and it assumes that homogeneous DSOs are geographically closer. The proposed spatial clustering analysis automatically estimates the number of clusters and the DSOs located in each cluster. The RJMCMC algorithm uses prior information of the number of clusters; i.e., expert information can be used to tune the prior information in order to improve precision. Weakly informative prior distribution can be used in order to get maximum likelihood estimates.

Using Brazilian energy distribution data, two spatial clusters were estimated. The first cluster comprises DSOs with lower efficiencies; the second cluster comprises DSOs with higher efficiencies. It may be claimed that the estimated spatial distribution is due to spatial clustering of public and private DSOs. There is also evidence of a secondary cluster with highly efficient DSOs located in the state of São Paulo. This important information can be used by the regulator to estimate future cost incentives.

Future work aims to include the SFA regression model in the RJMCMC algorithm; thus, estimating simultaneously the number of clusters, their locations and the SFA regression parameters in each cluster. Consequently, the proposed clustering method will include input, output and environmental information, as opposed to clustering efficiencies estimated using a previous benchmarking model.

2.6 Acknowledgements

The authors thank CAPES, CNPq [Grant numbers: PQ-308361/2014-8 and CNPq 402070/2016-0], FAPEMIG and FAPEMIG/CEMIG [Grant number: APQ-03165-11] for financial support.

3 A novel clustering-based spatial regression model applied to consumer power outage indicator

3.1 Introduction

The electricity distribution market in Brazil, as in most countries, operates in a natural monopoly. Consequently, end consumers cannot choose the energy distributor with low tariffs and high quality. Without proper electricity regulation, the lack of competition allows the energy distributors to charge abusive prices without improving the quality of the service.

In Brazil, the National Electricity Energy Agency (ANEEL – *Agência Nacional de Energia Elétrica*), created in 1996, is the electricity energy regulator in charge of tariff calculations, quality assessment of electricity services, among other activities related to electricity generation, transmission, distribution and commercialization (ANEEL, 1996). The distribution service comprises the delivery of electricity energy to residential and small business consumers.

The activities of an electricity distribution company, hereafter named distribution service operator (DSO), involve different and complex processes such as maintenance of the electricity assets, customer services, energy delivery, among others. Thus, the Brazilian regulator has proposed effective key performance indicators for monitoring the quality of the electricity distribution service. Among the proposed key performance indicators, the Brazilian regulator evaluates the lack of supplied energy, or a consumer power outage indicator, named DEC (*Duração Equivalente de Interrupção por Unidade Consumidora*) (ANEEL, 2016). The DEC indicator measures the average time a consumer has had electricity delivery service interrupted. In fact, the DEC indicator is calculated as the mean of the consumer power outage indicator among geographical electricity areas for a given company. Each geographical area, hereafter named electrical area, is defined by the regulator given the number of electrical assets and the number of consumers in each area, prior to calculating the DEC indicator.

Furthermore, for each electrical area, the Brazilian regulator estimates an upper bound threshold for the DEC indicator. If the observed DEC indicator surpasses this regulatory threshold, then fines are charged. In addition, the DSO must also compensate urban consumers if the power outage is greater than 2 hours, and rural consumers if the power outage is greater than 5 hours. In 2019, the Brazilian power outage compensations were estimated at R$ 617,718,741.81 (ANEEL, 2019) or US$ 150,296,530.85 considering an exchange rate of R$ 4.11/US$ 1 (December 1st, 2019).

As shown, the DEC indicator has major financial impacts. However, the DEC indicator has pros and cons. One of the main advantages is the simplicity, which makes it easy for the DSO to maintain quality control and intervene, if necessary. In contrast, the DEC indicator summarizes a complex distribution activity. Thus, it is not trivial to evaluate the financial impacts of individual management decisions on the consumer power outage indicator and, consequently, on the power outage compensations.

Based on technical evidence, environmental variables, such as precipitation, are suspected to affect the DEC indicator, as well as the size of the maintenance teams. Given limited resources, it is of utmost importance that DSOs evaluate, quantitatively, main environmental and operational drivers and their potential effects on the DEC indicator.

The use of geographical information in the analysis of Brazilian DSOs performance was first introduce by (GIL et al., 2017). Briefly, Brazil has major environmental and socioeconomic diversities mostly due to its continental dimension. Therefore, it is unlikely that only management factors affect the performance of DSOs. Nonetheless, as opposed to investigate as many environmental and socioeconomic factors as possible, a simpler alternative is to segregate the studied region into smaller geographical areas in which DSOs located in the same area are similar with respect to environmental and socioeconomic factors. The same approach can be applied to some of the Brazilian DSOs. For instance, some Brazilian DSOs have concession area larger than European countries. Thus, the concession area can be geographically divided to adjust environmental and socioeconomic heterogeneity. The estimate of geographical clusters using Brazilian DSOs was first proposed by (COSTA et al., 2019). Nonetheless, the studied aimed at identifying geographical clusters in which the mean efficient cost across the DSOs in the same cluster was similar. The number of clusters, their locations and the respective means were estimated using a Bayesian approach.

This work proposes a Bayesian clustering-based spatial regression model applied to the consumer power outage indicator. The regression model includes operational, financial and climatic variables as the independent variables. The clustering-based spatial regression allows geographical varying coefficients, which improves the prediction statistic of the model. The number and locations of spatial clusters are estimated using a Reversible-Jump Markov-Chain Monte Carlo (RJMCMC) algorithm (GREEN, 1995), inspired by epidemiological studies (KNORR-HELD; RASSER, 2000). The main motivation and the case study is the power outage indicator data from the main electricity distribution company in Brazil, named CEMIG. Results show that the proposed model achieves a predictive coefficient of determination of $R^2_{pred} = 67.6\%$, which comprises a reasonably accurate model. Based on the adjusted model, the distribution company can drive future management decisions in order to reduce both consumer energy outage and consumer compensations. To the best of our knowledge, this is the first proposal of a clustering-based spatial regression model applied to power outage indicator analysis.

This work is organized as follows. Section 3.2 presents the literature review, the Brazilian DEC indicator and the standard and Bayesian regression models. The proposed Bayesian regression model with spatial clusters, the respective algorithm, the simulation study and the Brazilian database is also presented in section 3.2. Simulation results and the Brazilian data set analysis using the proposed methodology are presented in section 3. Discussion is presented in section 3.4 and conclusion is presented in section 3.5.

3.2 Material and methods

3.2.1 Literature review

According to the Web of Science database, the term *power outage* was first used in 1970 and has appeared in 900 publications, of which 387 are papers and 449 are conference proceedings. The number of publications has grown at an average rate of 13.77%, and since 2016 has exceeded the threshold of 100 articles annually. Recent published papers show the importance of this topic. Next, selected publications based on the relevance of the theme to this work, impact factor of the journal and number of citations are briefly described.

(BEENSTOCK; GOLDIN; HAITOVSKY, 1997) present a new methodology for estimating the power outage cost in Israel. The authors use a two-limit Tobit model to estimate and simulate the economic cost of power outages. The method is based on the principle of revealed preference, using the data on investment in back-up generators to estimate the costs. They consider their model to be of more use in countries with relatively unreliable electricity systems.

(FUJITA; SHIRAI, 1997) propose a method to estimate how much power will drop after a severe generation outage. Their model aims to measure the generation outage in order to decide what proportion of the energy load will be missed following the outage. According to the authors, using dominating differential equations and simulations provides better power outage estimation than using the conventional method of second-order curve approximation.

(GUHA et al., 1999) tackle the problem of efficient recovery of an electricity system power outage following major disasters. These problems can be dealt with on two levels: the planning level, in which companies try to design more reliable and robust networks; and the operational level, in which companies try to recover their systems optimally, mainly managing the maintenance workforce. Even though the model has relevant restrictions (only the workforce resource is considered, and travel time is ignored), obtained results are satisfactory.

(MOELTNER; LAYTON, 2002) develop a model to estimate the power outage cost of firms in the U.S. using the Geweke-Hajivassiliou-Keane simulator and Halton sequences to estimate high order probabilities. Even though the model is considered better than current ones, it has restrictions regarding the specificity of the application.

(BAARSMA; HOP, 2009) analyze the Dutch energy regulatory system. The regulator uses the perceived costs of power outage as an indicator to motivate the transmission and distribution companies. The authors deal with the valuation of power-grid reliability by measuring the cost of a power outage of 2 hours for households and for small and medium enterprises. Results indicate a cost of almost 50 million euros to the Dutch society over 4 years.

(CARLSSON; MARTINSSON; AKAY, 2011) investigate willingness to pay (WTP) of the Swedish population before and after a storm hit the country in 2005. The storm caused power outages in 1/7 of Swedish households lasting from 24 hours to 3 weeks. The authors used an open-end contingent valuation with different random sample respondents. Results show a wide range of responses and, even though they cannot fully explain why, the authors propose several explanations.

(ZACHARIADIS; POULLIKKAS, 2012) study the power outage costs in Cyprus after a disaster compromised 60% of the power generating capacity of the country. The authors employed economic and engineering models to estimate the value lost by the economic sector during the outage. Results from the two proposed models are quite different, exposing the difficulties and uncertainties of such problem. Nevertheless, they consider that the emergency actions taken by the national energy authorities at the time were appropriate, even though they were not optimal.

(ANDERSEN; DALGAARD, 2013) analyze the correlation between the power outages and the economic growth in Sub-Saharan Africa between 1995 and 2007. Results indicate that a 1% increase in power outage implies a long-run reduction of the per capita GDP of 2.86%. Furthermore, if all African countries experienced the same power quality as South Africa, the per capita GDP of the continent would increase by 2%.

(MUKHERJEE; NATEGHI; HASTAK, 2018) developed a two-stage hybrid risk estimation model using data-mining techniques. The objective was to characterize the key predictors of power outages caused by weather. They used several categories of predictors, such as historical power outages, socio-economic data, climatological observations, electricity consumption patterns and land-use. Results indicate that the power outage risk depends on the type of natural hazard, the proportion of rural and urban areas and the levels of investments in operation/maintenance activities.

(REILLY; GUIKEMA, 2015) developed a tree-based statistical mass-balance Bayesian multiscale model to smooth the outage predictions. The authors allow spatially similar areas to reduce the spatial error and to yield estimates of spatial aggregation, in addition to the native model resolution. A generalized, density-based clustering algorithm is also developed. Results can improve infrastructure performance assessments, such as improved predictions for the utility operators and consumers. The model can also be applied to different spatial infrastructures (e.g., pipe breaks and road closures).

(CASTILLO, 2014) presents a literature survey of restoration strategies in response to power outages caused by hazards. The author concludes that even though there are plenty of studies focusing either on risk analysis or risk management, there are few incorporating both. One of the main reasons proposed is the lack of a unanimous approach in how to relate reliability and resiliency to market efficiency and economic loss.

(COLE et al., 2018) investigate the impact of power outages in the sales of firms from different African countries. The analysis includes firms with and without power generators. Results show a strong negative correlation between unreliable electricity supply and the sales of the firms, with stronger effects for those firms without power generators. The authors found that a reduction in the average power outage levels could increase overall sales of firms in Africa by 85.1%, going all the way up to 117.4%, for those firms without a power generator.

(BISWAS; GOEHRING, 2019) develop a model that shows a significant anti-correlation between the exponent value of the power-law outage size distribution and the load carried by the grid. Even though the results were satisfying, the authors affirm that if better outage data were available, it would be possible to draw a statistically significant map. That map would be able to mirror the health of the grid, thus allowing more effective risk management/mitigation strategies, enabling a more resilient and robust long-term power grid design.

(MORRISSEY; PLATER; DEAN, 2018) tackle the willingness to pay problem in Europe. Since the European electricity supply is considered exceptionally reliable, it is not possible to obtain data on the value of constant electricity supply. Thus, the authors propose a model to estimate the WTP in households in northwest England. Results show that the WTP changes depending on the period of the day, the weekday, the season and the duration of the power outage. The authors also used a mixed logit model to incorporate socio-demographic data, defining a price on the importance of constant electricity supply. Results can be used by both government and industry to guide future investments and policies.

(TAIMOOR et al., 2020) propose a two-stage model to estimate power outage intensity (first stage) and duration (second stage). The authors also use the model to define the three most critical cities in the U.S., considering the revenue loss due to power outage. The database contains historical power outage events, climatological annotations, socio-economic indicators and land-use data. Results indicate that the power outage interval is a function of climatological conditions, economic indicators, and time of the year.

(CARLSSON; MARTINSSON, 2007) use a contingent valuation survey to determine the willingness to pay to avoid nine different types of power outages of Swedish households. Results indicate that WTP is substantially lower as compared to the U.S., and that it increases with the duration of the outage, mainly for unplanned outages. Regarding the housing and socio-economic variables, they were not considered significant as compared to those directly related to the power outages.

Briefly, major findings in the literature are related to the prediction of the power outage economic impacts. Similarly, Brazilian DSOs face economic restrictions in their revenues if power outage levels are above regulatory levels. Thus, reliable models are required to estimate the impact of management, socioeconomic and environmental variables in the power outage levels.

3.2.2 The DEC indicator

The Brazilian regulator has applied several key performance indicators (kpi) to evaluate the quality of the services provided by the DSOs. The two main indicators named DEC and FEC (frequency of consumer power outage) evaluate the time in which customers were disconnected from the grid and the frequency of such events, respectively. Of the two indicators, the DEC is the most important.

The DEC indicator is estimated as the yearly average time of individual power outage of all consumers. The Brazilian DSOs must strive to achieve average interrupted time equal to or less than regulatory limits defined by ANEEL (ANEEL, 2016). Otherwise, fines are applied, and the operating license can even be suspended.

The DEC indicator is primarily evaluated at the electrical groups level, which comprises non-overlapping geographical areas in the concession region, defined by the regulator. Electrical groups have distinct characteristics. For instance, one electrical group may include several cities, while one city may include several electrical groups (ANEEL, 2016).

The DEC indicator is calculated using Equation 3.1,

$$DEC = \frac{\sum_{j=1}^{Cc} DIC_{(j)}}{Cc} \tag{3.1}$$

where DIC is the individual (electrical group level) interrupted time and Cc is the number of consumers in the electrical group.

However, the DEC indicator reflects a complex electrical energy activity. Mainly in a concession area larger than many European countries. Thus, effective managerial decisions based on a single indicator are, in general, naive. Consequently, such decisions may compromise investments.

Furthermore, each electrical group has its own regulatory limit. Urban, industrial and rural consumers are reimbursed differently if the power outage surpasses a time threshold. Urban and rural consumers have different compensation fees. In general, the time threshold for urban and industrial consumers are lower than for rural consumers. Thus, a complete analysis, evaluating regional factors as well as important drivers of the DEC indicator, is crucial to assess past decisions and guide future decision regarding the energy distribution quality. Consequently, if compensation fines are reduced, investments in the distribution system can be increased.

3.2.3 Multiple Linear Regression Model

The multiple linear regression model defines the relationship between the dependent variable Y and a set of k independent variables, $x_1, x_2, ..., x_k$, as follows:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i \tag{3.2}$$

where ϵ_i is a random variable following the Normal distribution with mean of zero and variance σ^2, and i is the sample index, i = 1,...,n. Using matrix notation, the model described in Equation 3.2 can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3.3}$$

Considering β_0 the intercept parameter of the model, each row in matrix $\mathbf{X}$ can be represented by a vector $\mathbf{x}_i = [1, x_{1i}, x_{2i}, ..., x_{ki}]$. Briefly, it is possible to show that the probability distribution of the vector $\mathbf{Y}$ follows a multivariate Normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix

$\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of dimension $n \times n$, that is, $\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}\right)$ (SEBER; LEE, 2012).

Assuming $\mathbf{y} = [y_1, \cdots, y_n]$ an observation sampled from the vector of random variables $\mathbf{Y}$, the maximum likelihood estimator for the parameter vector $\boldsymbol{\beta}$ is defined as $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$. The covariance matrix of the vector $\hat{\boldsymbol{\beta}}$ is also known and defined as $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (SEBER; LEE, 2012).

3.2.4 Bayesian Regression Model

In the subjective Bayesian context, the vector $\boldsymbol{\beta}$ and the variance parameter σ^2 are unknown and, thus, the prior uncertainty about their values should be expressed through *prior* probability distributions. In general, a joint *prior* distribution for the random variables $\boldsymbol{\beta}$ and σ^2 is assumed to take the form:

$$P(\boldsymbol{\beta}, \sigma^2) \propto P(\boldsymbol{\beta}|\sigma^2)P(\sigma^2). \tag{3.4}$$

In this context, $P(\boldsymbol{\beta}|\sigma^2)$ can be represented by a multivariate Normal distribution, denoted by:

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_{k+1}(\boldsymbol{\mu}_0; \sigma^2\boldsymbol{\Lambda_0}^{-1}).$$

In a weakly informative *prior* specification for $\boldsymbol{\beta}$, we set the mean and covariance matrix as $\boldsymbol{\mu_0} = \mathbf{0}$, $\boldsymbol{\Lambda_0} = \lambda_0\mathbf{I}$. Note that the lower the value of λ_0, the larger is the diagonal elements of the *prior* covariance matrix, that is, the larger is the *prior* variance for each element in $\boldsymbol{\beta}$. Similarly, a large value of λ_0 will lead to a more informative *prior* for the elements in $\boldsymbol{\beta}$.

Applying the conjugate concept in (GELMAN; others, 2006), the *prior* distribution for the variance parameter is defined by an Inverse-Gamma distribution with shape and scale parameters $a_0 > 0$ and $b_0 > 0$, respectively. Denote:

$$\sigma^2 \sim IG(a_0, b_0),$$
$$P(\sigma^2) \propto (\sigma^2)^{-a_0-1}e^{-b_0/\sigma^2}.$$

Using the Bayes Theorem, the *posterior* distribution is proportional to the likelihood and *prior* distribution product,

$$P(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \times P(\boldsymbol{\beta}, \sigma^2). \tag{3.5}$$

The application of Equation 3.5, using the mentioned *prior* probability distributions and the likelihood shown before, allows the identification of a Gaussian *posterior* distribution for $\boldsymbol{\beta}$ with the form

$$\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2 \sim \mathcal{N}_{k+1}\left(\left(\mathbf{X}^T\mathbf{X} + \lambda_0\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y};\ \ \sigma^2\left(\mathbf{X}^T\mathbf{X} + \lambda_0\mathbf{I}\right)^{-1}\right).$$

It can be noted that if $\lambda_0 = 0$, then the *posterior* distribution has similar properties as those of the maximum likelihood estimator,

$$\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\sigma^2,\lambda_0 = 0 \sim \mathcal{N}_{k+1}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y};\ \ \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

Finally, the *posterior* distribution for the variance parameter follows an Inverse-Gamma distribution denoted by $\sigma^2|\mathbf{y},\mathbf{X} \sim IG(a_n, b_n)$, with

$$a_n = a_0 + \frac{n}{2}$$
$$b_n = b_0 + \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - \boldsymbol{\mu}_n^T\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n\right)$$

where $\boldsymbol{\Lambda}_n = \left(\mathbf{X}^T\mathbf{X} + \lambda_0\mathbf{I}\right)$ and $\boldsymbol{\mu}_n = \left(\mathbf{X}^T\mathbf{X} + \lambda_0\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$. Furthermore, it is possible to show that

$$\boldsymbol{\mu}_n^T\boldsymbol{\Lambda}_n\boldsymbol{\mu}_n = \mathbf{y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda_0\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

3.2.5 Bayesian Regression Model with Spatial Clusters

A geographical area is represented by a set of n non-overlapping regions, each one with its own dependent variable Y_i and a vector of p covariates, including the 1 for the intercept, $\mathbf{x}_i = [1, x_{i1},...,x_{ip}]$ for $i = 1,...,n$. Therefore, a cluster $C_j \subset \{1,...,n\}$ is defined as a set of adjacent regions sharing the vector of coefficients $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jp})^T$. The definition of cluster implies that the groups C_1,...,C_k cover all the studied area and there is no overlap among them: $C_1 \bigcup ... \bigcup C_k = \{1,...,n\}$. Normality is assumed for the random variables Y_i. In other words, denote $Y_i \sim \mathcal{N}(\mathbf{x}_i\boldsymbol{\beta}_j,\sigma^2)$.

The number of clusters can vary between $k = 1$, where all regions are within the same cluster, and $k = n$ where each region characterizes a cluster having its own set of parameters. Defining n_j as the number of regions in each cluster C_j, $\mathbf{X}_j$ is the design matrix $n_j \times p+1$, where the rows are $\{\mathbf{x}_i : i \in C_j\}$. Assuming that the response variables, Y_i, $i = 1,...,n$, are conditionally independent given the coefficient matrix $\mathbf{B}_k = [\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_k]$, the likelihood function for the response variable vector $\mathbf{y} = (y_1,..,y_n)$ is defined by:

$$L(\mathbf{y} \mid \mathbf{X}, \mathbf{B}_k, \sigma^2) = \prod_{j=1}^{k}\prod_{i\in C_j}\frac{1}{\sigma}\phi\left(\frac{y_i - \mu_i}{\sigma}\right) \tag{3.6}$$

where $\mu_i = \mathbf{x}_i\boldsymbol{\beta}_{j_{(i)}}$ and $\phi(.)$ is the standard normal density.

The clusters model

As the first step in the definition of a configuration with k clusters, k regions g_1,...,g_k are selected as centers. Each center $g_j \in \{1,...,n\}$ defines a cluster C_j with $g_j \in C_j$. The vector of centers $G_k = (g_1,...,g_k)$ defines a clustering configuration, i.e., every region belongs to a cluster.

Furthermore, let $d(i_1, i_2)$ be the measure of distance between regions i_1 and i_2, defined as the minimum number of geographical boundaries that have to be crossed to move from i_1 to i_2. This measure can be calculated using the adjacency matrix as presented in (CRESSIE, 2015). The distance $d(i_1, i_2)$ is used to assign each area to one of the clusters. Each region i is assigned to the nearest cluster center. Nonetheless, the order of the cluster centers in vector G_k creates priority for selecting areas. For example, center g_1, which occupies the first position in vector G_k, has priority to select the nearest areas. In sequence, center g_2 has preference in selecting the remaining nearest areas, and so on. Therefore, a cluster configuration defined by vector $G_2 = (1,2)$ is, in general, different from the cluster configuration $G_2^* = (2,1)$.

To estimate the number of clusters, the regression coefficients within each cluster and the variance parameter, using the aforementioned cluster model, a reversible jump markov chain monte carlo (RJMCMC) algorithm (GREEN, 1995) is proposed.

3.2.5.1 Prior distribution for the number of clusters

As suggested by Knorr-Held and Raßer (KNORR-HELD; RASSER, 2000), the *prior* distribution for the number of clusters, $Pr(k)$, $k = 1,...,n$ is proportional to $(1-c)^k$, where the parameter $c \in [0,1)$ assumes a positive value defined by the user. A Small value of the parameter c represents non-informative *prior* distributions, whereas a large value of c indicates a *prior* distribution in which a small number of clusters is preferable.

$$P(k) \propto (1-c)^k. \tag{3.7}$$

3.2.5.2 Prior distribution for the coefficients and variance

As defined in section 3.2.4, a weakly informative Normal prior distribution is assumed for the vectors $\boldsymbol{\beta}_j$:

$$\boldsymbol{\beta}_j | \sigma^2 \sim \mathcal{N}_{p+1}(\mathbf{0}; \sigma^2(\lambda_0 \mathbf{I})^{-1}). \tag{3.8}$$

It is assumed that the vectors $\boldsymbol{\beta}_j$ are independent. For the variance parameter, the Inverse-Gamma distribution with shape $a_0 > 0$ and scale $b_0 > 0$ is used, denote:

$$\sigma^2 \sim IG(a_0, b_0).$$

The hyperparameters λ_0, a_0 and b_0 are predefined by the analyst.

3.2.6 Reversible Jump Markov Chain Monte Carlo

The proposed RJMCMC algorithm for the parameter estimation is similar to that proposed by Costa *et al.* (COSTA et al., 2019). However, the number of possible steps in the Reversible Jump Markov Chain Monte Carlo was reduced from five to three. Costa *et al.* (COSTA et al., 2019) claims that the *shift* and *switch* steps can be removed without affecting the

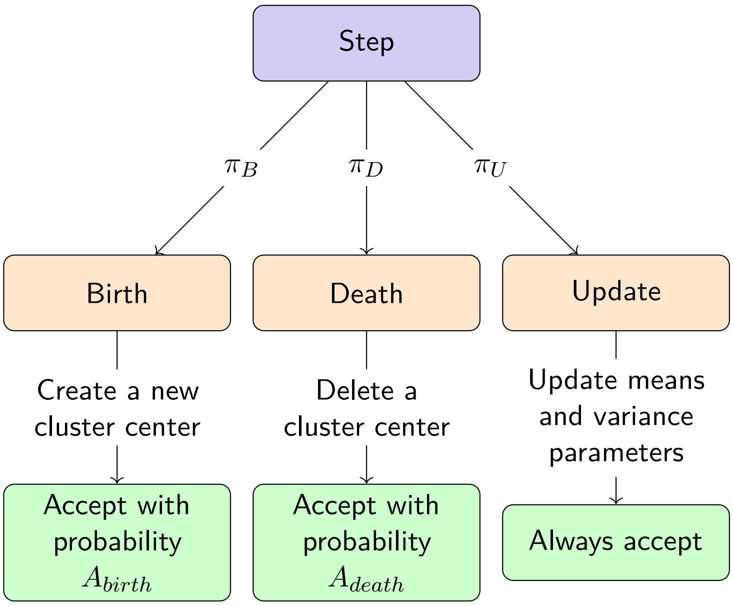

Step
π_B
π_D
π_U
Birth
Death
Update
Create a new cluster center
Delete a cluster center
Update means and variance parameters
Accept with probability A_{birth}
Accept with probability A_{death}
Always accept

Figura 3.1 – RJMCMC algorithm flowchart.

performance of the algorithm. Thus, only *birth*, *death* and *update* steps are implemented. Figure 3.1 presents a diagram of the proposed algorithm.

Similar to Costa *et al.* (COSTA et al., 2019), the proposed RJMCMC randomly chooses one of three available steps: *Birth*, *Death* or *Update* steps. In sequence, given the selected step a new configuration of the geographical partition is generated or the regression model parameters in each partition are updated. The new configurations generated using *Birth* and *Death* steps are accepted based on calculated probabilities. The algorithm is iterated using a pre-defined number of steps known as *burn-in*. After the *burn-in*, the algorithm is used to generate samples of the regression parameters and the geographical partitions. Thus, generating empirical posterior distributions. The proposed algorithm is available in the R package *gbdcd* (MINETI; COSTA, 2018). Further details about RJMCMC sampler are presented in the appendix.

3.2.6.1 Birth Step

If the birth step is selected, a new cluster configuration is created by randomly choosing a new cluster center in non-cluster center areas. This new area is added to the center vector at a random position, creating a new center vector G_{k+1}. Given the new center position r, $r \in 1,...,k+1$, a new vector of means, $\mathbf{B}_{k+1}$, is created. The mean parameter $\boldsymbol{\beta}_r$ of the new cluster center is generated from the following *posterior* multivariate normal distribution:

$$\boldsymbol{\beta}_r|\mathbf{y}_r,\mathbf{X}_r,\sigma^2 \sim \mathcal{N}_{p+1}\left(\left(\mathbf{X}_r^T\mathbf{X}_r + \lambda_0\mathbf{I}\right)^{-1}\mathbf{X}_r^T\mathbf{y}_r;\ \ \sigma^2\left(\mathbf{X}_r^T\mathbf{X}_r + \lambda_0\mathbf{I}\right)^{-1}\right). \tag{3.9}$$

The proposed distribution, $\varphi(\boldsymbol{\beta}_r)$, is the conditional distribution, that is, the *prior* distribution for $\boldsymbol{\beta}_r$, called $\varphi_0(\boldsymbol{\beta}_r)$, multiplied by the partial likelihood, that considers only the data points observed in the new cluster $(\mathbf{y}_r, \mathbf{X}_r)$, and by a normalization constant. The new cluster configuration with dimension $k+1$ is accepted with probability given by:

$$A_{birth} = \frac{L(\mathbf{y} \mid \mathbf{X}, \mathbf{B}_{k+1}, G_{k+1}, \sigma^2)}{L(\mathbf{y} \mid \mathbf{X}, \mathbf{B}_k, G_k, \sigma^2)} \cdot (1-c) \cdot \frac{\varphi_0(\boldsymbol{\beta}_r)}{\varphi(\boldsymbol{\beta}_r)}, \tag{3.10}$$

where $(1-c) = \frac{Pr(k+1)}{Pr(k)}$ is the *prior* distribution ratio of the number of clusters, penalizing steps from k to $k+1$.

If accepted, the new cluster configuration (G_{k+1} and $\mathbf{B}_{k+1}$) replaces the previous configuration (G_k and $\mathbf{B}_k$). Therefore, the RJMCMC state dimension becomes $k+1$.

3.2.6.2 Death Step

If the death step is selected, a new cluster configuration is created by randomly removing one of the current cluster centers. Thus, the cluster center g_r ($r \in$ 1,...,k) in the vector G_k is removed from the vector of centers, creating a new vector G_{k-1}. The mean parameter $\boldsymbol{\beta}_r$ associated with the selected center is also removed from the vector of means $\mathbf{B}_k$, creating a new vector $\mathbf{B}_{k-1}$. The new cluster configuration with dimension $k-1$ is accepted with probability given by:

$$A_{death} = \frac{L(\mathbf{y} \mid \mathbf{B}_{k-1}, G_{k-1}, \sigma^2)}{L(\mathbf{y} \mid \mathbf{B}_k, G_k, \sigma^2)} \cdot \frac{1}{(1-c)} \cdot \frac{\varphi(\boldsymbol{\beta}_r)}{\varphi_0(\boldsymbol{\beta}_r)}, \tag{3.11}$$

where $\frac{1}{(1-c)} = \frac{Pr(k-1)}{Pr(k)}$. If accepted, the new cluster configuration (G_{k-1} and $\mathbf{B}_{k-1}$) replaces the previous configuration (G_k and $\mathbf{B}_k$). Thus, the RJMCMC state dimension reduces to $k-1$.

3.2.6.3 Update Step

If the update step is chosen, first, only the elements of vector $\mathbf{B}_k$ are updated, without changing the dimension. Each element of vector $\mathbf{B}_k$ is updated according to Equation 3.9. Second, the variance parameter σ^2 is updated using a Gibbs sampling step, i.e., conditioned on vector $\mathbf{B}_k$ a new value for σ^2 is generated from an $IG(a_n, b_n)$ with:

$$\begin{aligned} a_n &= a_0 + \tfrac{n}{2} \\ b_n &= b_0 + \tfrac{\sum_{i=1}^{n}(y_i-\mu_i)^2}{2} \end{aligned} \tag{3.12}$$

where $\mu_i = \mathbf{x}_i\boldsymbol{\beta}_{j_{(i)}}$. Let $a_0 = 2.1$ and $b_0 = 1.1$ to indicate the prior information $E(\sigma^2) = 1$ and $V(\sigma^2) = 10$. This variance magnitude is large, suggesting high uncertainty *a priori.*

3.2.7 Estimating the location of the clusters

Using the RJMCMC samples, the locations of the clusters are estimated using the marginal frequency of pairs of electrical areas sharing geographical boundaries, as presented in

(COSTA et al., 2019) and (FENG et al., 2016). Briefly, a similarity matrix $S_{[n\times n]}$ stores the empirical probability that the electrical areas i and j are grouped in the same cluster, regardless of the estimated number of clusters. Using the Ng-Jordan-Weiss spectral clustering algorithm (NG; JORDAN; WEISS, 2002), and given a point estimate $\hat{k}$ of the number of clusters, clustering memberships for all electrical areas are calculated. Further details are found in (COSTA et al., 2019) and (FENG et al., 2016).

In general, the proposed cluster location estimate requires one multiple spatial regression model with one dependent variable and multiple independent variables. However, the proposed RJMCMC algorithm relies on the multivariate *a priori* distribution for the regression parameters, $\boldsymbol{\beta}_j$, which can be difficult to tweak if highly correlated independent variables are used. Furthermore, weakly informative prior distribution affects the acceptance rate of the algorithm. In general, the more similar the prior and the proposed distributions, the higher the acceptance rate. On the contrary, it is much easier to adjust prior distributions for univariate spatial regression models, since only two regression parameters are estimated. Furthermore, each univariate spatial regression model may indicate a different spatial cluster partition. As opposed to estimating a multivariate spatial cluster model, we propose to estimate the final spatial cluster partition by combining the results of the univariate spatial regression models, as follows. First, a final similarity matrix $S_{n\times x}$ is calculated by summing the elements of the similarity matrices for each univariate spatial model. Second, given different numbers of clusters, the respective spatial partitions of the electrical areas are generated using the spectral clustering algorithm, previously mentioned. Third, a cross-validation approach using multiple linear regression models for each cluster, using all independent variables, is applied to select the optimal number of clusters providing maximum predictive performance. The leave-one-out cross validation (FRIEDMAN; HASTIE; TIBSHIRANI, 2001) and the predictive coefficient of determination ($R^2_{prediction}$) (MONTGOMERY; PECK; VINING, 2012) is proposed to estimate the optimal number of clusters.

3.2.8 The database

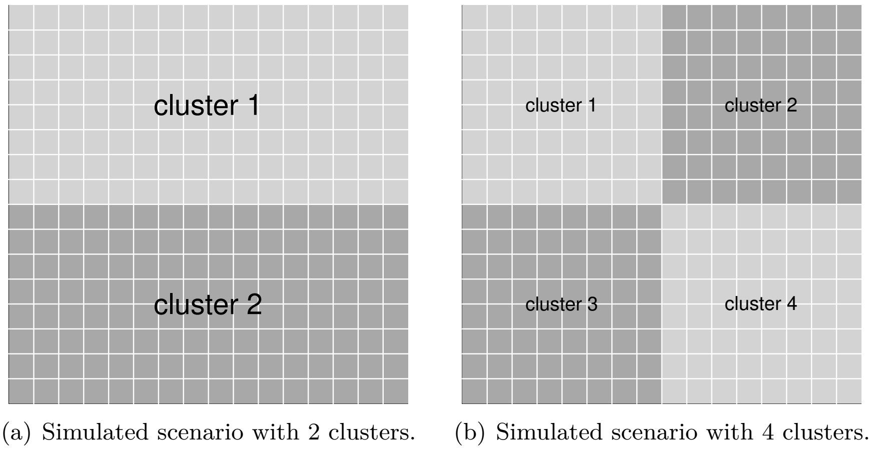
The database comprises 267 electrical areas or sub-regions of a Brazilian electricity distribution company located in southeast Brazil in the state of Minas Gerais. The power outage indicator is provided by the Brazilian electricity regulator, ANEEL (Agência Nacional de Energia Elétrica). In addition, a total of 25 predictor variables associated with each electrical area are available. These variables were originally investigated by a focus group with managers, engineers and electrical technicians from the electricity company and represent known drivers of power outage. Initially, these 25 variables were grouped into five groups: (i) geographical assets, (ii) electrical assets, (iii) demand for electrical services, (iv) climate variables and (v) operational and capital costs. Due to the high correlation among the predictor variables, a multivariate statistical analysis (dimensionality reduction) was applied. Initially, a statistical factor analysis (JOHNSON; WICHERN; others, 2002) was applied to each group, listed above, in order to represent each group by a single variable, i.e., the first principal component. Second, based on cross correlation analysis among the variables within each group, some variables were reallocated and the electrical assets and demand for electrical services groups were subdivided. Thus, the original 25 variables were divided into seven groups in which the first principal component was estimated. Consequently, the seven estimated latent variables were used as potential predictors of the power outage indicator. The proposed groups and variables within each group are shown in Table 3.1.

Tabela 3.1 – Available predictor variables and technical groups in which the first principal component is estimated.

Technical Groups	Variables within each group
Geographical assets	Service area (km^2) Extension of roads in the service area (km) Number of municipalities in the service area Number of locations served according to the electrical company definition
Electrical assets I	Extension of distribution lines (km) Extension of distribution network (km) Number of consumers
Electrical assets II	Number of substations Number of electrical protective equipment Number of automated equipment
Climate variable	Humidity index (%) Average temperature (^{o}C) Average precipitation (mm)
Demand for electrical services I	Number of working (maintenance) teams Number of commercial services Number of emergency services
Demand for electrical services II	Number of interruptions due to falling trees on the distribution lines Number of interruptions due to falling trees at substations Number of interruptions due to falling trees in the distribution network
Operational and capital costs	Operational expenditures (OPEX/R$) Capital expenditures (CAPEX/R$)

3.2.9 Simulation study

A simulation study was proposed to investigate the performance of the proposed spatial cluster regression model to detect clusters and the regression coefficients in each cluster. Simulated data was generated using a regular grid with 16 rows and 16 columns, as shown in Figure 3.2. Thus the sample size is $n = 256$. Three different scenarios with one, two and four clusters were simulated. Figures 2(a) and 2(b) show simulated scenarios with two and four clusters, respectively.



(a) Simulated scenario with 2 clusters. (b) Simulated scenario with 4 clusters.

Figura 3.2 – Simulated scenarios with 2 and 4 clusters.

The data generating process is described as follows. It is assumed that the data is generated using the following univariate linear regression equation: $Y_i = \beta \times x_i + \epsilon_i$, where ϵ_i follows a normal distribution with mean of zero and variance of σ^2. The minimum least squares estimate of β, say $\hat{\beta}$, can be written as

$$\hat{\beta} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Given the statistical data generating process above, the variance of $\hat{\beta}$ can be written as:

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_i x_i^2}$$

Using a regular grid with 256 observations, a two cluster simulation scenario (cluster A and cluster B,) assumes that in cluster A,

$$\hat{\beta}^{[A]} \sim Normal\left(\beta^{[A]}; \frac{\sigma^2}{\sum_i (x_i^{[A]})^2}\right)$$

Similarly, for cluster B,

$$\hat{\beta}^{[B]} \sim Normal\left(\beta^{[B]}; \frac{\sigma^2}{\sum_i (x_i^{[B]})^2}\right)$$

It is assumed a regular grid of size 128 for $x_i^{[A]}$ between 0 and 1. For cluster B, $x_i^{[B]} = -x_i^{[A]}$. Thus, $\sum_i (x_i^{[A]})^2 = \sum_i (x_i^{[B]})^2$. Consequently, it can be shown that the statistical distribution of the difference between $\hat{\beta}^{[A]}$ and $\hat{\beta}^{[A]}$ is written as

$$\hat{\beta}^{[A]} - \hat{\beta}^{[B]} \sim Normal\left(\beta^{[A]} - \beta^{[B]}; 2k^2\sigma^2\right) \tag{3.13}$$

where $k^2 = \frac{1}{\sum_i (x_i^{[A]})^2}$. Thus, assuming an α confidence level and the null hypothesis $H_0 : \beta^{[A]} = \beta^{[B]}$, the minimum distance between $\hat{\beta}^{[A]}$ and $\hat{\beta}^{[B]}$ that rejects the null hypothesis is

$$|\beta^{[A]} - \beta^{[B]}| \geq z_{\alpha/2} \cdot k\sigma\sqrt{2}$$

Our simulated scenario comprises the alternative hypothesis (H_a) in which $\beta^{[A]} > 0$ and $\beta^{[B]} = -\beta^{[A]}$. Thus,

$$\hat{\beta}^{[A]} - \hat{\beta}^{[B]} | H_a \sim Normal\left(\beta^{[A]} - \beta^{[B]}; 2k^2\sigma^2\right)$$

Consequently, rewritten the hypothesis testing as a one-sided test and assuming that the error type I (α) is equal to the error type II ($\gamma = \alpha$ or $z_\gamma = z_\alpha$), it can be shown that

$$\beta^{[A]} = z_\alpha \cdot k\sigma\sqrt{2}$$

where z_α is the z-score statistic. In our simulation study, the following values of z_α were used: $z_\alpha = 1.96$, $z_\alpha = 3.09$ and $z_\alpha = 4.01$. In addition, for each simulated data, the coefficient of determination (R^2), hereafter named as simulated coefficient of determination (R^2_{simul}), was calculated using Equation 3.14.

$$R^2_{simul} = \frac{\sum_{i=1}^{n} \left(\beta_{j(i)} x_i - \bar{y}\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.14}$$

where n is the sample size and β_j is the regression coefficient of cluster j. For one cluster scenario two values of β were used: $\beta = 0.45$ and $\beta = 0.10$. For four clusters simulated scenarios, the same data generating process using two clusters was applied. In sequence, each cluster data was divided into two clusters, as shown in Figure 2(b). Furthermore, the simulated coefficient of determination indicates the proportion of the simulated response y_i which is related to the regression equation. It is worth mentioning, that the statistical assumptions regarding the proposed data generating process does not include the intercept (β_0).

For each scenario, 200 simulations were evaluated using different values for the c parameter of the prior cluster distribution: $c = 0.01$ (weakly informative distribution) and $c = 0.30$ (informative distribution with a smaller prior mean). The RJMCMC algorithm was executed for 600,000 iterations using 300,000 iterations as the burn-in.

3.3 Results

3.3.1 Simulation results

Table 3.2 shows the simulated results using scenarios with one, two and four clusters; using informative ($c = 0.35$) and weakly informative ($c = 0.01$) prior distributions for the number of clusters and using different values for the simulated coefficient of determination (R^2_{simul}). In general, the larger the value of R^2_{simul} the greater the information conveyed by the regression model and the larger the detection rate of the true cluster, mainly if the informative prior distribution is applied. Furthermore, weakly informative distribution generates larger HPD intervals, as expected. For two clusters, even if lower values of R^2_{simul}, such as 2,0% or 4,0% as used, the proposed method achieves an estimated average number of clusters closer to the true value and a large proportion of simulations in which the true value is within the HPD interval. For four clusters, the weakly informative distribution achieves a larger proportion of simulations in which the true value is within the HPD interval, mainly for lower values of the R^2_{simul} statistic. Furthermore, detection rates are improved for larger values of R^2_{simul} and using the informative prior distribution. As mentioned, if weakly informative prior distributions are applied then larger HPD intervals are created and consequently, the more likely the true cluster size be found in the HPD interval. Scenarios with a lower number of clusters are more likely to be detected than scenarios with four number of clusters. This is because the larger the number of clusters the lesser the information of the local regression model, i.e., the smaller number of observations in each cluster. Thus, in order to correctly detect a large number of clusters, larger values of R^2_{simul} are required.

3.3.2 Analysis of the consumer power outage indicator

Initially, a multiple linear regression model was estimated using the seven predictor variables and the logarithm of the power outage indicator as the dependent variable. The logarithm transformation of the dependent variable was required in order to adjust the heteroscedasticity of the regression model. In addition to the estimated coefficients and the respective P-values, Table 3.4 shows the expected correlation between the dependent and each independent variable, based on technical information. Therefore, it is expected that: (i) the larger the climate variable, the larger the power outage; (ii) the larger the demand for electrical services I, the larger the power outage; (iii) the larger the demand for electrical services II, the larger the power outage; (iv) the larger the geographical assets, the larger the power outage; (v) the larger the electrical assets I, the larger the power outage; (vi) the larger the electrical assets II, the lesser the power outage; and, (vii) the larger the operational and capital costs the lesser the power outage. Results show that the expected correlation and the estimated coefficients do not match for the climate variable, demand for electrical services I, electrical assets I, and operational and capital costs. It is worth noticing that only the estimate for demand for electrical services I is not statistically significant (P-value > 0.05). The multiple linear regression model has a coefficient of determination of $R^2 = 0.5702$.

As opposed to using the proposed Bayesian spatial regression model including all seven variables, univariate models were primarily used to investigate the estimated number of clusters, Tabela 3.2 – Simulated results using scenarios with one, two and four clusters, and using informative and weakly informative prior distributions for the number of clusters.

Number of	Prior c		Average	$\bar{k}\|\mathbf{Y},\mathbf{X}$	Average	HPD
clusters	parameter	Z_α	R^2_{simul}		HPD size	proportion
1	0.01	NA	21.9%	1.26	6.48	98.0%
	0.35	NA	21.8%	1.08	3.01	100.0%
	0.01	NA	1.7%	3.6	12.7	96.0%
	0.35	NA	1.6%	1.13	3.02	99.0%
2	0.01	1.96	1.9%	4.5	16.2	97.5%
	0.35	1.96	1.9%	1.5	3.8	100.0%
	0.01	3.09	4.2%	2.1	12.5	99.5%
	0.35	3.09	4.2%	1.9	4.2	100.0%
	0.01	4.01	6.7%	2.3	11.9	98.5%
	0.35	4.01	6.8%	2.1	4.3	99.5%
4	0.01	1.96	3.4%	3.9	16.3	94.0%
	0.35	1.96	3.3%	1.4	4	74.5%
	0.01	3.09	7.4%	3.6	18.1	98.0%
	0.35	3.09	7.4%	2.5	6.3	99.0%
	0.01	4.01	11.7%	4.8	16	98.0%
	0.35	4.01	11.7%	3.5	6.3	99.5%

their locations and the spatial varying coefficients for each variable. Results are presented in Figures 3.3 to 3.9, showing the *a posteriori* distribution of the number of clusters, the most likely partitions of the electrical areas map using the mode as the point estimate, and the fitted univariate regression model for each cluster.

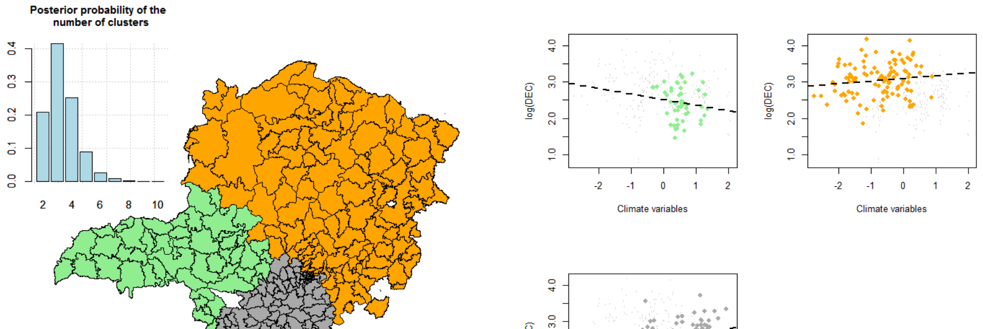
Using the climate variable, Figure 3.3 shows that three clusters were estimated. The largest cluster comprises the north region, in which low precipitation rates are generally observed. A second cluster comprises the south region, in which high precipitation rates are generally observed. The third cluster comprises the west (left) region in which high precipitation rates are also observed. The estimated coefficients for clusters located in the north and south are positive, as technically expected (see Table 3.4); i.e., the larger the climate variable, the larger the power outage. The estimated coefficient for the cluster located in the west is negative.

Tabela 3.3 – Mode of the estimated number of clusters and number of areas in each cluster sorted in increasing order.

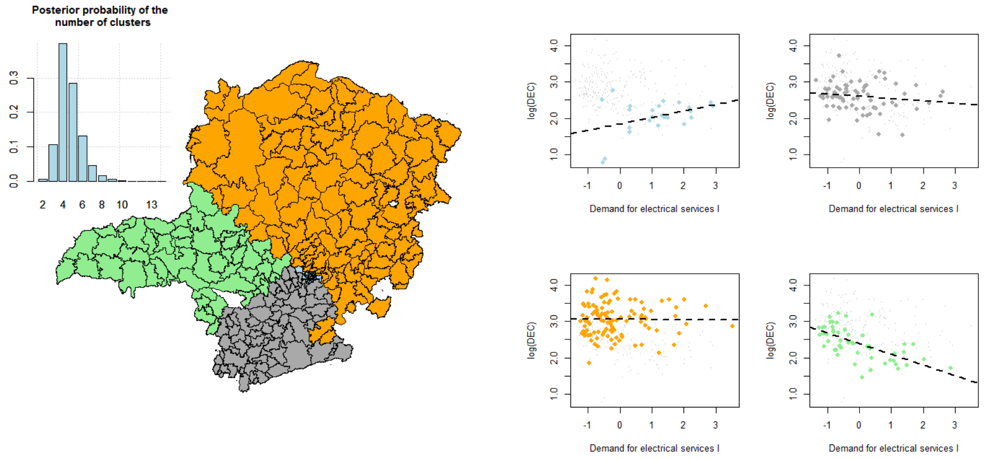
evaluated variables	Mode of the number of clusters	Number of areas in each cluster			
		cluster 1	cluster 2	cluster 3	cluster 4
Geographical assets	3				
Electrical assets I	4				
Electrical assets II	4				
Climate variable	3				
Demand for electrical services I	4				
Demand for electrical services II	4				
Operational and capital costs	4				

Tabela 3.4 – Multiple linear regression results using the estimated latent variables and the logarithm of the power outage indicator.

Predictor variable	Expected correlation	Coefficient Estimate	P-value
Intercept		6.103e-17	1.000
Climate variable	positive	-0.1831	2.49e-05
Demand for electrical services I	positive	-0.1297	0.1056
Demand for electrical services II	positive	0.8606	$<$ 2e-16
Electrical assets I	positive	-0.5087	5.36e-06
Electrical assets II	negative	-0.6563	1.96e-08
Geographical assets	positive	0.3645	1.30e-05
Operational and capital costs	negative	0.2451	0.0003



Using the demand for electrical services I, Figure 3.4 shows that four clusters were estimated. The largest cluster comprises the north region. The second cluster comprises the south region. The third cluster comprises the west (left) region and the fourth cluster, a small cluster, comprises electrical areas in the state capital. The estimated coefficient for the cluster located in the north is close to zero. The estimated coefficients for clusters located in the south and west are negative and the estimated coefficient for the small cluster located in the state capital is positive. The expected correlation is positive (see Table 3.4).

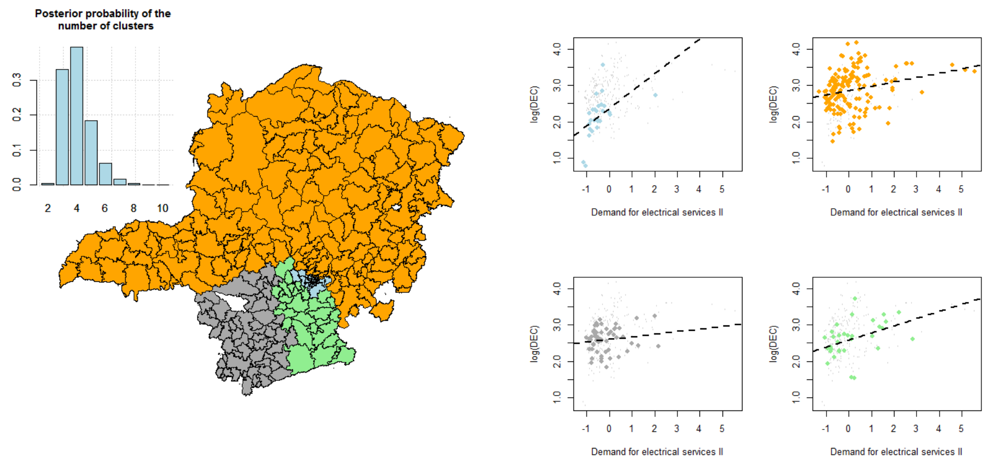


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figura 3.4 – Results using the univariate Bayesian spatial regression model and the demand for electrical services I as the predictor.

Using the demand for electrical services II, Figure 3.5 shows that four clusters were estimated. The largest cluster comprises the north and west regions. The second cluster comprises the southwest region. The third cluster comprises the southeast (bottom right) region and the fourth cluster, the smallest cluster, comprises electrical areas in the state capital and surrounding areas. The estimated coefficients for all clusters are positive, as technically expected. However, the values of the coefficients vary among the clusters showing that in some clusters, such as the smallest cluster and the cluster located in the southeast, the correlation between demand for electrical services II and the power outage is larger as compared to the remaining clusters.

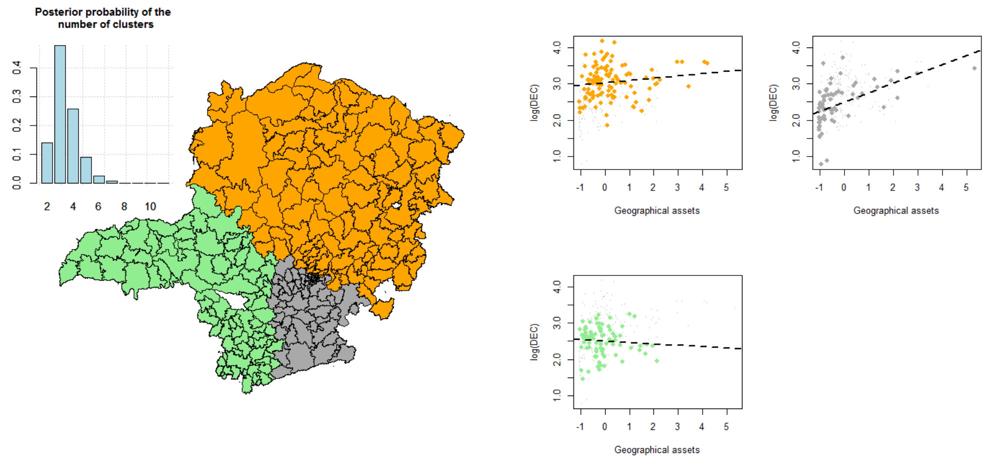


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figura 3.5 – Results using the univariate Bayesian spatial regression model and the demand for electrical services II as the predictor.

Using the geographical assets, Figure 3.6 shows that three clusters were estimated. The largest cluster comprises the north region. A second cluster comprises the southeast (bottom right) region. The third cluster comprises the west and southwest regions. The estimated coefficients for clusters located in the north and southeast regions are positive, as technically expected (see Table 3.4); i.e., the larger the geographical assets, the larger the power outage. The estimated coefficient for the cluster located in the west/southwest region is slightly negative.

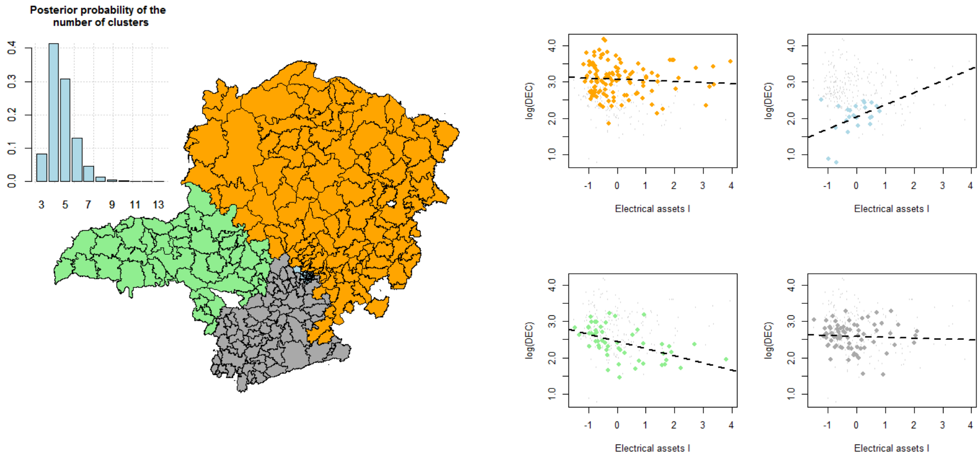


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figura 3.6 – Results using the univariate Bayesian spatial regression model and the geographical assets as the predictor.

Using the electrical assets I, Figure 3.7 shows that four clusters were estimated. Results are similar to those findings using the electrical services I. The largest cluster comprises the north region. The second cluster comprises the south region. The third cluster comprises the west (left) region and the fourth cluster, the smallest cluster, comprises electrical areas in the state capital. The estimated coefficients for the clusters located in the north and south regions are close to zero. The estimated coefficient for the cluster located in the west is negative, and the estimated coefficient for the small cluster located in the state capital is positive. The expected correlation is positive (see Table 3.4).

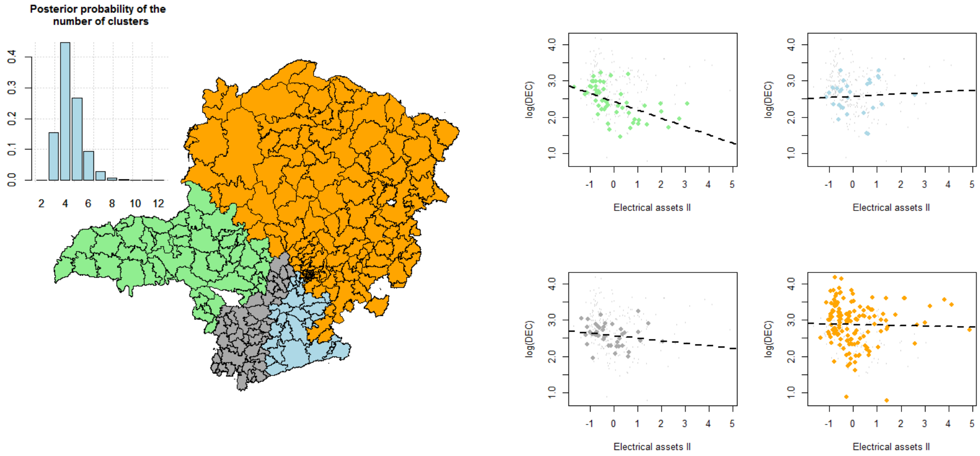


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figura 3.7 – Results using the univariate Bayesian spatial regression model and the electrical assets I as the predictor.

Using the demand for electrical services II, Figure 3.8 shows that four clusters were estimated. The largest cluster comprises the north region. The second cluster comprises the northwest region. The third cluster comprises the south (bottom right) region, and the fourth cluster, the smallest cluster, comprises electrical areas in the state capital and surrounding areas. The estimated coefficients for clusters located in the west and close to the state capital are negative, as technically expected. The estimated coefficients for the remaining clusters are close to zero.

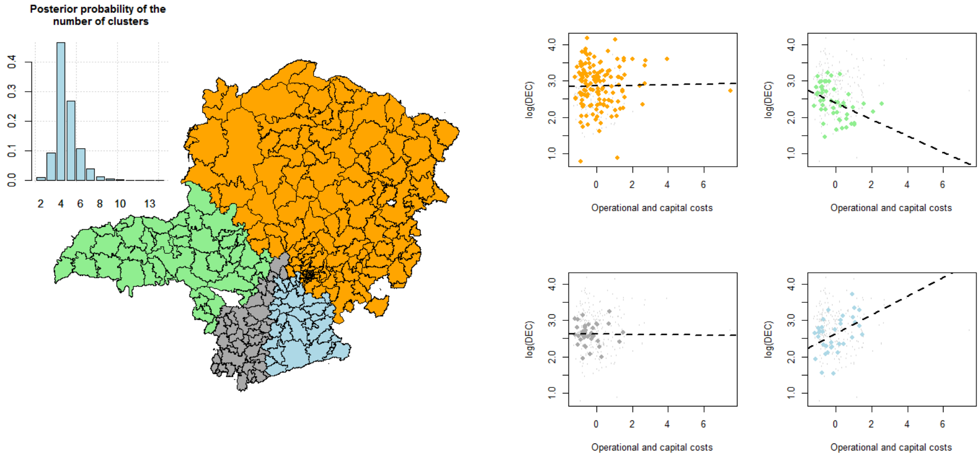


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figura 3.8 – Results using the univariate Bayesian spatial regression model and the electrical assets II as the predictor.

Using the operational and capital costs, Figure 3.9 shows that four clusters were estimated. The largest cluster comprises the north region. The second cluster comprises the west (left) region. The third cluster comprises the southwest region, and the fourth cluster comprises the southeast region. The estimated coefficients are close to zero for clusters located in the north and west regions. The estimated coefficient in the west region is negative, as technically expected. The estimated coefficient in the southeast cluster is positive indicating that, in this region, the larger the operational and capital costs, the larger the power outage.



(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

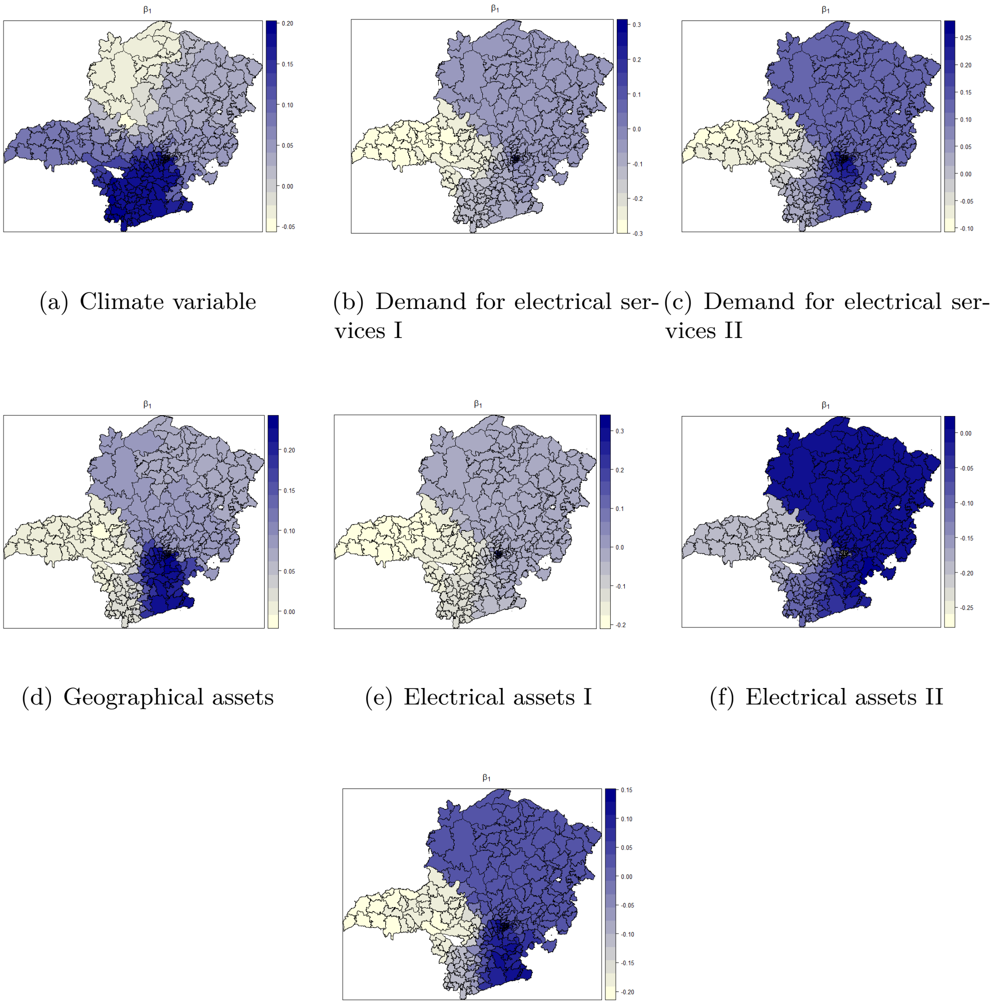
(b) Univariate regression models estimated for each cluster.

Figura 3.9 – Results using the univariate Bayesian spatial regression model and the operational and capital costs as the predictor.

A close look at the estimated clusters for each variable shows similar patterns. In general, all univariate models indicate a large cluster located in the north region. A second cluster is located in the west region. Some univariate models indicate a third cluster located either in south/southeast or southwest region, and some univariate models indicate a fourth cluster, with a smaller number of electrical areas, located closer to the state capital. Interestingly, by dividing the data into spatial clusters, some of the estimated coefficients had their signs changed. For example, without any spatial partition, the regression model presents a negative and statistically significant coefficient for the climate variable, as shown in Table 3.4. By dividing the data into spatial clusters, the estimated coefficients within some clusters are positive, as technically expected. These results indicate a concept known as the Simpson's paradox, or reversal paradox (SIMPSON, 1951), in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. The reversal paradox can also be partially observed for demand for electrical services I, electrical assets I, and operational and capital costs. These findings indicate that the spatial partition is an important variable in the model.

Figure 3.10 shows the spatial estimate of the regression coefficient (β_1) for each electrical area and predictor variable. Results indicate spatial locations in which the regression equation is more pronounced. Values close to zero indicate a zero slope, i.e, the absence of the regression effect.

Figure 3.12 illustrates the Bayesian univariate spatial regression results using a simulated scenario with no spatial clusters, i.e., only one cluster. In this case, the proposed model correctly identified a single partition on the map and that the spatial estimates of the regression coefficient (β_1) are homogeneous.



(a) Climate variable

(b) Demand for electrical services I

(c) Demand for electrical services II

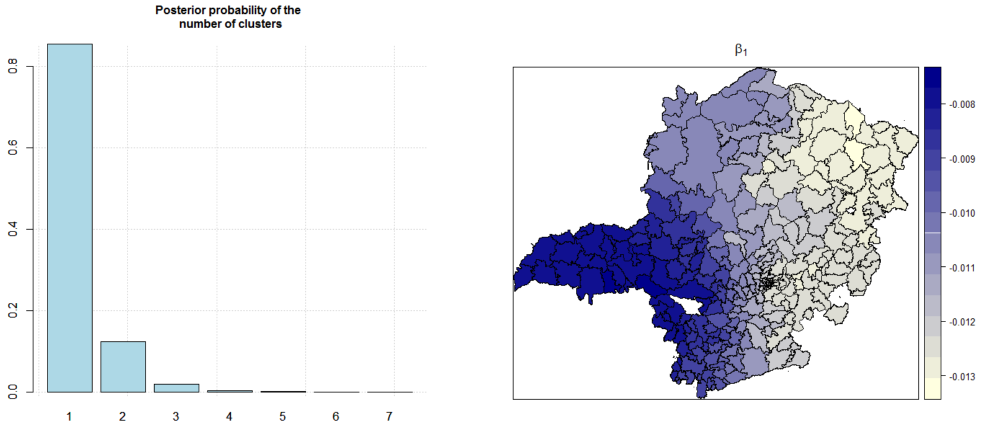
(d) Geographical assets

(e) Electrical assets I

(f) Electrical assets II

(g) Operational and capital costs

Figura 3.10 – Spatial estimate of the regression coefficient β_1 for each electrical area.



(a) The *a posteriori* distribution of the number of clusters.

(b) Spatial estimate of the regression coefficient β_1.

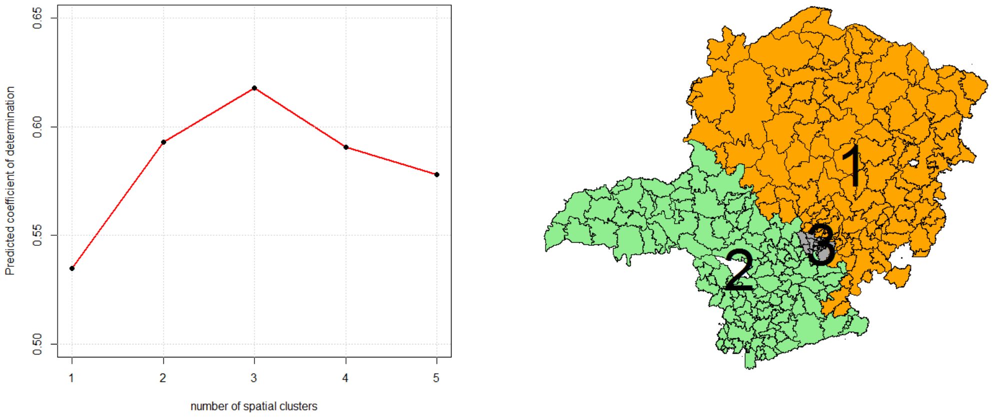
Figura 3.11 – Results using the univariate Bayesian spatial regression model and a simulated scenario with no clusters.

Table 3.5 shows the *birth* and *death* acceptance rates for models with different number of independent variables (predictors) and different values for the parameter c, which tunes the *a priori* distribution of the number of clusters. The larger the value of c, the more informative the *a priori* distribution with probability mass towards smaller number of clusters. Whereas lower values of c comprise weakly informative distributions, i.e., a flat *a priori* distribution. Results show that the more independent variables are included in the model the lower the *birth* acceptance rate. Furthermore, a weakly informative *a priori* distribution for the number of clusters achieves larger acceptance rates as compared to more informative *a priori* distribution. Using the complete number of independent variables, i.e., seven predictor variables, the *birth* acceptance rate is the lowest wheres the *death* acceptance rate is large. Consequently, the *a posteriori* distribution of the number of clusters has a probability mass towards the lowest value, which is one spatial cluster. Future studies aim at proposing different *a priori* distributions for vector $\boldsymbol{\beta}_j$, which can improve the *birth* acceptance rates. Therefore, the *a priori* distribution shown in Equation 3.8 has limitations if the number of independent variables is large. Alternatively, one may combine the univariate spatial regression results into a multiple spatial regression analysis as described below.

For each univariate spatial regression model, the clustering algorithm was applied varying the number of clusters from 1 to 5. For each cluster configuration, a multiple linear regression model using all predictor variables was adjusted for each partition. Finally, the predictive coefficient of determination (R^2_{pred}) (MONTGOMERY; PECK; VINING, 2012) was calculated and the cluster configuration achieving the maximum value of R^2_{pred} was selected. Figure 12(a) shows the R^2_{pred} values using different number of clusters. Results show that the maximum value of $R^2_{pred} = 61.79\%$ is achieved using three clusters. Figure 12(b) shows the best configuration of clusters with one cluster comprising the north region, one cluster comprising the south and west regions, and one cluster comprising electrical areas located in the state capital and surrounding areas.

Tabela 3.5 – Birth and death acceptance rates for models with different number of independent variables and varying *a priori* parameter c.

number of	$c = 0.35$		$c = 0.001$	
predictors	birth	death	birth	death
1	7.4%	7.4%	9.3%	9.3%
3	2.4%	2.4%	2.8%	2.8%
4	2.3%	2.3%	3.1%	3.1%
7	0.9%	6.3%	0.9%	8.7%



(a) Predictive coefficient of determination (R^2_{pred}) for different number of clusters.

(b) Electrical areas divided into 3 spatial clusters.

Figura 3.12 – Final selection of the number of clusters based on the predictive coefficient of determination for different number of clusters.

Table 3.6 shows the expected sign, the univariate coefficient estimates and the multivariate coefficient estimates using the data from the first cluster, i.e., using the electrical areas in the north region. Only three predictor variables were statistically significant (P-value < 0.05). The demand for electrical assets variable presented a positive coefficient, as expected. On the contrary, the electrical assets I variable presented a negative coefficient for both univariate and multiple linear regression models. The Variance Inflation Factor (VIF) statistic was large for electrical assets II, indicating multicollinearity.

Table 3.7 shows results using data from the second cluster, located in the south and west regions. In this second cluster, three predictor variables were statistically significant (P-value < 0.05). The demand for electrical services II and de geographical assets variables presented positive coefficients, as expected. On the contrary, the electrical assets I variable presented a negative coefficient. Similarly to results found in cluster one, the VIF statistic presented a large

Tabela 3.6 – Estimated univariate and multiple liner regression coefficients for the spatial cluster 1.

Predictor	Expected	Cluster 1			
variable	sign	univariate	coefficient	P-value	VIF
Climate variable	positive	0.0671	-0.0186	0.6840	1.49
Demand for electrical services I	positive	-0.0068	-0.0027	0.9764	6.21
Demand for electrical services II	positive	0.1118	0.3720	0.0000	3.93
Electrical assets I	positive	-0.0265	-0.3589	0.0002	9.04
Electrical assets II	negative	0.0059	-0.1601	0.1654	13.38
Geographical assets	positive	0.0531	0.0993	0.1090	4.40
Operational and capital costs	negative	0.0359	0.0980	0.0778	2.70
Sample size: 115 electrical areas					

value for electrical assets II.

Tabela 3.7 – Estimated univariate and multiple liner regression coefficients for the spatial cluster 2.

Predictor	Expected	Cluster 2			
variable	sign	univariate	coefficient	P-value	VIF
Climate variable	positive	0.0266	-0.1029	0.0651	1.18
Demand for electrical services I	positive	-0.1806	-0.1931	0.0274	9.43
Demand for electrical services II	positive	0.0322	0.4039	0.0000	5.19
Electrical assets I	positive	-0.1479	-0.2470	0.0019	8.68
Electrical assets II	negative	-0.1412	-0.1649	0.0912	12.30
Geographical assets	positive	0.0811	0.2140	0.0008	3.81
Operational and capital costs	negative	-0.0882	0.0398	0.5334	3.50
Sample size: 120 electrical areas					

Table 3.8 shows results using data from the third cluster, the smallest cluster, located in the in the state capital and surrounding areas. Despite the small sample size, three statistically significant variables are found. The demand for electrical services II variable presents positive coefficient as expected. The electrical assets II variable presents a negative coefficient, as expected. The geographical assets variable presents a negative coefficient in the multiple linear regression model, even though the univariate model estimated a positive coefficient (as expected). The electrical assets I and geographical assets presented larger values of the VIF statistics.

Tabela 3.8 – Estimated univariate and multiple liner regression coefficients for the spatial cluster 3.

Predictor	Expected	Cluster 3			
variable	sign	univariate	coefficient	P-value	VIF
Climate variable	positive	0.6781	0.4922	0.0795	1.54
Demand for electrical services I	positive	0.1044	-0.1770	0.1922	8.61
Demand for electrical services II	positive	0.3971	1.4204	0.0000	9.26
Electrical assets I	positive	0.1979	0.1284	0.6455	15.02
Electrical assets II	negative	-0.0022	-0.3788	0.0034	6.30
Geographical assets	positive	0.3184	-1.3093	0.0105	10.88
Operational and capital costs	negative	0.0942	0.1127	0.1199	4.93
Sample size: 31 electrical areas					

Results shown in Tables 3.6, 3.7 and 3.8 provide evidence that the predictive performance of the available variables with respect to the power outage indicator varies geographically. Thus, models using different variables are required in order to improve the predictive performance of the DEC indicator.

3.4 Discussion

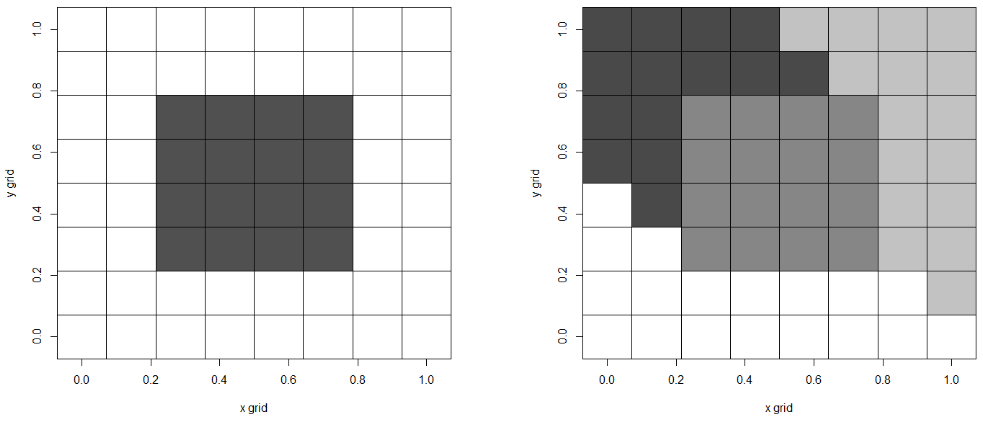
As previously mentioned, Brazil has continental dimensions and some of the Brazilian DSOs has concession areas larger than many european countries. Consequently, the electrical distribution service faces many challenges related to weather, vegetation and socioeconomic factors. In the case study, the geographical heterogeneity was technically known for engineers and management staff. Nonetheless, providing a proper treatment was a difficult tasks since standard statistical analysis do not rely on simultaneous estimation of spatial clusters and regression coefficients.

The estimated number of cluster as three and their respective locations do show consistent result, as expected by experts. The cluster located in the north (cluster one) comprises a drier region with little precipitation and old assets. The second cluster located in the west and south regions is mostly related to agricultural production. Large agricultural industries are located in the west whereas the precipitation index is larger in the south. Nonetheless, the variables associated with the electrical assets were also statistically significant but with different coefficients. Finally, cluster three comprises a highly industrialized and populated electrical areas. All detected clusters indicated the strong correlation between the power outage indicator and the variables associated with the electrical assets.

Figure 3.13 illustrates one limitation of the proposed clustering-based Bayesian spatial model. The original spatial partition algorithm, as proposed by Knorr-Held and Raßer (KNORR-

HELD; RASSER, 2000), may overestimate the number of clusters if their locations do not fit the spatial partition algorithm. For instance, Figure 13(a) shows a simulated scenario with two clusters in which one cluster is located in the center of the study region. The original spatial Bayesian algorithm does not allow such partition. Nevertheless, the central cluster is detected by creating additional clusters, as shown in Figure 13(b). Consequently, the algorithm overestimates the number of clusters, but the estimate of the regression parameters between the outer clusters are similar. Thus, future studies aim at developing more flexible spatial partition algorithms.

In addition, as shown in the case study, the more predictor variables are included in the spatial regression model, the lower the *birth* acceptance rate. Consequently, the proposed Bayesian spatial regression model may not detect any spatial partition. An alternative is to adjust univariate spatial regression models. In sequence, combine the univariate spatial partition information and use multiple regression models and cross-validation analysis to find the optimal number of partitions. Results using the power outage data suggest that this approach may overcome the lower acceptance rates and the spatial clustering limitation, previously mentioned.



(a) Simulated scenario with two spatial clusters.

(b) Detected clusters.

Figura 3.13 – Overestimation of the number of clusters using the spatial Bayesian approach. In order to detect the central cluster, two additional clusters are created.

3.5 Conclusion

The unexpected failure of electrical energy supply generates major production and financial losses to industrial, local market and residential consumers. In general, main causes of power outage can be attributed to both managerial and non-managerial factors. Precipitation, lightning, wind gusts are known environmental factors related to power outage. Likewise, socioeconomic factors may also affect the electricity supply, mainly in vulnerable socioeconomic areas. Thus, DSOs with large concession areas have a difficult task to evaluate the different factors, as well as their different impacts, in the power outage behavior across the concession region. Consequently, adjusting statistical regression models to geographical clusters captures the geographical heterogeneity with respect to both managerial and non-managerial factors. However, reliable estimates of the number of geographical clusters, their respective locations and the local regression coefficients is overwhelming and can be overcome using the proposed statistical Bayesian approach.

This work has successfully proposed a spatial Bayesian linear regression model which estimates spatial clusters and the respective regression coefficients. The main motivation and the case study is the prediction of the power outage indicator, named DEC, in the largest Brazilian DSO located in the southeast region. Results provide strong statistical evidence that the proposed geographical clustering approach improves the predictive accuracy of the DEC indicator. Briefly, three geographical clusters were estimated. Most important drivers are related to electrical and geographical assets. Secondary drivers are related to climate variables, operational and capital costs and demand for electrical services. Furthermore, the estimated effects of the drivers, i.e., the regression coefficients, do vary among the different clusters. The estimated coefficients of the models can drive future management decisions to reduce the DEC indicator and, consequently, reduce compensation paid to consumers. Thus, the studied DSO can increase future investments in network expansion and quality of the services.

Acknowledgements

The authors thank CEMIG-D (grant number: CEMIG-D 0636/2018) and CNPq (grant number 303119/2019-5) for financial support.

Referências

AFRIAT, S. N. Efficiency estimation of production functions. *International Economic Review*, p. 568–598, 1972. Publisher: JSTOR. Citado na página 13.

AGRELL, P. J. et al. Unobserved Heterogeneous Effects in the Cost Efficiency Analysis of Electricity Distribution Systems. In: *The Interrelationship Between Financial and Energy Markets*. [S.l.]: Springer, 2014. p. 281–302. Citado na página 13.

AIGNER, D.; LOVELL, C. K.; SCHMIDT, P. Formulation and estimation of stochastic frontier production function models. *journal of Econometrics*, v. 6, n. 1, p. 21–37, 1977. Publisher: Elsevier. Citado 2 vezes nas páginas 11 e 13.

ANDERSEN, T. B.; DALGAARD, C.-J. Power outages and economic growth in Africa. *Energy Economics*, v. 38, p. 19–23, 2013. Publisher: Elsevier. Citado na página 33.

ANEEL. *Agência Nacional de Energia Elétrica*. 1996. Published: aneel.gov.br. Citado 2 vezes nas páginas 8 e 30.

ANEEL. *Metodologia de Custos Operacionais*. [S.l.], 2015. Citado na página 19.

ANEEL. *Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST*. [S.l.], 2016. Disponível em: <https://www.aneel.gov.br/prodist>. Citado na página 34.

ANEEL. *Procedimentos de Regulação Tarifária - PRORET - ANEEL*. 2016. Disponível em: <http://www.aneel.gov.br/procedimentos-de-regulacao-tarifaria-proret>. Citado 2 vezes nas páginas 9 e 30.

ANEEL. *Qualidade do Serviço*. [S.l.], 2016. Disponível em: <https://www.aneel.gov.br/qualidade-do-servico2>. Citado na página 35.

ANEEL. *Painel de Desempenho das Distribuidoras de Energia Elétrica*. 2019. Published: www.aneel.gov.br/painel-de-desempenho. Citado 2 vezes nas páginas 9 e 30.

BAARSMA, B. E.; HOP, J. P. Pricing power outages in the Netherlands. *Energy*, v. 34, n. 9, p. 1378–1386, 2009. Publisher: Elsevier. Citado na página 32.

BANKER, R. D.; CHARNES, A.; COOPER, W. W. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, v. 30, n. 9, p. 1078–1092, 1984. Publisher: INFORMS. Citado na página 11.

BANKER, R. D.; MOREY, R. C. Efficiency analysis for exogenously fixed inputs and outputs. *Operations research*, v. 34, n. 4, p. 513–521, 1986. Publisher: INFORMS. Citado na página 14.

BEENSTOCK, M.; GOLDIN, E.; HAITOVSKY, Y. The cost of power outages in the business and public sectors in Israel: revealed preference vs. subjective valuation. *The Energy Journal*, v. 18, n. 2, 1997. Publisher: International Association for Energy Economics. Citado na página 32.

BISWAS, S.; GOEHRING, L. Load dependence of power outage statistics. *EPL (Europhysics Letters)*, v. 126, n. 4, p. 44002, 2019. Publisher: IOP Publishing. Citado na página 33.

BOGETOFT, P. *Comments on the Brazilian benchmarking model for energy distribution regulation. Forth cycle of tariff review, NT 192 2014*. [S.l.], 2014. Citado na página 11.

BOGETOFT, P.; LOPES, A. L. M. *Comments on the Brazilian Benchmarking model for energy distribution regulation Forth cycle Tariff review Technical Note 407/2014*. [S.l.], 2015. Citado na página 11.

BOGETOFT, P.; OTTO, L. *Benchmarking with DEA, SFA, and R*. New York: Springer, 2011. (International series in operations research &amp; management science, v. 157). ISBN 978-1-4419-7961-2. Citado na página 11.

BROECK, J. Van den et al. Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics*, v. 61, n. 2, p. 273–303, 1994. Publisher: Elsevier. Citado na página 13.

BROOKS, S. P.; GIUDICI, P.; ROBERTS, G. O. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v. 65, n. 1, p. 3–39, 2003. Publisher: Wiley Online Library. Citado na página 14.

CARLSSON, F.; MARTINSSON, P. Willingness to pay among Swedish households to avoid power outages: a random parameter Tobit model approach. *The Energy Journal*, v. 28, n. 1, 2007. Publisher: International Association for Energy Economics. Citado na página 34.

CARLSSON, F.; MARTINSSON, P.; AKAY, A. The effect of power outages and cheap talk on willingness to pay to reduce outages. *Energy Economics*, v. 33, n. 5, p. 790–798, 2011. Publisher: Elsevier. Citado na página 32.

CASTILLO, A. Risk analysis and management in power outage and restoration: A literature survey. *Electric Power Systems Research*, v. 107, p. 9–15, 2014. Publisher: Elsevier. Citado na página 33.

CHARNES, A.; COOPER, W.; RHODES, E. Measuring the efficiency of decision making units. *European Journal of Operational Research*, v. 2, n. 6, p. 429–444, nov. 1978. ISSN 03772217. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/0377221778901388>. Citado na página 11.

CHEN, M.-H.; SHAO, Q.-M. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, v. 8, n. 1, p. 69–92, 1999. Publisher: Taylor & Francis Group. Citado 2 vezes nas páginas 14 e 21.

CLAYTON, D. G.; BERNARDINELLI, L. Bayesian methods for mapping disease risk. *Geographical and environmental epidemiology: methods for small-area studies*, p. 205–220, 1992. Publisher: Oxford Univ Press. Citado na página 14.

COLE, M. A. et al. Power outages and firm performance in Sub-Saharan Africa. *Journal of Development Economics*, v. 134, p. 150–159, 2018. Publisher: Elsevier. Citado na página 33.

COOK, W.; ZHU, J. *Data Envelopment Analysis: balanced benchmarking*. [S.l.]: Cook and Zhu, San Bernardino, 2013. Citado na página 19.

COOPER, W. W.; SEIFORD, L. M.; ZHU, J. Data envelopment analysis. In: *Handbook on data envelopment analysis*. [S.l.]: Springer, 2004. p. 1–39. Citado na página 19.

COSTA, M. A. et al. Bayesian detection of clusters in efficiency score maps: An application to Brazilian energy regulation. *Applied Mathematical Modelling*, v. 68, p. 66–81, abr. 2019. ISSN 0307-904X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0307904X18305365>. Citado 4 vezes nas páginas 31, 38, 39 e 41.

CRESSIE, N. *Statistics for spatial data*. [S.l.]: John Wiley & Sons, 2015. Citado 2 vezes nas páginas 15 e 38.

DAI, X.; KUOSMANEN, T. Best-practice benchmarking using clustering methods: Application to energy regulation. *Omega*, v. 42, n. 1, p. 179–188, 2014. Publisher: Elsevier. Citado na página 13.

ELAVARASI, S. A.; AKILANDESWARI, J.; SATHIYABHAMA, B. A survey on partition clustering algorithms. *International Journal of Enterprise Computing and Business Systems*, v. 1, n. 1, p. 1–14, 2011. Citado na página 22.

FARRELL, M. J. The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, v. 120, n. 3, p. 253–290, 1957. Publisher: JSTOR. Citado na página 11.

FENG, W. et al. Spatial regression and estimation of disease risks: A clustering-based approach: Spatial Regression and Estimation of Disease Risks. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, v. 9, n. 6, p. 417–434, dez. 2016. ISSN 19321864. Disponível em: <http://doi.wiley.com/10.1002/sam.11314>. Citado 3 vezes nas páginas 21, 22 e 41.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. Citado na página 41.

FUJITA, G.; SHIRAI, G. Estimation of power outage size based on the dominating differential equation. *Electrical engineering in Japan*, v. 118, n. 3, p. 39–49, 1997. Publisher: Wiley Online Library. Citado na página 32.

GELMAN, A. *Bayesian data analysis*. Third edition. Boca Raton: CRC Press, 2014. (Chapman & Hall/CRC texts in statistical science). ISBN 978-1-4398-4095-5. Citado na página 14.

GELMAN, A.; others. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, v. 1, n. 3, p. 515–534, 2006. Publisher: International Society for Bayesian Analysis. Citado na página 36.

GIL, G. D. R. et al. Spatial statistical methods applied to the 2015 Brazilian energy distribution benchmarking model: Accounting for unobserved determinants of inefficiencies. *Energy Economics*, v. 64, n. Supplement C, p. 373–383, maio 2017. ISSN 0140-9883. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0140988317301160>. Citado 7 vezes nas páginas 10, 12, 13, 19, 25, 27 e 31.

GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, v. 82, p. 711–732, 1995. Citado 5 vezes nas páginas 9, 14, 28, 31 e 38.

GUHA, S. et al. Efficient recovery from power outage. In: *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. [S.l.: s.n.], 1999. p. 574–582. Citado na página 32.

HEATON, M. J.; CHRISTENSEN, W. F.; TERRES, M. A. Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics*, v. 59, n. 1, p. 93–101, 2017. Publisher: Taylor & Francis. Citado na página 26.

JOHNSON, R. A.; WICHERN, D. W.; others. *Applied multivariate statistical analysis*. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002. v. 5. Citado 2 vezes nas páginas 25 e 41.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009. v. 344. Citado na página 25.

KNORR-HELD, L.; RASSER, G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, v. 56, n. 1, p. 13–21, mar. 2000. ISSN 0006-341X. Citado 8 vezes nas páginas 8, 14, 15, 28, 29, 31, 38 e 59.

KUOSMANEN, T. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics*, v. 34, n. 6, p. 2189–2199, 2012. Publisher: Elsevier. Citado 2 vezes nas páginas 11 e 13.

LAZAR, J. *Electricity regulation in the US: A guide*. [S.l.]: Regulatory Assistance Project, 2011. Citado na página 11.

LLORCA, M.; OREA, L.; POLLITT, M. G. Using the latent class approach to cluster firms in benchmarking: An application to the US electricity transmission industry. *Operations Research Perspectives*, v. 1, n. 1, p. 6–17, 2014. Publisher: Elsevier. Citado na página 13.

LOPES, A. L. M. et al. Critical evaluation of the performance assessment model of Brazilian electricity distribution companies. *Revista Gestão & Tecnologia*, v. 16, n. 3, p. 5–30, 2016. Citado 2 vezes nas páginas 11 e 19.

MCLACHLAN, G. J.; BASFORD, K. E. *Mixture models: Inference and applications to clustering*. [S.l.]: Marcel Dekker, 1988. v. 84. Citado na página 13.

MINETI, L.; COSTA, M. *leandromineti/gbdcd: An R package implementing the Bayesian Detection of Clusters and Discontinuities*. 2018. Disponível em: <https://doi.org/10.5281/zenodo.1291501>. Citado 2 vezes nas páginas 19 e 39.

MOELTNER, K.; LAYTON, D. F. A censored random coefficients model for pooled survey data with application to the estimation of power outage costs. *Review of Economics and Statistics*, v. 84, n. 3, p. 552–561, 2002. Publisher: MIT Press. Citado na página 32.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. 5th ed. ed. Hoboken, NJ: Wiley, 2012. (Wiley series in probability and statistics, 821). ISBN 978-0-470-54281-1. Citado 2 vezes nas páginas 41 e 55.

MORRISSEY, K.; PLATER, A.; DEAN, M. The cost of electric power outages in the residential sector: A willingness to pay approach. *Applied energy*, v. 212, p. 141–150, 2018. Publisher: Elsevier. Citado na página 34.

MUKHERJEE, S.; NATEGHI, R.; HASTAK, M. A multi-hazard approach to assess severe weather-induced major power outage risks in the US. *Reliability Engineering & System Safety*, v. 175, p. 283–305, 2018. Publisher: Elsevier. Citado na página 33.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2002. p. 849–856. Citado 2 vezes nas páginas 22 e 41.

RAY, S. C. Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. *Socio-Economic Planning Sciences*, v. 22, n. 4, p. 167–176, 1988. Publisher: Elsevier. Citado na página 14.

RAZI, M. A.; ATHAPPILLY, K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, v. 29, n. 1, p. 65–74, 2005. Publisher: Elsevier. Citado na página 13.

REILLY, A.; GUIKEMA, S. Bayesian multiscale modeling of spatial infrastructure performance predictions with an application to electric power outage forecasting. *Journal of infrastructure systems*, v. 21, n. 2, p. 04014036, 2015. Publisher: American Society of Civil Engineers. Citado na página 33.

SAMOILENKO, S.; OSEI-BRYSON, K.-M. Increasing the discriminatory power of DEA in the presence of the sample heterogeneity with cluster analysis and decision trees. *Expert Systems with Applications*, v. 34, n. 2, p. 1568–1581, 2008. Publisher: Elsevier. Citado na página 13.

SEBER, G. A.; LEE, A. J. *Linear regression analysis*. [S.l.]: John Wiley & Sons, 2012. v. 329. Citado na página 36.

SILVA, A. V. da et al. A close look at second stage data envelopment analysis using compound error models and the Tobit model. *Socio-Economic Planning Sciences*, p. 1–16, 2018. Publisher: Elsevier. Citado na página 12.

SIMAR, L.; WILSON, P. W. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management science*, v. 44, n. 1, p. 49–61, 1998. Publisher: INFORMS. Citado 2 vezes nas páginas 19 e 20.

SIMPSON, E. H. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 13, n. 2, p. 238–241, 1951. Publisher: Wiley Online Library. Citado na página 53.

TAIMOOR, N. et al. Power Outage Estimation: The Study of Revenue-led Top Affected States of US. *IEEE Access*, 2020. Publisher: IEEE. Citado na página 34.

WINSTEN, C. Discussion on Mr. Farrel's Paper. *Journal of the Royal Statistical Society*, A, n. 120, p. 282–284, 1957. Citado na página 11.

ZACHARIADIS, T.; POULLIKKAS, A. The costs of power outages: A case study from Cyprus. *Energy Policy*, v. 51, p. 630–641, 2012. Publisher: Elsevier. Citado na página 33.

A Apêndice do primeiro artigo

A.1 Details of the RJMCMC sampler

Algorithm 1 RJMCMC_Sampling(Y,$Steps$,$Neigh$,π_B,π_D,π_{Sh},π_{Sw},π_{Up})

```
Require: $Y$: input data; $Steps$: number of RJMCMC steps; $Neigh$: neighboring data; $\pi_B$, $\pi_D$, $\pi_{Sh}$, $\pi_{Sw}$ and $\pi_{Up}$: probabilities
  of the five moves.
Ensure: RJMCMC samples: $M_k$ and $G_k$ at each step.
  $\mu_0 \leftarrow 0$; $\sigma_0^2 \leftarrow 1$
  $k \leftarrow$ prior mean of the number of clusters.
  $G_k \leftarrow$ randomly select $k$ cluster centers.
  $M_k \leftarrow$ apply the update move (Eqs. (2.10) and (2.11))
  for s=1 to Steps do
      Randomly choose one of the five moves with probabilities
          $\pi_B$,$\pi_D$,$\pi_{Sh}$,$\pi_{Sw}$,$\pi_{Up}$.
      if birth move is selected then
          $G^*_{k+1} \leftarrow$ randomly choose a new cluster center and randomly
                  impute in vector $G_k$.
          $\mu_r \leftarrow$ random normal value using Eqs. (2.4) and (2.5).
          $M^*_{k+1} \leftarrow$ update $M_k$ with $\mu_r$.
          Use $M^*_{k+1}$, $G^*_{k+1}$, $M_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Birth}$ (Eq. (2.6))
          $\alpha \leftarrow \min(1, A_{Birth})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $M_k \leftarrow M^*_{k+1}$
              $G_k \leftarrow G^*_{k+1}$
          end if
      end if
      if death move is selected then
          $G^*_{k-1} \leftarrow$ delete one random cluster center from $G_k$.
          $M^*_{k-1} \leftarrow$ delete the respective mean parameter from $M_k$.
          Use $M^*_{k-1}$, $G^*_{k-1}$, $M_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Death}$ (Eq. (2.7))
          $\alpha \leftarrow \min(1, A_{Death})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $M_k \leftarrow M^*_{k-1}$
              $G_k \leftarrow G^*_{k-1}$
          end if
      end if
      if shift move is selected then
          $G^*_k \leftarrow$ shift a randomly selected cluster center in $G_k$ to a random new area
                  in the same cluster.
          $M^*_k \leftarrow M_k$.
          Use $M^*_k$, $G^*_k$, $M_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Shift}$ (Eq. (2.8))
          $\alpha \leftarrow \min(1, A_{Shift})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $M_k \leftarrow M^*_k$
              $G_k \leftarrow G^*_k$
          end if
      end if
      if switch move is selected then
          $G^*_k \leftarrow$ switch two random cluster centers in $G_k$.
          $M^*_k \leftarrow$ switch the respective mean parameters from $M_k$.
          Use $M^*_k$, $G^*_k$, $M_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Switch}$ (Eq. (2.9))
          $\alpha \leftarrow \min(1, A_{Switch})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $M_k \leftarrow M^*_k$
              $G_k \leftarrow G^*_k$
          end if
      end if
      if update move is selected then
          $M_k \leftarrow$ Update the mean parameters in vector $M_k$ using a normal random numbers with
            mean $\mu_*$ (Eq. (2.10)) and variance $\sigma^2_*$ (Eq. (2.11)), $\mu_j \sim Normal(\mu_*, \sigma^2_*)$.
          $\sigma^2 \leftarrow$ Update the variance parameter using an inverse-Chi squared random number,
            $\sigma^2 \sim$ Inv-$\chi^2_{(n-k,s^2)}$ (Eq. (2.12))
      end if
  end for
```

Tabela A.1 – Operational costs, ANEEL DEA efficiencies, and changes after applying DEA to each cluster, sorted in decreasing order of cost differences between clusters and original costs. Major changes in efficiency indexes are highlighted.

DSO	OPEX (R$)	DEA	Cluster	DEA	Efficient	Cluster eff.	OPEX	Cluster
			DEA	change	OPEX (R$)	OPEX (R$)	change (R$)	
LIGHT	722,222,026.67	0.7839	0.8585	0.0746	566,164,085.62	620,035,208.38	53,871,122.75	2
CEMAT	423,266,803.33	0.7611	0.8872	0.1261	322,137,914.71	375,505,981.35	53,368,066.64	1
ELETROPAULO	1,255,830.566.67	0.8692	0.8978	0.0285	1,091,601,201.50	1,127,442,012.68	35,840,811.19	2
COPEL	1,225,581.910.00	0.6363	0.6577	0.0214	779,838,353.26	806,069,990.62	26,231,637.35	1
CEMIG	2,041,586.440.00	0.6904	0.7019	0.0115	1,409,558,101.29	1,432,963,631.09	23,405,529.80	2
CELG	762,130.693.33	0.6854	0.7130	0.0276	522,375,199.97	543,409,078.53	21,033,878.56	2
CELPA	577,061.083.33	0.5618	0.5971	0.0353	324,181,293.18	344,568,130.82	20,386,837.64	1
CERON	239,274.793.33	0.4824	0.5505	0.0681	115,417,359.45	131,722,731.56	16,305,372.10	1
ESCELSA	302,786.963.33	0.7105	0.7640	0.0535	215,121,723.61	231,329,840.81	16,208,117.19	2
ENERSUL	331,261.320.00	0.6638	0.7111	0.0473	219,895,151.68	235,569,603.77	15,674,452.09	1
CPFL PAULISTA	720,481.060.00	0.9463	0.9667	0.0204	681,804,974.32	696,481,235.96	14,676,261.63	2
CEEE	597,813.956.67	0.4109	0.4312	0.0203	245,661,139.49	257,770,570.30	12,109,430.81	1
CEB	333,767.260.00	0.5244	0.5571	0.0327	175,026,169.25	185,945,986.66	10,919,817.41	2
AMPLA	479,317.470.00	0.6998	0.7175	0.0177	335,446,567.37	343,928,567.10	8,481,999.72	2
BANDEIRANTE	327,364.556.67	0.8173	0.8418	0.0245	267,543,956.62	275,561,328.86	8,017,372.24	2
ELEKTRO	463,617.950.00	0.9382	0.9492	0.0110	434,969,778.88	440,076,792.69	5,107,013.81	2
ELETROACRE	89,155.566.67	0.5096	0.5659	0.0563	45,432,241.24	50,451,410.61	5,019,169.37	1
CELESC	842,382.040.00	0.6191	0.6228	0.0037	521,490,795.12	524,602,257.91	3,111,462.79	1
COSERN	196,500.780.00	0.9192	0.9307	0.0115	180,614,479.03	182,878,056.59	2,263,577.56	2
CHESP	12,527.823.33	0.7948	0.9640	0.1692	9,957,058.29	12,076,365.05	2,119,306.76	2
ENE. BORBOREMA	35,458.666.67	0.7306	0.7897	0.0591	25,907,284.95	28,001,418.71	2,094,133.76	2
CELPE	549,361.833.33	0.8692	0.8729	0.0037	477,528,928.99	479,552,828.82	2,023,899.84	2
ENE. SERGIPE	164,595.263.33	0.5999	0.6116	0.0117	98,735,547.87	100,663,288.75	1,927,740.88	2
AME	374,980.226.67	0.3239	0.3289	0.0050	121,468,720.88	123,332,231.06	1,863,510.18	1
CEAL	312,737.580.00	0.4351	0.4405	0.0054	136,063,148.88	137,752,552.65	1,689,403.77	2
AES SUL	270,415.596.67	0.8131	0.8183	0.0053	219,867,105.97	221,293,566.13	1,426,460.16	1
MOCOCA	9,542.343.33	0.9152	1.0000	0.0848	8,733,233.06	9,542,343.33	809,110.27	2
SULGIPE	36,088.133.33	0.6624	0.6802	0.0177	23,906,494.05	24,546,300.27	639,806.22	2
BRAGANTINA	38,394.170.00	0.6854	0.7018	0.0164	26,314,825.44	26,944,307.55	629,482.11	2
CFLO	14,326.320.00	0.6714	0.7117	0.0403	9,618,035.01	10,195,671.94	577,636.94	1
DME-PC	30,082.166.67	0.4164	0.4342	0.0178	12,526,454.31	13,060,971.52	534,517.21	2
ENE. PARAÍBA	249,989.183.33	0.8210	0.8230	0.0020	205,245,777.08	205,737,856.66	492,079.58	2
CPEE	13,787.390.00	0.8876	0.9146	0.0271	12,237,063.94	12,610,116.20	373,052.26	2
NACIONAL	29,008.853.33	0.6726	0.6846	0.0120	19,511,007.50	19,859,199.75	348,192.25	2
CEPISA	334,005.756.67	0.5893	0.5903	0.0010	196,839,326.96	197,180,192.36	340,865.40	2
ENE. MINAS GERAIS	95,472.570.00	0.8291	0.8326	0.0035	79,155,784.18	79,487,813.12	332,028.94	2
ELETROCAR	13,944.003.33	0.5109	0.5318	0.0209	7,124,250.89	7,415,991.34	291,740.45	1
COOPERALIANÇA	9,767.343.33	0.6302	0.6546	0.0244	6,155,163.57	6,393,630.73	238,467.17	1
SANTA MARIA	28,950.813.33	0.8087	0.8161	0.0074	23,411,296.57	23,626,859.65	215,563.08	2
IGUAÇU	13,072.183.33	0.5535	0.5673	0.0138	7,234,987.80	7,415,274.47	180,286.67	1
DEMEI	8,913.560.00	0.5737	0.5836	0.0098	5,113,872.60	5,201,522.23	87,649.62	1
CAIUA	56,770.493.33	0.7362	0.7376	0.0014	41,794,327.49	41,875,755.46	81,427.96	2

B Apêndice do segundo artigo

B.1 Details of the RJMCMC sampler

Algorithm 2 RJMCMC_Sampling($\mathbf{Y}$,$\mathbf{X}$,$Steps$,$Neigh$,λ_0,a_0,b_0,π_B,π_D,π_{Up})

```
Require: $\mathbf{Y}$, $\mathbf{X}$: input data; $Steps$: number of RJMCMC steps; $Neigh$: neighboring data; $\pi_B$, $\pi_D$ and $\pi_{Up}$: probabilities of the three moves.
Ensure: RJMCMC samples: $M_k$ and $G_k$ at each step.
  $k \leftarrow$ prior mean of the number of clusters.
  $G_k \leftarrow$ randomly select $k$ cluster centers.
  $\mathbf{B}_k \leftarrow$ apply the update move (Eq. (2.4))
  for s=1 to Steps do
      Randomly choose one of the three moves with probabilities
          $\pi_B$,$\pi_D$,$\pi_{Up}$.
      if birth move is selected then
          $G^*_{k+1} \leftarrow$ randomly choose a new cluster center and randomly
                  impute in vector $G_k$.
          $\boldsymbol{\beta}_r \leftarrow$ multivariate normal random value using Eq. (2.4).
          $\mathbf{B}^*_{k+1} \leftarrow$ update $\mathbf{B}_k$ with $\boldsymbol{\beta}_r$.
          Use $\mathbf{B}^*_{k+1}$, $G^*_{k+1}$, $\mathbf{B}_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Birth}$ (Eq. (2.6))
          $\alpha \leftarrow \min(1, A_{Birth})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $\mathbf{B}_k \leftarrow \mathbf{B}^*_{k+1}$
              $G_k \leftarrow G^*_{k+1}$
          end if
      end if
      if death move is selected then
          $G^*_{k-1} \leftarrow$ delete one random cluster center from $G_k$.
          $\mathbf{B}^*_{k-1} \leftarrow$ delete the respective mean parameter from $\mathbf{B}_k$.
          Use $\mathbf{B}^*_{k-1}$, $G^*_{k-1}$, $\mathbf{B}_k$, $G_k$, $Neigh$ and $Y$ to calculate $A_{Death}$ (Eq. (2.7))
          $\alpha \leftarrow \min(1, A_{Death})$
          $u \leftarrow$ uniform random number between 0 and 1.
          if $u \leq \alpha$ then
              $\mathbf{B}_k \leftarrow \mathbf{B}^*_{k-1}$
              $G_k \leftarrow G^*_{k-1}$
          end if
      end if
      if update move is selected then
          $\mathbf{B}_k \leftarrow$ Update the mean parameters in vector $\mathbf{B}_k$ using multivariate normal random numbers with
            mean of $\left(\mathbf{X}_r^T\mathbf{X}_r + \lambda_0\mathbf{I}\right)^{-1}\mathbf{X}_r^T\mathbf{y}_r$ and variance of $\sigma^2\left(\mathbf{X}_r^T\mathbf{X}_r + \lambda_0\mathbf{I}\right)^{-1}$ (Eq. (2.4)).
          $\sigma^2 \leftarrow$ Update the variance parameter using an inverse-Gamma random number,
            $\sigma^2 \sim \mathrm{IG}(a_n, b_n)$ (Eq. (2.12))
      end if
  end for
```