Álvaro Ledo Ferreira

Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico

Belo Horizonte

Novembro de 2021

Álvaro Ledo Ferreira

Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico

Trabalho apresentado ao Programa de Pós-Graduação em Engenharia de Produção da UFMG como parte dos pré-requisitos para obtenção do título de Doutor em Engenharia de Produção

Universidade Federal de Minas Gerais

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Produção

Orientador: Marcelo Azevedo Costa

Belo Horizonte

Novembro de 2021

Ferreira, Álvaro Lédo.
F383m Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando base de dados e conhecimento técnico [recurso eletrônico] / Álvaro Lédo Ferreira. - 2021.
1 recurso online (246 f. : il., color.) : pdf.

Orientador: Marcelo Azevedo Costa.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndices: f. 133-246.
Bibliografia: f. 124-132.

1. Engenharia de produção - Teses. 2. Framework (Programa de computador) - Teses. 3. Modelos de equações estruturais – Teses. 4. Serviços de eletricidade – Teses. 5. Sustentabilidade – Teses. 6. 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

ATA DA DEFESA DE TESE DO ALUNO
ALVARO LÉDO FERREIRA

Realizou-se, no dia 23 de novembro de 2021, às 14:00 horas, online em https://bityli.com/Lss8Z0, da Universidade Federal de Minas Gerais, a 55ª defesa de tese, intitulada *Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico*, apresentada por ALVARO LÉDO FERREIRA, número de registro 2018692598, graduado no curso de ENGENHARIA DE PRODUÇÃO, como requisito parcial para a obtenção do grau de Doutor em ENGENHARIA DE PRODUÇÃO, à seguinte Comissão Examinadora: Prof(a). Marcelo Azevedo Costa - Orientador (DEP/UFMG), Prof(a). Anderson Laécio Galindo Trindade (DEP/UFMG), Prof(a). Frederico Gualberto Ferreira Coelho (UFMG), Prof(a). José Francisco Moreira Pessanha (Universidade Estadual do Rio de Janeiro), Prof(a). Sidney Lino de Oliveira (PUC Minas).

A Comissão considerou a tese:

(**X**) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Belo Horizonte, 23 de novembro de 2021.

Prof(a). Marcelo Azevedo Costa (Doutor)

Prof(a). Anderson Laécio Galindo Trindade (Doutor)

Prof(a). Frederico Gualberto Ferreira Coelho (Doutor)

José Francisco Moreira Pessanha
Prof(a). José Francisco Moreira Pessanha (Doutor)

Sidney Lino de Oliveira
Prof(a). Sidney Lino de Oliveira (Doutor)

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

FOLHA DE MODIFICAÇÕES

As modificações exigidas na tese de ALVARO LÉDO FERREIRA, número de registro 2018692598, de número 55, em 23 de novembro de 2021, são relacionadas a seguir:

Os comentário e sugestão serão encaminhados pelos membros da banca para alteração na versão final da tese.

O prazo para entrega da versão final do trabalho com as modificações exigidas acima é de:

( ) 7 (sete) dias (**X**) 30 (trinta) dias ( ) 90 (noventa) dias

( ) Outro (a critério da comissão): _____ dias ficando responsável pela verificação destas o professor:

Álvaro Lédo Ferreira
Assinatura do Aluno

Assinatura do Professor Responsável

Atesto que as modificações exigidas FORAM integralmente cumpridas.

Belo Horizonte, 9 de dezembro de 2021.

Assinatura do Professor Responsável

Este trabalho é dedicado a minha família, minha namorada, meus amigos e especialmente ao meu orientador Marcelo pela paciência e ensinamentos.

Agradecimentos

Agradeço primeiramente à minha mãe Sandra, meu pai Álvaro e minha irmã Lorena por serem a minha base para todo o meu desenvolvimento como ser humano e profissional.

Agradeço ao meu orientador e amigo Professor Marcelo pela confiança, paciência e ensinamentos ao longos dos últimos 3 anos. Que essa amizade cresça e se fortaleça nos próximos anos.

Agradeço à minha namorada Lígia pela amizade, cumplicidade, carinho e companhia durante tempos tão difíceis. O seu impacto e importância na minha vida são imensuráveis. Que continuemos ouvindo Salif Keita, dançando, criando asas e com os olhos brilhando por muito tempo.

Agradeço ao Professor Anderson por me iniciar na pós-graduação e ser um exemplo de pessoa e profissional.

Agradeço aos meus amigos Alan, Danilo e Felipe por serem fonte de alegria e diversão há mais de 15 anos.

Agradeço aos meus amigos da graduação Diana, Nayáry e Tiago por trilharem junto comigo o caminho da Engenharia de Produção.

Agradeço aos meus amigos do LADEC Leandro, Tiago, Rodrigo e Cassius pelos momentos de discussão e descontração que tornaram o trajeto do doutorado mais humano e acolhedor.

Agradeço aos companheiros de projeto Professor Vinícius, Professora Ana, Tomás, Luiz, Mateus e Francisco por me acompanharem no processo do P&D.

Agradeço à equipe da CEMIG-D, principalmente ao Sérgio e Iguatinan, por todo o apoio no desenvolvimento do P&D.

Agradeço à UFMG, a qual considero minha segunda casa, por ter me possibilitado realizar o Mestrado e Doutorado em um local com uma das melhores infraestrutura, corpo técnico e docente.

Por fim, agradeço a todas as pessoas que direta ou indiretamente contribuíram para a realização da minha pesquisa.

"A ciência nunca resolve um problema sem criar pelo menos outros dez."
(George Bernard Shaw)

Resumo

O setor elétrico brasileiro se diferencia de outros setores de serviços mais tradicionais. Ele se caracteriza como um monopólio natural e portanto deve ser regulado para garantir a sua eficiência. O órgão responsável pela regulação no Brasil é a Agência Nacional de Energia Elétrica (ANEEL). Dentre as suas diversas atribuições, pode-se destacar a atuação como um agente mediador entre as empresas geradoras, transmissoras e distribuidoras de energia elétrica e os clientes. No caso das distribuidoras, existem outros detalhes que dificultam a compreensão do funcionamento do setor de distribuição. Da forma que os contratos são firmados com a ANEEL, as empresas distribuidoras não são proprietárias dos ativos elétricos dos quais fazem a gestão, são somente o possuidor já que o proprietário é a União. Essa condição, em conjunto com crises econômicas e políticas nos últimos anos, faz com que a gestão das empresas distribuidoras se torne um desafio. Dentre as distribuidoras do Brasil, a que possui a maior operação (número de clientes e receita) é a Companhia Energética de Minas Gerais (CEMIG). Essa condição torna a empresa o *benchmark* para as demais do setor. Ademais, a realização de um projeto de P&D com a empresa permitiu uma aproximação e acesso facilitado a seus dados. Assim, neste trabalho é proposto a aplicação de diferentes ferramentas de modelagem, estatística e computacionais no intuito de prover suporte à gestão de diferentes áreas da CEMIG. Para isso, primeiramente são utilizados *frameworks* para modelar, entender e avaliar o modelo de negócio das empresas distribuidoras. Os resultados encontrados permitem visualizar o funcionamento do setor e servem como base para a construção de outros modelos estatísticos e matemáticos. Em seguida foi realizada a análise do setor pelo ponto de vista da sustentabilidade tal qual foi definida pela ANEEL. No estudo foram utilizadas análises comparativas e geográficas, e modelos ordinais logísticos de regressão. Os resultados permitiram visualizar o estado da sustentabilidade do setor entre 2011 e 2018, além de avaliar o impacto na sustentabilidade de fatores como tipo de controle e região. Em seguida foi realizado um estudo sobre o índice de Duração Equivalente de Interrupção por Unidade Consumidora (DEC) utlizando MEE e Modelos híbridos multicamadas. Esse índice é utilizado pela ANEEL como o principal indicador do nível de qualidade das operações das empresas distribuidoras. O objetivo nessa etapa foi duplo: 1) identificar as principais variáveis que impactam no DEC; 2) encontrar o melhor modelo preditivo possível para o DEC. O primeiro objetivo permite à empresa identificar a melhor forma de diversificar os seus investimentos e tomar decisões com bases quantitativas. O segundo objetivo almeja possibilitar à empresa a realização de simulações e análises de cenários futuros ou incertos. Os resultados encontrados permitiram identificar as variáveis mais relevantes ao analisar o DEC. Também foi alcançado um modelo com poder preditivo considerável utilizando como input variáveis contábeis, operacionais, climáticas e geográficas. Este trabalho de pesquisa se caracteriza como uma proposta de doutorado em Engenharia de Produção pois não foi encontrado na literatura científica nacional e internacional um modelo estatístico computacional desenvolvido especificamente para auxiliar a gestão de uma empresa brasileira de distribuição de energia elétrica. Os resultados obtidos até o presente momento são inéditos e já foram objeto de premiação em um evento nacional do setor energético. Além disso, foi desenvolvida uma ferramenta computacional inovadora que permite o ajuste e a simulação das metodologias desenvolvidas em uma interface *user-friendly*. O modelo proposto

para o indicador DEC agrega modelos Bayesianos para regionalização, modelos de regressão múltipla e modelos de aprendizado de máquina (*machine learning*) sendo, portanto, caracterizado como um modelo híbrido. Este modelo pode ser facilmente aplicado a outros contextos como as compensações financeiras e a receita anual da CEMIG-D ou qualquer outra empresa Brasileira de distribuição de energia elétrica.

Palavras-chaves: Setor elétrico, *Framework*, Sustentabilidade, Regionalização, Equações estruturais, Modelos híbridos.

Abstract

The Brazilian electricity sector differs from other more traditional service sectors. It is characterized as a natural monopoly and therefore must be regulated to guarantee its efficiency. The agency responsible for regulation in Brazil is the National Electric Energy Agency (ANEEL). Among its various attributions, it can be highlighted the acting as a mediating agent between the generators, transmission and distribution companies of electric energy and the customers. In the case of distributors, there are other details that make it difficult to understand how the distribution sector works. As the contracts are signed with ANEEL, the distribution companies do not own the electrical assets they manage, they are only the owner since the owner is the Union. This condition, together with economic and political crises in the last few years, makes the management of distribution companies a challenge. Among the distributors in Brazil, the one with the largest operation (number of customers and revenue) is Companhia Energética de Minas Gerais (CEMIG). This condition makes the company the benchmark for others in the sector. Furthermore, carrying out an R&D project with the company allowed for an approach and easier access to its data. Thus, in this work, the application of different modeling, statistical and computational tools is proposed in order to support the management of different areas of CEMIG. For this, first frameworks are used to model, understand and evaluate the business model of the distribution companies. The results found allow us to visualize the functioning of the sector and serve as a basis for the construction of other statistical and mathematical models. Then, an analysis of the sector was carried out from the point of view of sustainability as defined by ANEEL. The study used comparative and geographic analyses and ordinal logistic regression models. The results allowed us to visualize the state of sustainability of the sector between 2011 and 2018, in addition to evaluating the impact on sustainability of factors such as type of control and region. Then, a study was carried out on the Index of Equivalent Outage Duration per Consumer Unit (DEC) using SEM and multilayer hybrid models. This index is used by ANEEL as the main indicator of the quality level of the operations of the distribution companies. The objective in this step was twofold: 1) to identify the main variables that impact the DEC; 2) find the best possible predictive model for DEC. The first objective allows the company to identify the best way to diversify its investments and make decisions on a quantitative basis. The second objective aims to enable the company to carry out simulations and analyses of future or uncertain scenarios. The results found allowed us to identify the most relevant variables when analysing the DEC. A model with considerable predictive power was also achieved using accounting, operational, climatic and geographic variables as input. This research work is characterized as a proposal for a doctoral degree in Production Engineering as it was not found in the national and international scientific literature a computational statistical model developed specifically to assist the management of a Brazilian electricity distribution company. The results obtained so far are unprecedented and have already been awarded at a national event in the energy sector. In addition, an innovative computational tool was developed that allows the adjustment and simulation of the developed methodologies in a user-friendly interface. The proposed model for the DEC indicator aggregates Bayesian models for regionalization, multiple regression models and machine learning models (machine learning) and is, therefore, characterized as a hybrid

model. This model can be easily applied to other contexts such as financial compensation and annual revenue of CEMIG-D or any other Brazilian electricity distribution company.

Keywords: Electric sector, Framework, Sustainability, Regionalization, Structural equations, Hybrid models.

Disseminação da pesquisa

Artigos submetidos em periódicos

Ferreira, Álvaro, Costa, Marcelo, Castro, Tomás, Ribeiro, Sérgio, Monteiro, Iguatinan. "Desenvolvimento do modelo de negócio de uma distribuidora brasileira." Revista Gestão & Produção.

Ferreira, Álvaro, Castro, Tomás, Costa, Marcelo, Ribeiro, Sérgio, Monteiro, Iguatinan. "An analysis of the financial-economic sustainability in Brazil's electricity distribution sector." Energy, Sustainability and Society, Springer.

Ferreira, Álvaro, Costa, Marcelo, Ribeiro, Sérgio, Monteiro, Iguatinan. "Multilayer Hybrid Models for the power outage index." Expert Systems with Applications, Springer.

Costa, Marcelo, Mineti, Leandro, **Ferreira, Álvaro**. "A novel clustering-based spatial regression model applied to consumer power outage indicator." Expert Systems with Applications, Springer.

Artigos submetidos em congressos

Costa, Marcelo, Mayrink, Vinicius, Lopes, Ana, **Ferreira, Álvaro**, Mello, Mateus, Neto, Francisco, Castro, Tomás, Oliveira, Luiz, Ribeiro, Sérgio e Monteiro, Iguatinan. "Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico." 21º Seminário de Planejamento Econômico-Financeiro e de Regulação do Setor Energético Brasileiro (SEPEF), 2021.

Ferreira, Álvaro, Costa, Marcelo, Ribeiro, Sérgio e Monteiro, Iguatinan. "Análise de sustentabilidade econômico-financeira das empresas distribuidoras de energia elétrica no Brasil." Seminário Nacional de Distribuição de Energia Elétrica (SENDI), 2020.

Costa, Marcelo, **Ferreira, Álvaro**, Ribeiro, Sérgio, Pepe, Renan, Monteiro, Iguatinan. "Desenvolvimento do Modelo de Negócios de uma Empresa de Distribuição de Energia Elétrica." 20º Seminário de Planejamento Econômico-Financeiro e de Regulação do Setor Energético Brasileiro (SEPEF), 2019.

Sumário

Parte I

Introdução

1 Considerações Iniciais

1.1 Contextualização

1.1.1 O Setor Elétrico Brasileiro

O conhecimento sobre o funcionamento do setor de distribuição de energia elétrica brasileiro pelo cidadão comum e leigo por vezes se mostra falacioso, não é difícil ouvir inverdades ou afirmações duvidosas. Este é um segmento com diversas peculiaridades, detalhes e que, por tais motivos, pode causar confusão sobre sua operação. Assim, esta introdução tem o propósito de desmistificá-lo para melhor entendimento do leitor.

A começar que, diferente de outros mercados nos quais existe concorrência entre as empresas, no setor de distribuição elétrica existe um monopólio natural. Alguns motivos para isso são o alto investimento necessário para entrar no mercado e questões logísticas; considere o caos nas cidades caso existissem diversos postes e cabos cruzando as ruas, cada um pertencente a uma empresa diferente. Apesar de, nesse cenário, cada cliente ter a opção de escolher o seu fornecedor de energia a cada mês, tal qual ocorre atualmente com a telefonia por exemplo, a tecnologia necessária para implementá-la de forma eficaz ainda não existe.

Entretanto, a existência de um monopólio natural, conforme estudado por diversos pesquisadores, como por exemplo Posner (1999), tende a resultar em mais desvantagens do que vantagens para os consumidores. Clientes em um mercado monopolista não possuem poder de negociação e ficam reféns dos preços e ações dos fornecedores; os fornecedores, por sua vez, por não possuírem concorrência, podem definir preços abusivos e não possuem incentivos para operar de forma eficiente ou até mesmo melhorar suas operações.

Essa condição leva à suposição errônea pelo consumidor de que, por falta de concorrência direta, está totalmente à mercê das decisões do seu fornecedor de energia. O governo federal, para mitigar os impactos negativos dos monopólios, criou um órgão responsável por regulamentar e fiscalizar as atividades das empresas de geração, transmissão, distribuição e comercialização de energia elétrica.

Esse órgão é a Agência Nacional de Energia Elétrica (ANEEL), uma autarquia vinculada ao Ministério de Minas e Energia criada em 1996 e que iniciou suas atividades em dezembro de 1997 (ANEEL, 2019). A sua missão é de “proporcionar condições favoráveis para que o mercado de energia elétrica se desenvolva com equilíbrio entre os agentes e em benefício da sociedade” e dentre as suas principais atribuições pode-se listar (Brasil, 1996):

- Estabelecer tarifas;
- Regular a geração (produção), transmissão, distribuição e comercialização de energia elétrica;

- Fiscalizar, diretamente ou mediante convênios com órgãos estaduais, as concessões, as permissões e os serviços de energia elétrica;
- Implementar as políticas e diretrizes do governo federal relativas à exploração da energia elétrica e ao aproveitamento dos potenciais hidráulicos;
- Dirimir as divergências, na esfera administrativa, entre os agentes e entre esses agentes e os consumidores; e,
- Promover as atividades de outorgas de concessão, permissão e autorização de empreendimentos e serviços de energia elétrica, por delegação do Governo Federal.

Essa lista contrapõe uma segunda suposição inexata de leigos de que a ANEEL é um órgão cujo papel é proteger as empresas do setor elétrico, quando na verdade ela atua como mediadora entre os interesses de consumidores e fornecedores. Uma prática comum nesse sentido é a abertura de Consultas Públicas em que qualquer agente, seja ele individual, público ou privado, pode dar a sua opinião e contribuir ativamente na tomada de decisões pela agência. É frequente também a sua preocupação com melhorias para o consumidor, como por exemplo a desoneração tarifária (FIRJAN, 2019) e a qualidade do serviço prestado (Energia, 2019).

Abordando especificamente o item "Estabelecer tarifas", o modo escolhido pela ANEEL para executá-lo é conhecido como "Regulação pelo preço" (*Price Cap*), ou também como "Regulação por incentivos". De forma sucinta, compara-se o desempenho e eficiência operacional de todas as empresas distribuidoras do Brasil, e a partir desse resultado é estipulado o preço da tarifa que será praticado por cada uma e, consequentemente, a receita máxima que cada uma poderá obter no próximo ano. Caso os seus custos praticados sejam menores do que os estipulados, parte do superávit é mantido pela empresa como recompensa pelo desempenho eficiente, o restante é utilizado para subsidiar uma redução da tarifa no próximo ano; caso contrário, a empresa arca com o prejuízo do déficit em razão de sua ineficiência.

Esse cálculo é realizado anualmente através de dois mecanismos: o Reajuste Tarifário Anual (RTA) e a Revisão Tarifária Periódica (RTP) (ANEEL, 2016d). O RTA ocorre anualmente no período entre duas RTPs e tem como objetivo manter a estabilidade econômico-financeira das distribuidoras ajustando suas receitas em função da inflação, da variação do desempenho operacional, de variações de preço da energia e de ganhos de produtividade. Tais ganhos são avaliados através do Fator X, uma variável definida para compartilhar com a sociedade parte dos ganhos de produtividade obtidos pelas distribuidoras e reduzir as tarifas (ABRADEE, 2018). Maiores informações sobre o Fator X podem ser encontradas em ANEEL (2016d).

O RTP é realizado em média a cada 5 anos, tempo chamado de Ciclo de Revisão Tarifária Periódico (CRTP), e é o momento no qual a concessão é renegociada e recalculada de forma a adequar a sua tarifa tal que seja suficiente para a empresa operar de forma eficiente e auferir dividendos satisfatórios e dentro da sua realidade econômica (ANEEL, 2016d).

Essa revisão tem como um de seus principais objetivos garantir o poder de compra das concessionárias ao corrigir os valores de acordo com as variações da economia, já que os contratos firmados vigoram por vários anos. No cálculo são considerados os investimentos realizados

em infraestrutura, qualidade do serviço prestados e do produto (energia) entregue, eficiência operacional dos custos, ganhos de escala (i.e., aumento de consumo e de consumidores) através do Fator X, assim como a variação inflacionária do ano anterior (ABRADEE, 2018).

No RTP calcula-se para cada empresa, com base nos dados históricos e na sua eficiência, qual a receita necessária para cobrir os seus custos, denominada Receita Requerida (RR). A ANEEL entende que uma parte dos custos de operação dependem diretamente da eficiência da gestão das empresas enquanto que outra parte é influenciado por forças que fogem ao controle das distribuidoras. No primeiro caso, elas devem se responsabilizar financeiramente caso ocorra uma má gestão, no segundo caso elas não arcam com prejuízos decorrente de ações de agentes externos e repassam essa diferença diretamente aos consumidores. Essa diferença entre os custos é importante, pois refuta uma terceira opinião incorreta de que todo aumento de custos é repassado diretamente ao consumidor.

O custo que não depende das empresas é denominado Parcela A e consiste principalmente pelos custos com aquisição da energia elétrica junto às empresas geradoras, com a transmissão dessa energia através da infraestrutura das empresas transmissoras e com os encargos pagos ao Governo, conforme apresentado na Figura 1.1. Já o segundo tipo de custo que depende da gestão das empresas é denominado Parcela B e consiste principalmente dos custos com pessoal, manutenção, serviços de terceiros, instalações móveis e imóveis, entre outros.

As empresas distribuidoras são responsáveis por arrecadar o dinheiro através da cobrança das contas de luz e de repassar os valores devidos aos demais agentes do setor. Essa divisão ocorre, aproximadamente, nas seguintes proporções: 32% da fatura é repassada para as empresas geradoras, 8% para as transmissoras e 41% para o Governo, ou seja, 81% para pagamento da Parcela A, restando 19% para as distribuidoras cobrirem os custos da Parcela B. Esse fato rebate outro engano comum, de que as distribuidoras retêm a maior parte do capital arrecadado com as contas de luz, o que se mostra distante da realidade.

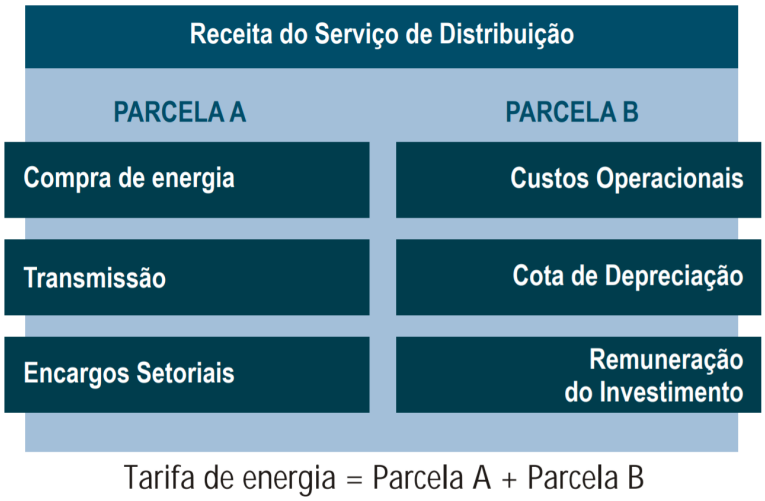


Figura 1.1 – Representação das tarifas divididas em dois grupos (Parcela A e B).

Assim, considere um exemplo em que o valor efetivamente pago pela distribuidora para a aquisição da energia junto às geradoras foi maior do que o estipulado no RTP/RTA; nesse caso,

a RR da empresa é incrementada na quantidade exata e suficiente para cobrir essa diferença através do aumento das tarifas.

Caso esse aumento inesperado de custos esteja relacionado, por exemplo, com a contratação ineficiente de mão-de-obra para a realização de serviços na empresa, nem a RR nem a tarifa sofrem alterações e resulta em uma redução dos lucros obtidos no período. A diferença entre o RTP e o RTA nesse procedimento é que no RTP as Parcelas A e B são totalmente recalculadas a partir dos resultados alcançados pela empresa no último ciclo de revisão, ao passo que no RTA a Parcela A é totalmente revisada (anualmente) e a Parcela B é somente atualizada pelos indicadores já apresentados.

De posse da RR, a ANEEL também realiza uma projeção da variação do mercado no próximo ano para definir a tarifa para os clientes. Neste cálculo, de forma simplificada, divide-se a RR pela quantidade estimada de energia que será consumida no ano; dessa forma, obtém-se a taxa (R$/kWh) que deverá ser paga pelos clientes para cobrir os custos e permitir à empresa alcançar o nível de receita proposto. Claramente, em razão dos cálculos serem realizados a partir de bases históricas para realizar projeções futuras, ocorrem discrepâncias entre o valor determinado e o realmente praticado; essa diferença é inserida no cálculo do RTP/RTA do ano subsequente.

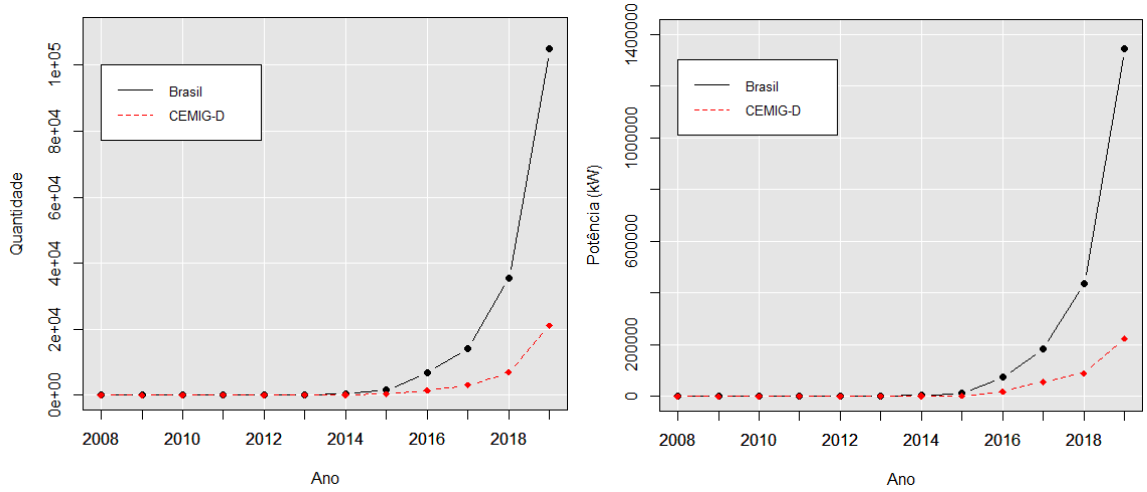
Soma-se a tudo isso o surgimento e expansão da Geração Distribuída (GD) nos últimos anos, uma modalidade de cliente que gera parte da sua energia através, por exemplo, de fonte solar ou eólica. Em 2012 a ANEEL emitiu a Resolução Normativa nº 482 que, entre outras definições, estabeleceu um conjunto de isenções para os micro e minigeradores de energia, o que resultou num crescimento elevado nos anos posteriores, conforme apresentado na Figura 1.2.

No início de 2019 foi aberta uma Consulta Pública para revisar a Resolução Normativa nº 482, discutindo a necessidade e a melhor forma de valorar a energia que é injetada no sistema pelos clientes GD, podendo até mesmo resultar na extinção das isenções e criação de novas taxas. Apesar de ser do interesse das distribuidoras que o mercado GD se desenvolva, principalmente por conta do desenvolvimento de fontes de energia renováveis, essa nova relação cliente-fornecedor traz à tona questões interessantes sobre o futuro. Por exemplo, considere o cenário em que a tecnologia se desenvolva a tal ponto que baterias residenciais se tornem acessíveis; nessa situação, os clientes optarão por se desconectar do sistema elétrico, reduzindo o mercado e consequentemente a receita das empresas.

1.1.2 A CEMIG

A Companhia Energética de Minas Gerais S/A (CEMIG) é uma *holding* responsável, entre outras atividades, pela distribuição da energia elétrica na maior parte do estado de Minas Gerais. Ela foi fundada em 1952 com o nome de Centrais Elétricas de Minas Gerais e atualmente possui 8,4 milhões de clientes em 774 municípios (20 milhões de habitantes), correspondendo a aproximadamente 10% do total de energia distribuída e receita no Brasil. Também é a maior fornecedora de energia para clientes livres do país, com 18% do mercado e um dos maiores grupos geradores, com 89 usinas operando a uma capacidade instalada de 6,1 GWatt.

A CEMIG-D, especificamente, é a divisão do grupo responsável pela distribuição da



(a) Quantidade de GD **(b)** Potência instalada (kW)

Figura 1.2 – Quantidade de unidades de GD instaladas (a) e potência instalada em kW (b) à cada ano no Brasil e na área de concessão da CEMIG-D.

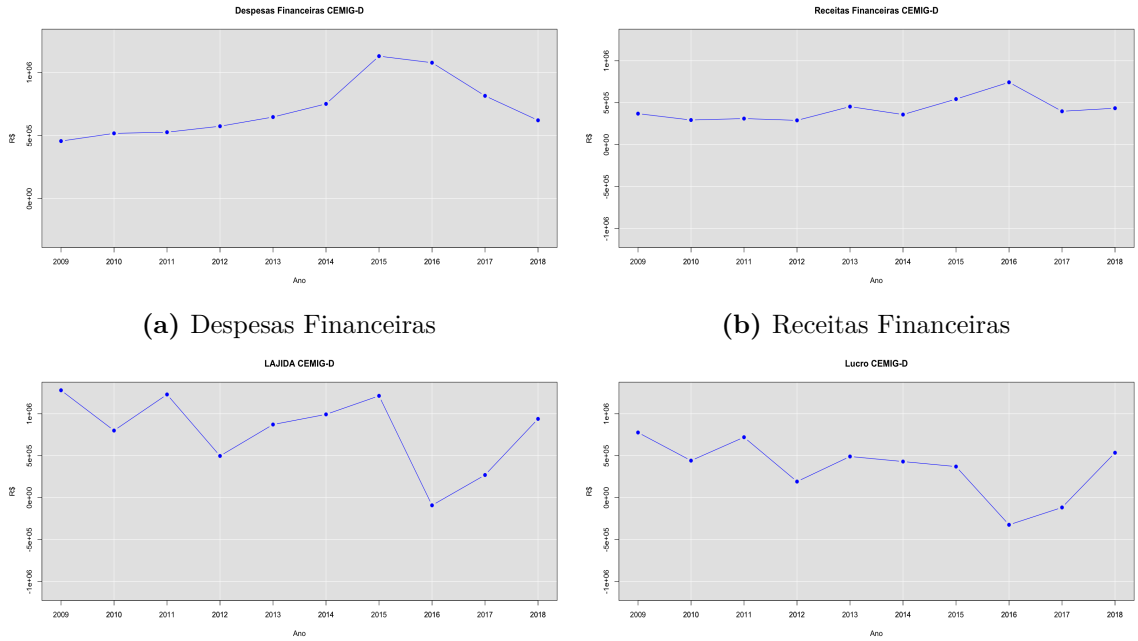
energia até os consumidores (CEMIG, 2012b). É o maior grupo de distribuição de energia da América do Sul e também atua em 31 municípios do Estado do Rio de Janeiro, por meio da empresa de distribuição de energia Light. A sua área de concessão abrange 567,4 mil km^2, aproximadamente 96% do Estado de Minas Gerais. Em extensão de rede, conta com 536,562 mil km de redes de distribuição (108,576 mil km de rede urbana e 410,486 mil km de rede rural) (CEMIG, 2012a).

Em razão do contexto regulatório e da magnitude de suas operações, a CEMIG-D enfrenta desafios para se manter competitiva e economicamente sustentável. Tais desafios podem ser observadas ao se analisar o seu Demonstrativo de Resultado do Exercício (DRE). O DRE é um compilado de informações contábeis que as empresas divulgam à cada trimestre com o seu desempenho no período. Dentre as diversas informações ali contidas, foram consideradas as quatro mais importantes para análise: Lucros Antes de Juros, Impostos, Depreciação e Amortização (LAJIDA); Receitas Financeiras; Despesas Financeiras; e, Lucro.

A Figura 1.3a apresenta as despesas financeiras da CEMIG-D no período entre 2009 e 2018. É possível identificar um aumento no valor absoluto das suas despesas ao longo dos anos, principalmente em 2015 e 2016, quando o maior valor absoluto é alcançado. Entretanto, nos dois anos subsequentes a empresa apresentou melhoria considerável, com o valor absoluto da sua despesa reduzindo e voltando ao patamar dos anos de 2012 e 2013.

Já a sua receita financeira no período pode ser vista na Figura 1.3b, e de forma contrária à despesa, ela apresenta tendência de crescimento. Destaca-se o ano de 2016, em que a empresa alcançou sua maior receita histórica, com resultados 100% maiores dos que os alcançados entre 2010 e 2012. Nos anos seguintes, a empresa retornou aos valores médios apresentados em 2013 e 2014.

As grandes flutuações apresentadas pelas Figura 1.3a e Figura 1.3b indicam uma tendência de grandes variações no seu LAJIDA. A Figura 1.3c apresenta o LAJIDA da CEMIG-D no



(a) Despesas Financeiras **(b)** Receitas Financeiras

(c) LAJIDA **(d)** Lucro

Figura 1.3 – Análise do DRE da CEMIG-D entre os anos de 2009 a 2018: (a) Depesas Financeiras; (b) Receitas Financeiras; (c) LAJIDA; (d) Lucro.

período entre 2009 e 2018, e como esperado, ele apresenta grande variabilidade com tendência de decrescimento. Mesmo com a tendência de crescimento da receita (Figura 1.3b), há uma queda, com atenção ao ano de 2016, no qual o lucro foi negativo mesmo com a empresa atingindo receita recorde nesse ano. A empresa se recuperou dessa grande diferença negativa no ano seguinte, voltando à um valor próximo da média dos anos anteriores.

Por fim, a Figura 1.3d apresenta o real lucro por período obtido pela empresa. O seu comportamento é bastante similar ao encontrado na Figura 1.3c, com exceção dos anos de 2014 e 2015. O LAJIDA nesses anos apresentou valor crescente e maior que o alcançado no ano de 2013, enquanto que na Figura 1.3d, percebe-se um decrescimento para os mesmos anos.

Resumindo, nos últimos anos a empresa apresentou dificuldade de manter estabilidade financeira, seja por motivos que fogem do seu controle (i.e. questões políticas, ambientais e governamentais) ou por dificuldades de gerência. Quanto ao primeiro motivo, pouco ou quase nada pode ser feito pela empresa para mudar o ambiente em que atua. Já o segundo é consequência direta das suas decisões de gestão e investimento.

Assim, talvez até mesmo mais do que em outros segmentos, fica clara a necessidade de uma administração de recursos eficaz baseada na combinação do alto grau de conhecimento técnico/tácito presente na empresa, personificado nos seus colaboradores, com a aplicação de ferramentas estatístico-computacionais baseadas em dados e séries históricas. Tal agregação pode se mostrar um diferencial para a companhia avaliar, por um ponto de vista objetivo, o impacto passado e presente de suas ações e permiti-la definir um modelo de negócios sustentável e próspero para o futuro.

Para alcançar este objetivo, a CEMIG-D iniciou em 2018 o projeto de Pesquisa e

Desenvolvimento (P&D) 0636 cujo título é "Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico". Esse P&D está intimamente associado ao desenvolvimento deste projeto de doutorado, cujo autor e respectivo orientador fazem parte da equipe de pesquisadores.

1.2 Problema de Pesquisa

Com base em todas as questões levantadas sobre o setor elétrico, buscou-se responder a seguinte questão: Como a modelagem estatístico-computacional pode impulsionar/ajudar na gestão de empresas do setor de distribuição elétrica?

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem por objetivo realizar a modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico. A modelagem estatístico-computacional comporta todas as técnicas e metodologias utilizadas, incluindo, mas não limitado a: *frameworks*, equações estruturais, simulação, *machine learning* e regressões lineares. Elas são aplicadas tanto em bases de dados públicas de todas as distribuidoras do Brasil, em grande parte disponibilizadas pela ANEEL, quanto em bases de dados próprias da CEMIG-D, algumas com restrições de compartilhamento. Um terceiro ponto chave para garantir a eficácia do projeto é a incorporação contínua do conhecimento técnico proveniente da experiência dos colaboradores, presentes em todas as etapas e sempre agentes participativos nas discussões.

Ao final, busca-se obter uma ferramenta que possibilite à empresa: avaliar os impactos no presente/futuro de suas decisões passadas; compreender as relações de causa/consequência de suas decisões em função de atributos financeiros e não-financeiros; projetar cenários futuros; simular o resultado de diferentes abordagens estratégicas; entre outras aplicações.

1.3.2 Objetivos Específicos

Assim, os objetivos específicos do trabalho são:

1. Pesquisar sobre o funcionamento do setor elétrico brasileiro e o papel das distribuidoras;
2. Revisar a literatura sobre ferramentas e técnicas com potencial de uso (*Canvas*, Equações Estruturais, *Machine Learning*, etc.);
3. Analisar o contexto da CEMIG-D e o seu modelo de negócio atual;
4. Coletar e explorar bases de dados públicas e próprias da CEMIG-D;
5. Propor e validar um modelo de negócio para a CEMIG-D contendo: 1) DEC, 2) Compensações financeiras, e 3) Receita;
6. Avaliar a qualidade, utilidade e praticidade dos modelos propostos.

1.4 Justificativa e Relevância

O trabalho configura como uma contribuição única para o desenvolvimento do setor de distribuição de energia elétrica brasileiro. Isso se deve tanto em razão do ineditismo no uso de algumas ferramentas nesse contexto quanto da aplicação de ferramentas inovadoras ou inexploradas.

A começar pela contribuição que resulta da modelagem do negócio da CEMIG-D, maior empresa distribuidora do Brasil. Nos últimos anos viu-se a popularização de *frameworks* de modelo de negócio (Canvas, Causa/Consequência), entretanto sem uma aplicação voltada especificamente para empresas distribuidoras. A união desses tópicos resultou em um artigo publicado no Seminário de Planejamento Econômico-Financeiro do Setor Elétrico (SEPEF) 2019. No evento, além da premiação como 2º melhor trabalho, ficou nítido o interesse de representantes de outras empresas no assunto, questionando a possibilidade de replicar tal metodologia em seu próprio local de trabalho. Assim, com a divulgação da metodologia e dos resultados obtidos, espera-se incitar um interesse maior no setor sobre o assunto.

Com este trabalho pretende-se também expandir o campo de aplicações de outras ferramentas, como Modelos de Equações Estruturais (MEE), ao aplicá-la em outros cenários. Tal ferramenta está intimamente ligada com áreas de saúde, psicologia e biologia, com uma quantidade limitada de uso em áreas relacionadas a engenharia. Essa transição se mostra ao mesmo tempo interessante e desafiadora ao exigir certo nível de interpretação e abstração não usuais ao aplicar outras ferramentas mais objetivas.

Paralelamente, busca-se também discutir a Nota Técnica nº 111/2016, que define a criação dos Indicadores de Sustentabilidade Econômico-Financeiros para empresas distribuidoras. Desde 2016 são disponibilizados, a cada trimestre, relatórios com informações críticas sobre todas as empresas do setor. Atualmente os resultados apresentados não são utilizados pela ANEEL como base para tomada de decisões sobre o setor, servindo apenas para acompanhamento do seu desenvolvimento. Apesar disso, os dados nele contidos permitem realizar análises profundas, como observar a saúde financeira das empresas desde 2011, como cada uma reagiu a momentos de crise e realizar uma comparação do seu desempenho.

Outra contribuição para o setor é uma análise objetiva do indicador Duração Equivalente de Interrupção por Unidade Consumidora (DEC). De forma resumida, o DEC indica, em média, o quanto os clientes ficaram sem acesso à energia elétrica durante o mês. Esse indicador é usado pela ANEEL como principal parâmetro para avaliar a qualidade das operações das distribuidoras. Essa abordagem possui pontos positivos e negativos: de um lado, é uma medida objetiva e fácil de calcular/acompanhar tanto pela ANEEL quanto pelas distribuidoras; por outro lado, é uma métrica que sintetiza de modo bastante simples (e com perda considerável de informações) uma infinidade de outras questões relevantes sobre as operações. Isso resulta em uma situação incômoda para as empresas, pois ao mesmo tempo que lhes é fácil calcular o DEC, é muito difícil identificar com precisão quais as variáveis que possuem maior impacto na sua variação, dificultando, consequentemente, o seu processo de tomada de decisão.

Para abordar essa questão na pesquisa, inicialmente utilizou-se a metodologia de Modelos

de Equações Estruturais. Em função da complexidade do problema enfrentado, o MEE, mesmo com uma base de dados com 25 variáveis, não se mostrou suficiente para explicar de maneira satisfatória essa questão.

Recorreu-se então aos Modelos Híbridos Multicamacdas, uma metodologia recentemente proposta por Costa et al. (2019a) adaptada do *bagging and boosting* (Friedman, 2001) . Nos Modelos Híbridos, busca-se agrupar dois ou mais modelos em camadas de tal forma que, com base em um primeiro modelo base, os modelos subsequentes tentem explicar os resíduos do modelo anterior, aumentando assim o poder explicativo total.

Essa metodologia se mostrou inovadora tanto pela própria abordagem objetiva do DEC, ao tentar identificar exatamente quais variáveis o impactam e o nível do impacto, quanto pela aplicação de Modelos Híbridos no contexto energético de distribuição brasileiro, algo nunca antes realizado dessa forma. Outra inovação foi no uso da regionalização. Além da análise regional gráfica, também foi aplicada uma nova metodologia de agrupamento por regionalização. Essa metodologia agregou informações aos modelos que até então estavam ocultas sob os dados geográficos.

A sociedade, a agência reguladora e as empresas distribuidoras se beneficiam deste trabalho. Ele provoca discussões sobre assuntos relevantes para o meio que não receberam devida atenção (Sustentabilidade e DEC). Tais discussões têm o potencial de alavancar o setor e potencializar as operações da empresas, resultando em uma prestação melhor de serviços e tarifas mais baixas para a sociedade.

Além disso, este trabalho também deixa como produto um conjunto de interfaces gráficas. Estas interfaces foram utilizadas durante a pesquisa para gerar diversos resultados aqui apresentados. Apesar de as interfaces terem sido construídas com foco na CEMIG-D, elas podem ser utilizadas com bases de dados de qualquer distribuidora e com outras variáveis além das que foram selecionadas neste trabalho. Somado à sua flexibilidade, as interfaces apresentam potencial ilimitado para a análise e desenvolvimento do setor elétrico.

1.5 Organização do trabalho

O trabalho está dividido em 7 capítulos. O Capítulo 1, Introdução, contém uma contextualização do setor de energia elétrica, informações relevantes sobre a CEMIG-D, problema de pesquisa, os objetivos e a justificativa e relevância do trabalho.

O Capítulo 2 apresenta o artigo “Desenvolvimento do modelo de negócio de uma distribuidora brasileira” submetido à Revista Gestão & Produção. O artigo contém um estudo sobre o modelo de negócios da CEMIG-D. No artigo foram utilizados conceitos referentes a modelagem de negócios e diferentes tipos de *frameworks* (ciclos virtuosos, *CANVAS* e escolha e consequência).

O Capítulo 3 apresenta o artigo “An analysis of the financial-economic sustainability in Brazil’s electricity distribution sector” submetido à Revista Energy, Sustainability and Society. O artigo contém um estudo do setor de distribuição de energia elétrica pelo ponto de vista dos indicadores de sustentabilidade sócio-ecônomicos definidos em Nota Técnica pela ANEEL em 2016.

Esse artigo avalia os resultados das empresas distribuidoras divulgados nos últimos anos utilizando estatística descritiva, análises espaciais e modelos de regressão ordinal logística. O trabalho aborda com profundidade a diferença no desempenho decorrentes de controles público/privados e da localização da empresa.

O Capítulo 4 apresenta o artigo "Multilayer Hybrid Models for the power outage index" submetido à Revista Expert Systems with Applications. O artigo apresenta uma análise multi-camadas para previsão do indicador Duração Equivalente de Interrupção por Unidade Consumidora (DEC). Tenta-se identificar quais as variáveis de maior impacto na explicação do DEC de forma que a CEMIG-D seja capaz de entender como esse indicador reage às suas decisões. Isso permite identificar quais as melhores decisões de gerenciais.

O Capítulo 5 apresenta o artigo "Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico" submetido ao 21º Seminário de Planejamento Econômico-Financeiro e de Regulação do Setor Energético Brasileiro (SEPEF). O artigo apresenta um resumo dos resultados deste trabalho e do P&D-636 vinculado à pesquisa em três frentes: modelo de negócio, modelo para o DEC e modelo para a receita.

O Capítulo 6 apresenta um conjunto de interfaces gráficas web desenvolvidas ao longo da pesquisa. Essas interfaces foram utilizadas na realização da pesquisa e permitem um acesso facilitado às rotinas computacionais desenvolvidas. Este capítulo também apresenta o resumo do trabalho e os direcionamentos para trabalhos futuros.

O apêndice contém dois arquivos. O primeiro apêndice é o artigo intitulado "A novel clustering-based spatial regression model applied to consumer power outage indicator" submetido à Revista Expert Systems with Applications (Apêndice A). Este artigo, no qual consto como terceiro autor, apresenta a metodologia de regionalização utilizada nos modelos híbridos multicamadas. O segundo apêndice é o manual das interfaces gráficas desenvolvidas no projeto (Apêndice B).

Parte II

Artigos submetidos em periódicos

2 Desenvolvimento do Modelo de Negócios de uma Empresa de Distribuição de Energia Elétrica

Resumo

O setor de distribuição de energia elétrica apresenta diversos desafios. A resposta à questão de quem é o cliente abre margem para discussões. Esse fato, somado à complexidade e magnitude das operações do setor e à busca pela eficiência para se enquadrar nos custos operacionais regulatórios definidos pela ANEEL, faz com que a elaboração de um modelo de negócio que seja capaz de elucidar as dinâmicas do mercado e auxiliar na tomada de decisões se torne essencial para as empresas que atuam nele. Entretanto, os estudos a respeito desse tema no setor de energia elétrica ainda são marginais, denotando uma área com potencial pesquisas relevantes. Este trabalho buscou identificar e construir diferentes representações das dinâmicas características de uma empresa distribuidora de energia elétrica, cada uma com objetivos e conclusões únicos. Ademais, consolidou um modelo de negócio para uma empresa de destaque no Brasil no setor de distribuição de energia elétrica, que pode ser utilizado como referência para outras empresas do setor. Esse trabalho contribui de forma inovadora para a discussão da aplicação de modelos de negócio em um setor atípico e inclui análises profundas sobre o setor e o seu funcionamento, assim como representações gráficas inovadoras específicas para esse contexto.

Palavras-chave: Modelo de negócio. Distribuição de energia. Canvas.

2.1 Introdução

O setor elétrico brasileiro passa por um momento de incertezas decorrente principalmente das mudanças políticas, como a discussão desde 2017 do novo marco legal do setor, e dos problemas com reservas hídricas ocorridas nos últimos anos. As chuvas abaixo do esperado nos últimos anos prejudicaram a geração de energia hidrelétrica, a principal fonte do país, responsável por 66% da energia gerada no país (ANEEL, 2016a).

Nesse contexto, as empresas geradoras, transmissoras e distribuidoras de energia elétrica enfrentam um momento crítico em que as decisões estratégicas e sua eficiência operacional são determinantes para sua sobrevivência. Tal realidade é ainda mais delicada e expressiva para as empresas distribuidoras, que são o elo mais sensível da cadeia. O setor de geração se caracteriza como um mercado concorrencial, enquanto a transmissão e distribuição são monopólios e, portanto, estão sob regulação da Agência Nacional de Energia Elétrica (ANEEL) (Doege; Lakoski, 2012). O custo do setor de transmissão é repassado diretamente aos consumidores através da tarifa, enquanto a receita das distribuidoras está diretamente relacionada com a sua eficiência operacional.

A Companhia Energética de Minas Gerais – Distribuição (CEMIG-D) é uma distribuidora que faz parte do maior grupo de distribuição de energia da América do Sul. Tal empresa entrega 10% do mercado brasileiro de energia elétrica, atendendo mais de 8 milhões (10%) do total de clientes do sistema elétrico nacional. A sua área de concessão abrange 567,4 mil km^2, aproximadamente 96% do Estado de Minas Gerais, atendendo 774 municípios e aproximadamente 20 milhões de habitantes através de uma rede de distribuição de 525.224 km (CEMIG, 2012a). A dimensão da operação da CEMIG-D é outra barreira que a empresa precisa considerar ao pensar em competitividade e estratégia.

Dentre as alternativas existentes para dar suporte à CEMIG-D na análise e definição de qual direção estratégica seguir, destaca-se a metodologia de modelo de negócio (business model), que se prova útil tanto para organizações novas/inexistentes quanto para as que já operam há muito tempo (Magretta, 2002). Um bom modelo de negócio permite identificar todas as relações internas e externas de uma empresa, tornando possível traçar estratégias de modo mais objetivo e assertivo.

Assim, o objetivo deste trabalho é desenvolver modelos de negócios para a CEMIG-D que contemplem todas as características relevantes das suas operações. Tais características envolvem desde definições básicas (quem é o cliente da empresa), até questões mais profundas (como a empresa captura valor no mercado, como são as relações entre as diversas variáveis que compõem o seu negócio) (Magretta, 2002). Desse modo, almeja-se em trabalhos futuros utilizar tal modelo de negócios como base para aplicar ferramentas estratégicas e elaborar planejamentos que permitam à empresa aprimorar seus processos internos, otimizar a sua alocação de recursos e sua tomada de decisão — fatores cruciais dado o contexto econômico e ambiental vivido e o porte da empresa.

2.2 Modelo de negócios

2.2.1 Cadeia produtiva da energia elétrica no Brasil

A cadeia produtiva do negócio de energia elétrica no Brasil envolve diversos agentes, conforme demonstra a Figura 2.1. O início da cadeia é composto pelas empresas geradoras de energia elétrica, que podem ser divididas entre aquelas que atuam no mercado concorrencial e as que possuem tarifas reguladas. A energia gerada é transportada para as distribuidoras através dos ativos (instalações) das empresas transmissoras, que atuam em um mercado de tarifa regulada, tendo em vista que o negócio de transmissão de energia elétrica configura um monopólio natural. A energia chega então às distribuidoras, que também atuam em um mercado de monopólio natural, possuem tarifas reguladas e são responsáveis por entregar a energia aos seus clientes.

Os clientes das distribuidoras podem ser divididos em dois grupos: os consumidores cativos e os clientes livres. Os consumidores cativos são aqueles que não possuem a liberdade de negociar a compra da energia; são compostos por consumidores residenciais, comerciais, industriais, poder público, entre outros. Os clientes livres são aqueles que possuem um consumo de alta tensão (carga instalada acima de 3.000 kW) e que têm a possibilidade de negociar a energia tanto com as distribuidoras quanto diretamente com as empresas geradoras. Além desses,

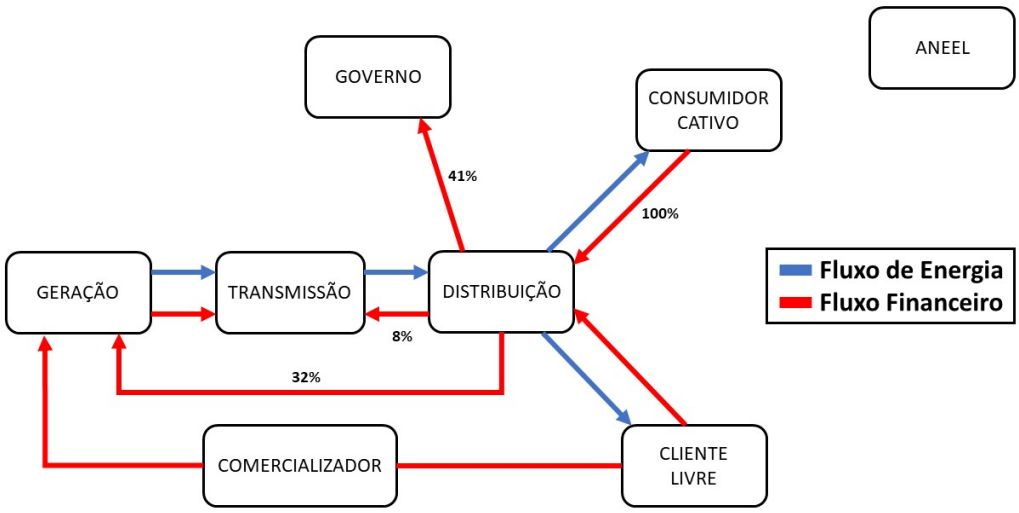


Figura 2.1 – Agentes e fluxo financeiro e de energia da cadeia produtiva de energia elétrica no Brasil.

é importante citar também a participação do estado, dos comercializadores (responsáveis por mediar o negócio entre os clientes livres e as empresas geradoras) e a agência reguladora, que no caso do Brasil é a ANEEL.

A Figura 2.1 também apresenta o fluxo de energia na cadeia (em azul), que segue um caminho linear saindo das empresas geradoras, passando pelas transmissoras, posteriormente para as distribuidoras e enviada para os consumidores cativos e clientes livres. O fluxo financeiro (em vermelho), por outro lado, flui em direções opostas, com a respectiva divisão das receitas. É importante frisar que essa divisão foge ao controle das distribuidoras pois, como exposto anteriormente, elas atuam em um mercado de monopólio natural, de tal forma que a ANEEL é responsável por definir, entre outras atribuições, qual a receita recebida por cada distribuidora.

Nesse contexto, pode-se definir as principais atividades das empresas distribuidoras brasileiras: comprar energia para o atendimento ao mercado cativo; receber a energia em tensão de transmissão; planejar, projetar e construir os ativos de Distribuição; relacionar, ligar e atender a novos clientes; distribuir a energia até os pontos de consumo; entregar, medir e faturar esta energia; faturar e arrecadar, repassando os valores de energia, encargos, tributos e garantir a receita da Distribuidora; operar e manter o sistema elétrico; assegurar a qualidade e continuidade do serviço; remunerar adequadamente os investidores.

2.2.2 Definições de Modelo de negócio

O estudo de modelos de negócio aplicado ao setor elétrico já foi realizado em outros países. Burger e Luke (2017) apresentam diferentes modelos de negócio para sistemas de reposta de demanda e gerenciamento de energia, armazenamento elétrico e térmico, e para energia solar fotovoltaica.

Magretta (2002) define modelos de negócio como histórias que explicam como as empresas

trabalham. Ele compara modelos de negócio com o método científico, no sentido em que inicia com uma hipótese que é testada e então revisada caso necessário. Nesse sentido, a principal força do modelo de negócio como uma ferramenta de planejamento é o seu foco em como todos os elementos de um sistema se encaixam. Ele atenta para que um modelo de negócio não seja construído sobre falsas suposições sobre o comportamento do consumidor, o que resulta em um modelo que é uma solução em busca de um problema. Também é importante diferenciar modelo de negócio e estratégia: o primeiro somente define como as peças de um negócio se encaixam e não considera a existência de competição, que fica a cargo do segundo.

Ovans (2015) faz um resumo de diversas teorias e definições acerca de modelos de negócio. O autor cita que um modelo de negócio possui duas partes: a primeira associada a criação de algo e a segunda com a sua venda. A primeira parte envolve o design, compra de materiais e a sua produção. A segunda parte envolve encontrar e alcançar consumidores, realizar a venda, distribuir um produto ou entregar um serviço. É importante identificar quando um modelo de negócio deve ser atualizado: quando suas inovações criam incrementos cada vez menores, quando se torna difícil desenvolver melhorias e quando seus clientes encontram cada vez mais alternativas no mercado.

Para Casadesus-Masanell e Ricart (2010) deve-se tomar cuidado ao definir modelo de negócio como a lógica de funcionamento de uma firma pois essa definição se confunde com a de estratégia. Portanto, eles definem modelo de negócio como o reflexo da estratégia realizada pela firma. Os autores sugerem a utilização de um framework que integre esses conceitos.

Johnson, Christensen e Kagermann (2008) afirmam que bons modelos de negócios podem modificar indústrias e trazer um grande crescimento, mas que muitas empresas têm dificuldade em aplicar essa metodologia. Os motivos são dois: pouco estudo formal e pouco conhecimento do modelo de negócio atual. Para mitigar esse problema, eles sugerem três ações: identificar uma oportunidade de satisfazer um consumidor, definir um guia de como satisfazer a necessidade obtendo lucro, comparar o modelo novo com o atual e avaliar o tamanho da mudança necessária para capturar essa oportunidade.

Segundo Wirtz et al. (2016), modelo de negócios é a representação simplificada e agregada das atividades relevantes de uma empresa, descrevendo como ela captura valor no mercado. É importante destacar a estrutura conceitual de um negócio definido a partir de três componentes principais: (a) a estratégia de negócio, que ocupa o topo da estrutura, (b) o(s) modelo(s) de negócio(s) e (c) o plano de negócio, conforme Figura 2.2.

A estratégia de negócio envolve uma missão, um posicionamento da empresa em relação ao cenário de negócios. Como concepção, a estratégia aponta para o futuro. O modelo de negócio estrutura a lógica de captura de valor da empresa, oferecendo maneiras coerentes para implantar a sua estratégia. O plano de negócio delineia as ações concretas que colocarão o modelo de negócio em prática.

Como mencionado, a atividade de uma empresa implica na captura de valor no mercado. Inicialmente, entende-se por valor criado como o benefício percebido pelo consumidor, a respeito de um produto, deduzido o custo da empresa para ofertá-lo. O valor capturado é a parcela do

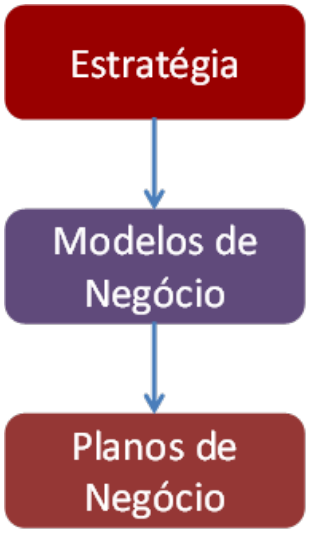


Figura 2.2 – Estrutura hierárquica conceitual de um negócio definida pela estratégia do negócio, o(s) modelo(s) de negócio e o plano de negócio.

valor criado que é capturado pelos agentes envolvidos em uma transação, como as empresas, consumidores e fornecedores, conforme pode ser verificado na Figura 2.3.

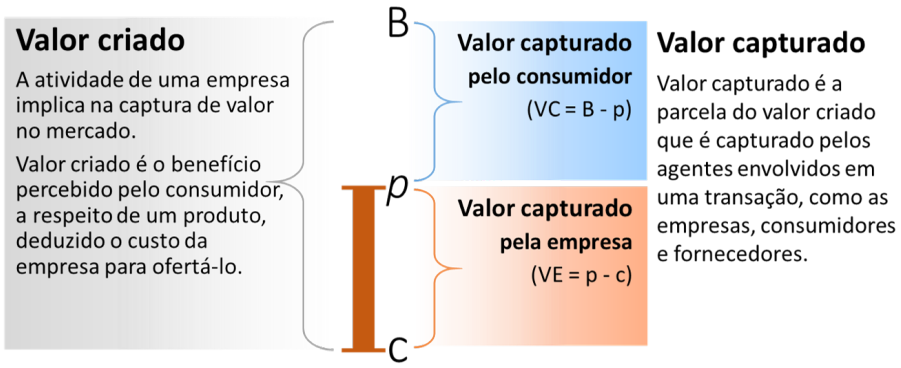


Figura 2.3 – Representação gráfica da diferença entre valor criado e valor capturado.

A relação da criação de valor por uma empresa e o seu modelo de negócio já foi explorado na literatura. Arend (2013) discute como o termo modelo de negócio não deveria mais ser utilizado como a simples descrição das operações de uma empresa e sim focar no modelo de criação de valor. Um modelo de negócio também não deve necessariamente focar em uma atividade (ou negócio) específica de uma empresa. Aspara et al. (2013) apresentam uma definição de modelo negócio para a corporação toda ao invés de somente para um de seus negócios, com uma aplicação prática na Nokia.

Existem, na literatura especializada, diferentes estruturas ou *frameworks* para estruturar, definir ou desenhar um modelo de negócio e a captura de valor. Neste sentido, *frameworks* são estruturas genéricas que apresentam os componentes centrais de um modelo de negócio. Um dos mais conhecidos e utilizados na atualidade é o *Business Model Canvas* (Osterwalder; Pigneur, 2010). Esse *framework* tem como foco a criação de novos modelos de negócio ou a avaliação de

modelos existentes.

Uma empresa pode existir sem um modelo de negócio. A existência de um modelo de negócio, entretanto, agrega valor, influencia e pode aumentar a sua chance de sobrevivência (Kauffman; Wang, 2008; Velu, 2015). A criação ou atualização de um modelo de negócio para uma empresa existente e em operação pode trazer novos desafios. Berends et al. (2016) apresentam duas formas de desenvolver modelos de negócio para empresas estabelecidas: a primeira parte da pesquisa cognitiva que muda posteriormente para a aprendizagem experimental; a segunda no outro sentido.

O *Business Model Canvas* apresenta nove componentes: (i) segmento de clientes, (ii) proposição de valor, (iii) canais, (iv) relacionamento com clientes, (v) fluxo de receitas, (vi) recursos chave, (vii) atividades chave, (viii) parcerias chave e (ix) estrutura de custos. Basicamente os itens (i) a (v) representam o fluxo de receitas do Business Model Canvas, ao passo que os itens (vi) a (ix) representam a estrutura de custos.

O preenchimento do *Canvas* segue uma ordem pré-estabelecida. Inicialmente, define-se o campo "Segmentos de Clientes", que representa grupos de clientes atendidos pela empresa. Procura-se defini-lo com base nas questões: para quem estou agregando valor? e onde estou capturando valor? O segundo campo é o "Proposição de Valor". Neste caso a pergunta chave é: quais demandas eu estou atendendo em cada grupo de clientes? O terceiro é o campo "Canais", que procura identificar como cada proposição de valor é entregue aos clientes. A pergunta chave é: como eu estou alcançando meus clientes para atender suas demandas? O quarto campo é o "Relacionamento com os Clientes", que procura descrever os meios pelos quais os clientes podem contactar a empresa para expor as suas demandas. A pergunta chave é: como minha empresa coleta e registra as mudanças nas necessidades dos clientes? O quinto é o campo "Fluxo de Receitas", que deve indicar como o modelo de negócio da empresa está capturando valor. A pergunta chave é: com o que nossa empresa está ganhando dinheiro? O sexto é o campo "Recursos-Chave", que engloba os ativos necessários para sustentar os componentes do fluxo de receitas. A pergunta chave é: quais são meus recursos-chave? O sétimo campo é o "Atividades-Chave", que forma o core business da empresa. A pergunta chave é: Quais atividades sustentam meu *core business*? O oitavo é o campo "Parcerias-Chave", que deve indicar os parceiros sem os quais o modelo de negócio não se sustenta. A pergunta chave é: quem são os meus parceiros imprescindíveis? Finalmente, o último campo é "Estrutura de Custos", que é determinado pela combinação dos recursos, atividades e parcerias-chave. A pergunta chave é: onde estão meus maiores custos?

Um exemplo de representação do *Canvas* é apresentado na Figura 2.4. Para preenchê-lo recomenda-se reunir um grupo de pessoas que estejam diretamente relacionadas com a operação que se deseja modelar. É importante valorizar nas discussões o conhecimento específico de cada participante. Por exemplo, ao discutir questões contábeis, deve-se dar maior destaque e importância para a opinião do colaborador do setor contábil, mas ponderar também ideias e perspectivas dos demais presentes.

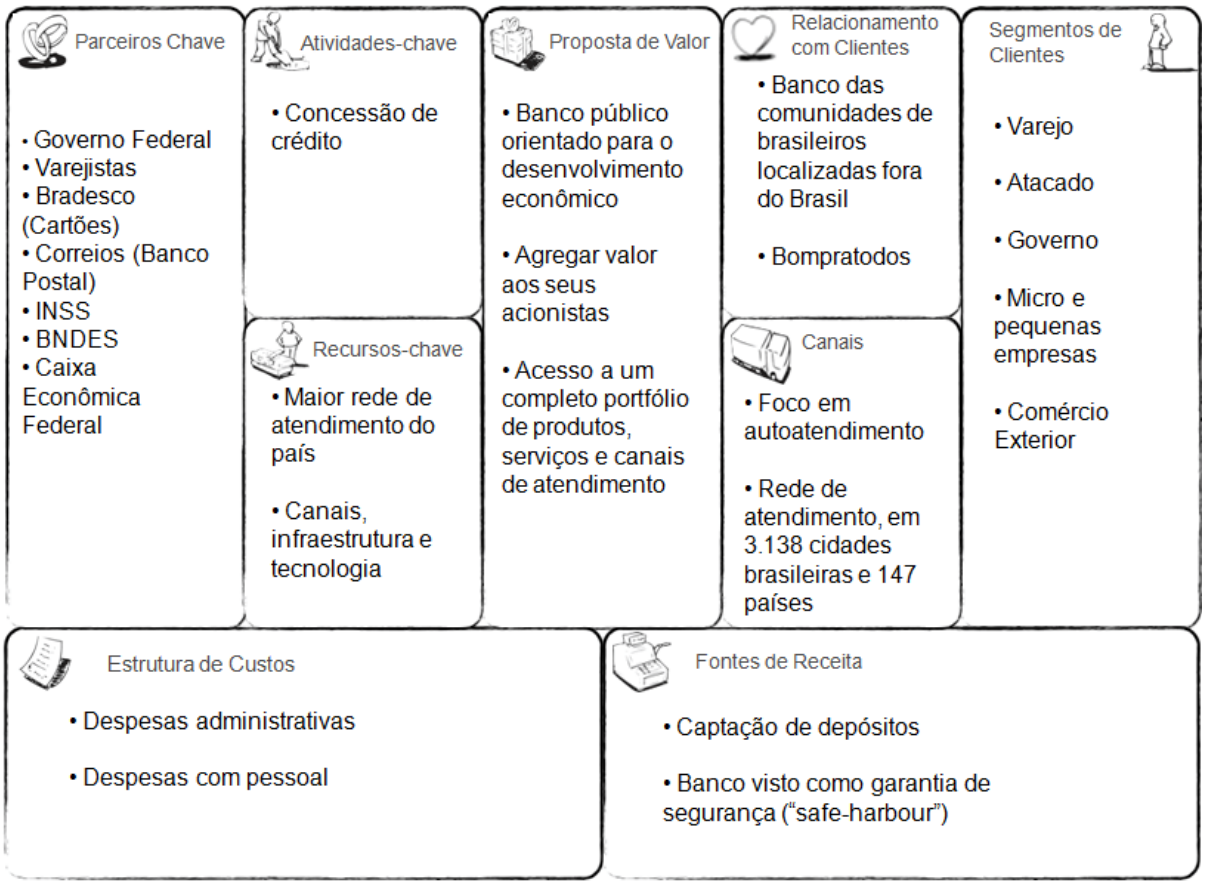


Figura 2.4 – Exemplo de uma ilustração de *framework* Canvas (Teixeira; Lopes, 2016).

2.2.3 Revisão de literatura

Cosenz e Noto (2018) relatam as críticas de diversos pesquisadores sobre o uso de representações estáticas de *business modelling*. Os autores argumentam que é vantajoso combinar esquemas de *Business Modelling* convencionais com sistemas dinâmicos de modelagem (SD - *System Dynamics Modelling*). Simulações são usadas para mapear elementos chave por trás do processo de geração de valor, determinando um sistema de interdependências. Isto permite ao analista aprender ou entender como o negócio reage a mudanças estratégicas. O suporte metodológico do SD é útil para situações envolvendo complexidade dinâmica e incerteza.

Kang, Shin e Lee (2012) apresentam um estudo de um sistema de controle de danos quando um navio sofre um acidente. Uma parte principal no sistema está relacionada às atividades da tripulação. Os autores argumentam que para ser efetivo como um sistema de análise, o *Business Model* deve incluir características relacionadas às tarefas que o sistema planejado suporta. No caso da aplicação do navio, o controle de danos envolve a ação de diversos tripulantes, os quais obedecem a um comandante. Este artigo usa uma abordagem de *Systems Engineering* (SE) para guiar o desenvolvimento de um *business model*. O *business model* foi usado para estruturar o processo de decisão, minimizando o número de alternativas para cada componente do sistema de controle de danos.

Abdelkafi e Täuscher (2016) fornecem uma nova perspectiva sobre *Business Models for*

Sustainability (BMfS) e combina insights da literatura em um modelo conceitual de notação dinâmica. O modelo conecta quatro modelos parciais: a empresa, o ambiente, o tomador de decisões e o cliente. O estudo demonstra o potencial de projetar um loop de realimentação entre os diferentes grupos de clientes, o ambiente e a geração de lucros da empresa.

Bonabeau (2002) propõe o *Agent-based modeling* (ABM), descrito como uma ferramenta poderosa de modelagem e simulação utilizada em diversas aplicações recentes (destaque aqui para problemas reais na área de negócios - *Business Modeling*). O ABM é utilizado em áreas como simulação de fluxo (evacuação de emergência, tráfego e administração de clientes), simulação de organizações (risco e planejamento operacional), simulação de mercado (bolsa de valores, softwares, planejamento estratégico) e simulação de difusão (propagação de inovação e sua dinâmica de adoção).

Liu e Wei (2013) definem o termo *Business Model* através de revisão de literatura e estudos de caso. Mais especificamente, eles apresentam quatro modelos de *business model*: *focused cost innovation*, *integrated cost innovation*, *focused value innovation* e *integrated value innovation*. Os autores discutem como cada tipo de modelo pode ajudar as empresas a se desenvolverem.

Horkoff et al. (2014) propõem um modelo de *Business Intelligence Model* (BIM), apresentando os principais conceitos relacionados (objetivos, situação, influência, indicadores). A metodologia proposta busca ajudar empresas a organizar e tirar conclusões sobre grandes quantidades de dados internos e externos da empresa. Esse modelo tenta capturar fatores internos e externos que afetam os objetivos estratégicos de uma organização e as medidas de performance para alcançar tais objetivos (incluindo análises probabilísticas). Eles desenvolveram uma metodologia que combina estratégias *top-down* e *bottom-up*.

Swan e Ugursal (2009) apresentam uma ampla revisão de técnicas de modelagem aplicada à modelagem do consumo de energia residencial. São revisados os conceitos das modelagens *top-down* e *bottom-up*. A abordagem *top-down* utiliza variáveis macroeconômicas, climáticas, entre outras. A principal vantagem da abordagem *top-down* é o uso de dados agregados, amplamente disponíveis. A abordagem *bottom-up* procura modelar o consumo de cada unidade, utilizando modelos estatísticos ou modelos de engenharia.

Castro et al. (2017) apresentam um estudo, utilizando metodologias estatísticas, para identificar qual o conjunto de indicadores que melhor caracteriza a sustentabilidade econômico-financeira de uma concessionária de distribuição de energia elétrica. O principal indicador encontrado foi a Margem Operacional Recorrente, que representa a razão entre o EBITDA Recorrente e a Parcela B Regulatória. Segundo os autores, empresas que apresentam capacidade de gerar um resultado operacional superior ao contemplado na tarifa ou possuem grande base de ativos remunerada via tarifa tendem a ter uma performance econômica e financeira melhor que as demais.

Kraus, Feuerriegel e Oztekin (2018) avaliam o estágio atual da aplicação de técnicas de *deep learning* a problemas de *business modeling*. Os autores descrevem as principais arquiteturas de redes neurais utilizadas hoje em dia e descrevem suas aplicações em três estudos de caso: gestão de risco, predição de uso de recursos e predição de vendas.

Zott, Amit e Massa (2011) apresentam uma extensa e elaborada revisão de literatura no tema *business models*. Analisando uma amostra final de 103 publicações em revistas relevantes da área, os autores concluem que não há um consenso sobre a definição de *business model*. Pelo contrário, há uma concentração nítida de definições. São apresentados quatro grandes grupos: (a) *business model* como uma nova unidade de análise; (b) *business model* como um sistema de níveis, uma abordagem holística para explicar como as empresas realizam seus negócios; (c) atividades fins das empresas desempenham um papel importante nas várias conceptualizações de *business model* propostos; (d) *business model* procura explicar como o conceito de valor (de um produto ou processo) é criado e capturado pelas empresas.

Teece (2010) procura entender a significância do *business model* e explorar suas conexões com *business strategy*, inovação e teoria econômica. O artigo destaca o foco no consumidor e as tecnologias facilitando a troca de informação e a busca por soluções. O autor indica que o conceito de *business model* não tem embasamento teórico dentro da economia. Os elementos principais de um *business model* seriam: (i) seleção de tecnologias e característica próprias do produto/serviço, (ii) determinação dos benefícios para o consumidor, (iii) identificação do segmento de mercado alvo, (iv) confirmação dos gastos, (v) planejamento de mecanismos de captura de valor.

Morris, Schindehutte e Allen (2005) apresentam uma revisão de literatura e tiram conclusões sobre a definição, natureza, estruturação e evolução temporal dos *business models*. O estudo propõe uma abordagem para caracterizar os modelos de negócios, composta por três níveis de tomada de decisão (fundação, propriedade e regras). Os autores indicam que não há uma definição única para *business model*, cujo conceito pode ser confundido com *business strategy*, *revenue model* e *economic model*. A definição mais simples diz que *business model* é o modelo econômico de uma empresa com lógica focada na geração de lucro com sustentação ao longo do tempo. No caso geral, alguns componentes que integram um *business model* são: (i) valor da empresa, (ii) modelo econômico, (iii) relacionamento com clientes, (iv) papéis das parcerias, (v) infraestrutura interna e (iv) mercados alvo.

Teixeira e Lopes (2016) aplicam o *business model Canvas* para descrever o modelo de negócio do Banco do Brasil e da Caixa Econômica Federal. São utilizados os relatórios da administração no período de 2002 a 2012. Segundo os autores, o *business model Canvas* (Osterwalder; Pigneur, 2010) considera que um modelo de negócios deve ser simples, intuitivo e relevante e procura descrever como uma organização cria, entrega e captura valor. Ainda segundo os autores, o modelo de negócio é uma representação abstrata dos elementos-chave de um negócio: o que será vendido (proposta de valor), a quem será comercializado, quais são os processos essenciais para o desenvolvimento do produto/serviço (incluindo a estrutura de custos) e como ocorrerá a interação mercadológica entre empresas e clientes.

Osterwalder e Pigneur (2012) explicam por que e como a pesquisa em Sistemas de Informação (SI) contribui para estudos estratégicos, e o estudo evidencia a necessidade de aumentar a compreensão sobre modelos de negócio. Segundo os autores, o problema essencial que organizações enfrentam com relação às suas estratégias de evolução não é o de escolher entre vários modelos de negócio existentes. A questão de fato central é a falta de um processo que permita alternativas para modelos inteiramente novos e viáveis. Nesse sentido, os autores

promovem a ideia de que a pesquisa em Sistemas de Informação é adequada para questões de estudos estratégicos, devido ao seu enfoque histórico nas técnicas de design e metodologia.

Hoptroff (1993) argumenta o uso de redes neurais do tipo *multilayer perceptron* (mlp) para previsão de séries temporais e modelagem de negócios (*business modelling*). Em seu estudo de caso, uma rede mlp com 4 neurônios foi utilizada para modelar a rentabilidade de empresas. As vantagens do uso de redes neurais são a capacidade de extrapolação; robustez a ruídos, mau condicionamento e dados insuficientes; e facilidade de uso.

2.3 Metodologia

Realizou-se um levantamento bibliográfico de artigos e livros sobre o tema de modelo de negócios para criar uma primeira definição dos conceitos referentes ao tema. Em seguida foi feito um estudo sobre o setor de energia elétrico brasileiro para entender todas as particularidades e características dos diferentes agentes que o compõem.

Para o desenho do modelo de negócio da CEMIG-D, foi organizada uma pesquisa qualitativa abrangendo um grupo focal de 12 técnicos e gerentes das áreas de regulação e operação da empresa. Os técnicos foram selecionados de forma a abranger todos os setores da CEMIG-D com interesse direto na construção do modelo de negócio, possibilitando que as discussões englobassem diferentes perspectivas. O grupo focal participou de um treinamento de 16 horas durante o qual foram apresentados os conceitos introdutórios sobre Modelos de Negócios e o *framework Canvas*. Entrevistas semi-estruturadas foram conduzidas utilizando formulários on-line para definir os ciclos eficientes de operação, os nove campos do *framework Canvas* e o diagrama de relações. Após o preenchimento dos formulários, as respostas foram discutidas e sintetizadas pelo grupo focal durante o treinamento.

Posteriormente, o *framework Canvas* foi apresentado a equipes de gerentes e superintendentes das áreas de regulação, tarifas, investimentos, entre outras áreas, e as alterações e atualizações pertinentes foram executadas, até que se chegasse ao formato final da ferramenta.

2.4 Resultados

2.4.1 Ciclos eficientes

Com base nos conceitos de modelo de negócio de uma empresa de distribuição de energia, a Figura 2.5 apresenta o ciclo eficiente de operação do modelo de negócio da CEMIG-D. Esta representação busca demonstrar quais os ganhos que a empresa de distribuição pode obter ao realizar mais e melhores investimentos. A construção do ciclo elucida as dinâmicas cruciais da empresa, e catalisa o preenchimento assertivo do *business model Canvas*, apesar de não estar diretamente atrelado à metodologia de modelos de negócios.

Para a compreensão do ciclo operacional eficiente da CEMIG-D torna-se necessário definir alguns termos técnicos do setor. O indicador de Duração Equivalente de Interrupção por Unidade Consumidora (DEC) contabiliza a quantidade média de horas que um cliente fica sem energia elétrica durante o mês; o indicador de Frequência Equivalente de Interrupção por Unidade

Consumidora (FEC) registra em média quantas vezes o cliente teve seu acesso à rede elétrica interrompido; a Duração Relativa da Transgressão de Tensão Precária (DRP) indica o percentual de tempo no qual a unidade consumidora permaneceu com tensão precária; a Duração Relativa da Transgressão de Tensão Crítica (DRC) indica o percentual do tempo no qual a unidade consumidora permaneceu com tensão crítica; a Base de Remuneração Regulatória (BRR) consiste no total de investimentos feitos pelas distribuidoras em ativos e operações que será restituído pelas tarifas pagas pelos clientes; a Base de Remuneração Líquida (BRL) é definida como a BRR deduzida de alguns valores contábeis, representando o real valor decorrente de suas operações que será recebido pelas distribuidoras; a Quota de Reintegração Regulatória (QRR) contabiliza a depreciação e amortização dos investimentos feitos de modo a repor os ativos utilizados na prestação do serviço ao final da sua vida útil; o OPEX representa todos os custos operacionais executados pelas empresas; o CAPEX representa todos os investimentos realizados pela empresa; os Lucros Antes de Juros, Impostos, Depreciação e Amortização (LAJIDA – em inglês EBITDA) representam o lucro da empresa apenas em função dos custos operacionais, sem descontar os impostos e demais despesas; as Despesas Não Reconhecidas (DNR) pelo regulador são despesas não restituídas pela ANEEL e são formadas principalmente por multas contratuais, multas por demissões, provisões e perdas/baixas de bens.

Idealmente, um maior investimento (CAPEX) permite: alcançar um maior fornecimento de energia; melhorar a qualidade dos indicadores de DEC e FEC; aumentar a BRR e BRL; aumentar a QRR e a remuneração do capital; aumentar a receita; reduzir o OPEX; maior LAJIDA. Esse incremento no LAJIDA, que ocorre em função do aumento da receita e redução dos custos, melhora a capacidade de reinvestir da empresa, fechando um ciclo virtuoso.

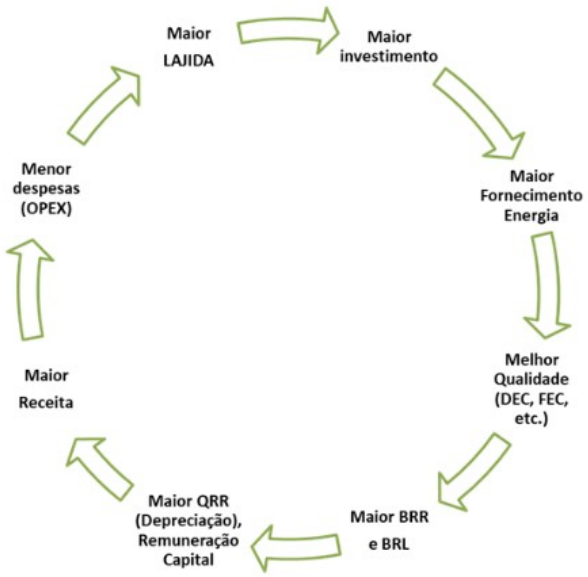


Figura 2.5 – Modelo ótimo de funcionamento de uma concessionária distribuidora de energia elétrica.

A partir desse ciclo eficiente de operação (Figura 2.5), podem ser criados ciclos complementares, como demonstrado na Figura 2.6. Esses ciclos possuem a finalidade de detalhar os ganhos decorrentes em algumas componentes do ciclo principal. Nesse caso, foram escolhidas as

três componentes consideradas mais importantes do modelo de negócio da distribuidora: maior CAPEX, maior receita e menor OPEX.

O ciclo complementar do OPEX demonstra que para alcançar menores despesas deve-se: contratar Serviços de Terceiros em menor quantidade e com maior qualidade; gerar compensações financeiras reduzidas devido às atividades de gestão de rotina diária de atendimento e serviço; reduzir as DNR; reduzir a frequência de serviços emergenciais (como por exemplo vandalismo, abalroamento, roubo, acidente, objeto na rede, defeito cliente afetando outros, ligação clandestina, interferência de terceiros) e comerciais (como por exemplo atendimentos por erro de leitura, cobrança por irregularidade, suspensão indevida, cadastro e alteração cadastral).

O ciclo complementar do CAPEX indica que para realizar um maior investimento deve-se: realizar uma melhor estratégia de gestão dos ativos; disponibilizar uma maior capacidade de fornecimento; realizar uma melhor O&M (Operação e Manutenção) dos Ativos; executar melhor gestão da base de ativos. Por fim, o ciclo da receita apresenta que para alcançar uma maior receita deve-se: aumentar a blindagem dos consumidores; reduzir perdas não técnicas; reduzir a inadimplência dos consumidores.

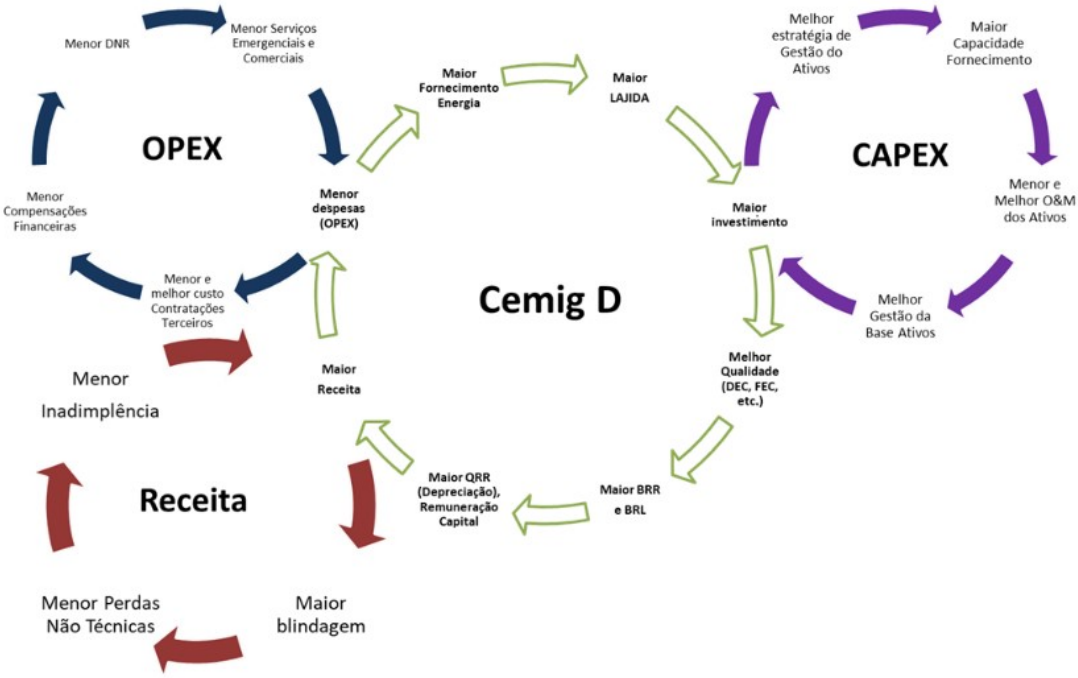


Figura 2.6 – Ciclo marginal do OPEX, CAPEX e Receita do modelo de negócio da CEMIG-D.

De forma geral, pode-se afirmar que uma empresa distribuidora possui como negócio principal (*core business*) a gestão intensiva de ativos. Os ativos impactam na BRR e, portanto, na remuneração do capital e remuneração regulatória. Um ativo mal gerido impacta negativamente na gestão do OPEX e na satisfação dos consumidores ao aumentar os custos de operação, piorar os indicadores de continuidade (DEC/FEC) e de qualidade de energia (DRC/DRP), aumentar gastos com cobertura de danos e reparações em instalações dos clientes, além de aumentar o número de reclamações, depreciar a imagem da concessionária, entre outros.

Nesse contexto, o pior cenário para uma distribuidora é aquele no qual ela possui um ativo

depreciado, mal mantido e em uso, por ele não gerar remuneração do capital, gerar altos gastos com manutenção corretiva e piorar os indicadores regulatórios — piorando consequentemente as compensações financeiras, o que pode em casos graves, causar até mesmo perda de concessão.

A gestão da rotina diária de uma distribuidora de energia é feita através de complexos processos comerciais, de Operação e Manutenção (O&M) e de suporte (administração, jurídico, pessoal, TI etc.). A gestão desses processos impacta diretamente na Gestão dos Ativos e todos devem atender os requisitos legais, regulatórios e empresariais definidos pelo governo, ANEEL e investidores. Conclui-se então que a excelência na operação de uma distribuidora é decorrência de uma boa gestão de ativos e dos processos inerentes ao negócio. Assim, define-se que o principal indicador de eficiência operacional deve ser o LAJIDA (Lucro Antes dos Juros, Impostos, Depreciação e Amortização) (ou EBITDA). O trabalho de Castro et al. (2017) também indica que o LAJIDA é o principal indicador para avaliar a eficiência econômico-financeiras das distribuidoras, ratificando as ideias apresentadas.

Além disso, há também a gestão da concessão junto à agência reguladora. Atualmente, a manutenção da concessão está ligada à manutenção de indicadores de sustentabilidade econômico-financeiras em níveis adequados. A gestão da concessão das distribuidoras sob a ótica dessa sustentabilidade econômico-financeira deve ser norteada pela percepção de todas as áreas no resultado operacional da empresa (LAJIDA). Todas as áreas das distribuidoras devem conhecer os impactos dos seus processos e atividades no LAJIDA e traçar planos de ações no intuito de aproximar da meta e evitar ficar acima do limite regulatório de gastos.

2.4.2 Desenvolvimento do business model Canvas

Com relação ao entendimento de cada um dos nove campos do *framework Canvas*, uma intepretação foi definida para a CEMIG-D, a partir da análise, compreensão e discussão do grupo focal de técnicos e gerentes, e está representada na Figura 2.7. Um diagrama do modelo de negócio relacionando as principais atividades e variáveis da CEMIG-D foi desenvolvido com base nesse *framework Canvas*. O principal objetivo deste diagrama é apresentar como se relacionam as diversas receitas, despesas e variáveis presentes nas suas operações.

Grande parte da discussão se concentrou na questão do segmento de clientes. Como já mencionado anteriormente, a CEMIG-D, assim como todas as outras distribuidoras no Brasil, são gestoras dos ativos elétricos (e.g. postes, subestações, transformadores) que na prática pertencem à União. As empresas distribuidoras são contratadas pela União para gerir tais ativos e fornecer energia elétrica para a sociedade. Por esse ponto de vista, pode-se depreender então que a União é o cliente das distribuidoras, já que elas são contratadas para prestar um serviço. Entretanto, após diversas discussões com o grupo focal, foi decidido que o principal segmento de clientes das distribuidoras são os clientes livres e cativos presentes ao final da cadeia, conforme apresentado na Figura 2.1. Pode-se dizer que esse segmento do *Canvas* foi o que mais consumiu tempo das discussões pois a definição do cliente da empresa define como todo o restante deve ser preenchido. Uma definição incorreta nessa primeira etapa compromete qualquer resultado que possa ser extraído da ferramenta.

Após definir os clientes, o restante do preenchimento ocorreu sem muitos imprevistos,

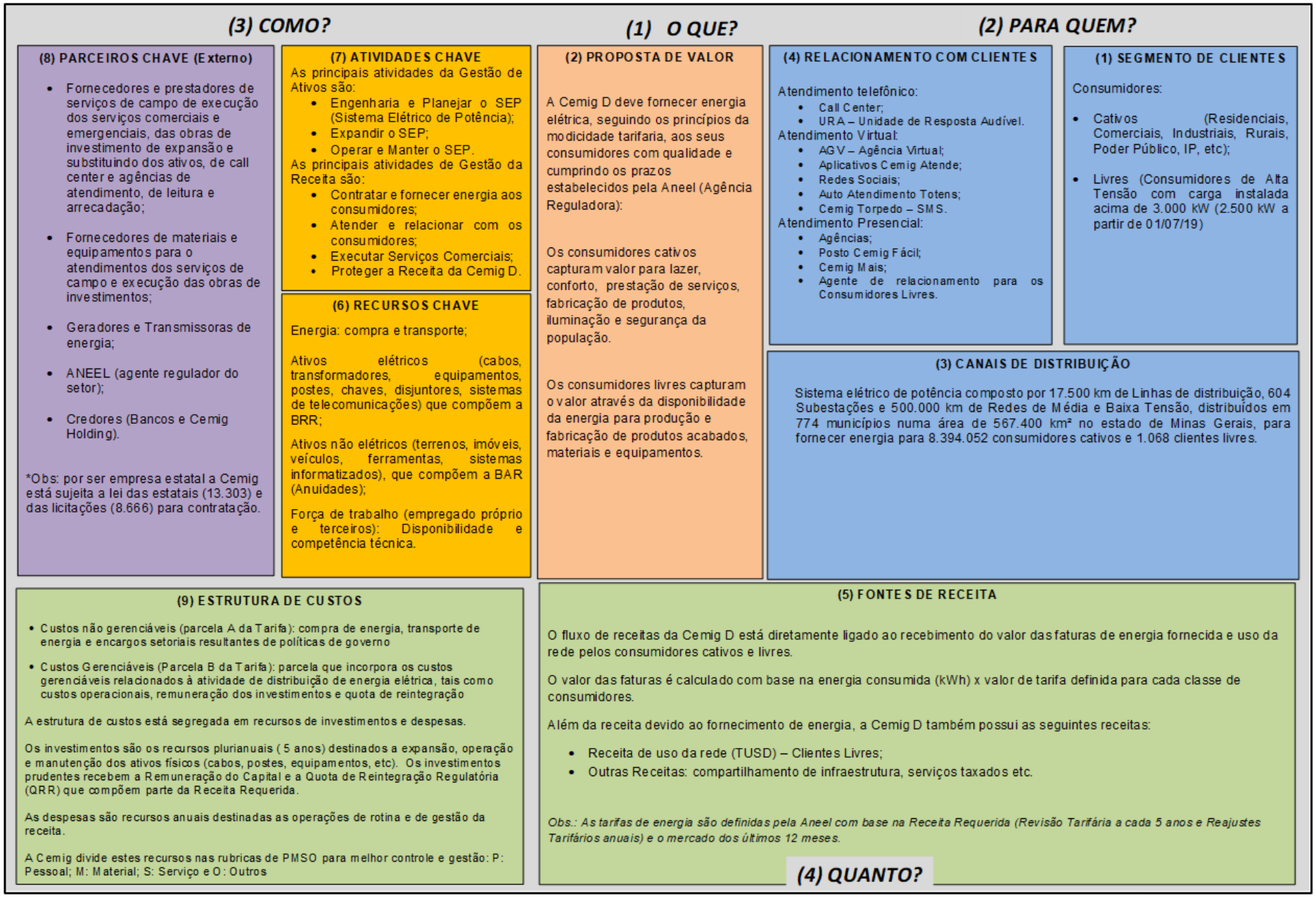


Figura 2.7 – O Modelo de Negócio da CEMIG-D.

mas vale destacar o campo Proposta de Valor. Em um primeiro momento a proposta pode ser definida de forma simples como "entregar energia". Entretanto, ao analisar mais profundamente, chegou-se à conclusão de que a proposta de valor da empresa envolve fornecer lazer, conforto, segurança, entre outros, para os seus clientes. A versão final do *Canvas* permitiu entender melhor o funcionamento do negócio da CEMIG-D com todas as suas atividades e variáveis.

De forma geral, pode-se dividir as atividades e variáveis da CEMIG-D em cinco grandes áreas, apresentadas na Figura 2.8. São essas: Receita Requerida; Investimentos; Variáveis vinculadas à Receita; Variáveis vinculadas ao OPEX; Variáveis vinculadas ao setor financeiro. Todas as cinco áreas estão relacionadas entre si direta ou indiretamente, ou seja, impactos em um campo reverberam em todos os demais. Dessa representação depreende-se que a divisão da Receita Requerida entre os investimentos (CAPEX), OPEX e Receitas são os principais fatores que impactam nos resultados financeiros da empresa.

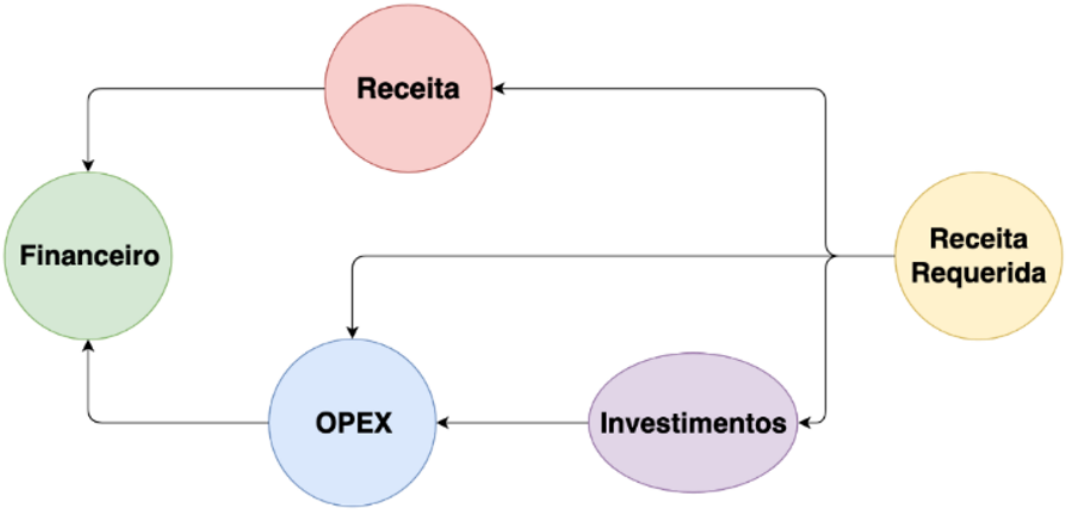


Figura 2.8 – Representação macro das variáveis do modelo de negócio da CEMIG-D.

A Figura 2.9 expande essa representação para todas as relações entre esses blocos e as variáveis relevantes. O bloco da Receita Requerida, apresentado em detalhes na Figura 2.9, reúne as 12 variáveis que formam a receita que a ANEEL permite que as distribuidoras arrecadem, calculado de forma que as permita cobrir seus custos e realizar os investimentos necessários. Nela impactam diretamente: Compra de Energia; Gastos com Transporte/Encargos; Perdas regulatórias; Receita Irrecuperável (RI); Outras Receitas; Ajustes Financeiros; PMSO Regulatório (CAOM); Remuneração de Capital (OE); QRR; Custo Anual das Instalações Móveis e Imóveis (CAIMI); Ultrapassagem de Demanda; Excedente Reativo.

O bloco de investimentos é formado pelas nove opções que compõem o portfólio de CAPEX da empresa. A receita obtida pela OE e pela QRR devem ser reinvestidas na empresa para renovar os seus ativos, garantir a sua expansão e implementar novas tecnologias em suas operações. Esse portfólio é composto por: Reforço; Mercado/Expansão; Perdas; Medição; Automação; Telecom; Melhoria da Qualidade; Segurança; O&M.

O grupo das variáveis vinculadas à receita envolve todas as etapas do processo que impactam diretamente na receita obtida pela distribuidora. Ela aborda desde a compra de

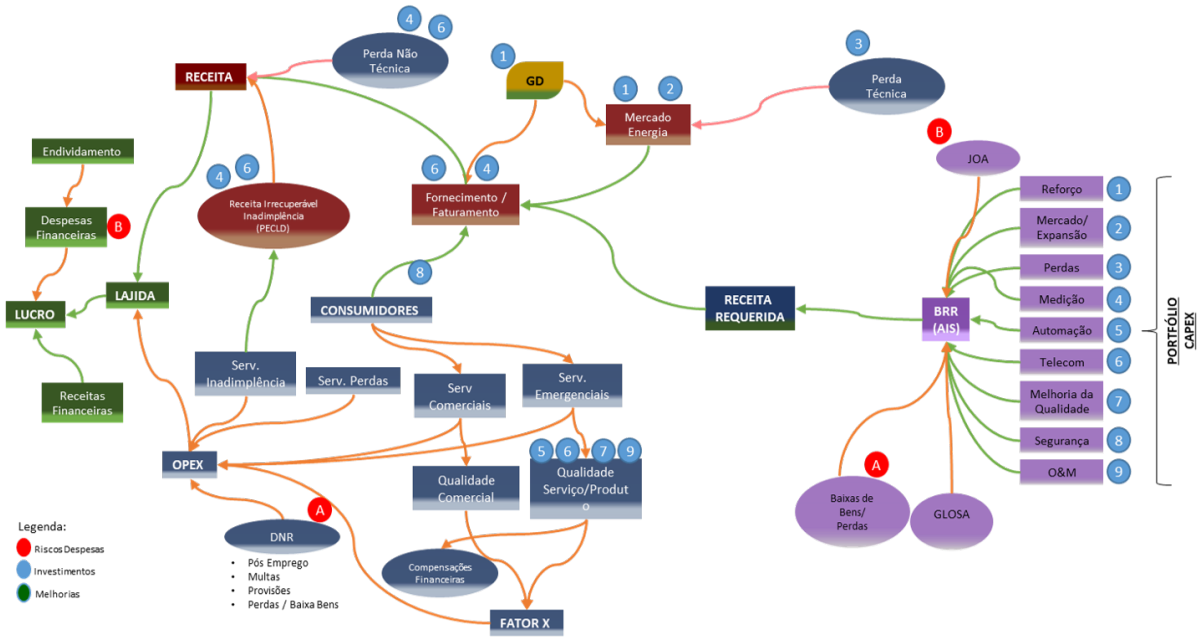


Figura 2.9 – Representação das principais variáveis do modelo de negócio da CEMIG-D.

energia no mercado, o fornecimento/faturamento das contas, as tarifas, a geração distribuída, a arrecadação e por fim o cálculo da receita real obtida.

O grupo das variáveis vinculadas ao OPEX envolve os próprios consumidores, os serviços emergenciais, comerciais, de perda e de inadimplência, a qualidade do serviço/produto e comercial, compensações financeiras, Fator X, as DNR e o cálculo do custo operacional real.

Por fim, o último bloco envolve as variáveis vinculadas ao setor financeiro e que tem como principal indicador o LAJIDA. Com base nesse lucro calculado, a CEMIG-D consegue calcular as suas receitas e despesas financeiras e o nível de endividamento desejado.

Fica claro após essa explanação, que a gestão de um negócio de distribuição de energia elétrica é extremamente complexa. A empresa precisa lidar com diversas variáveis, algumas controláveis e outras não, atender aos objetivos regulatórios definidos pela ANEEL e fornecer um serviço de qualidade para a sociedade. Alcançar esses objetivos sem o auxílio de ferramentas estatísticas e computacionais exige um esforço massivo, sem que haja garantia de sucesso. Assim, espera-se que esses *frameworks* possam servir como ponto de partida para o desenvolvimento de ferramentas que deem suporte a empresa nessa missão.

2.4.3 Modelo final (consolidado)

Com base no Canvas apresentado na Figura 2.7 e na representação interna da empresa apresentado na Figura 2.9, um modelo coeso e conciso de relação de variáveis foi construído, o qual está apresentado na Figura 2.10. Esse modelo tem dois objetivos: ser uma representação mais simples/eficiente daquela apresentada na Figura 2.9; expor a relação direta entre as variáveis quantificáveis e o LAJIDA. Na representação visual do modelo, é possível perceber que o LAJIDA (3) é diretamente relacionado às variáveis Receita (1) e OPEX (2). Essas representações, por sua vez, são compostas por uma série de variáveis relacionadas a diferentes esferas que influenciam o

resultado da operação da distribuidora, que podem ser quantificadas, e, portanto, usadas como base para análises e modelagens estatísticas.

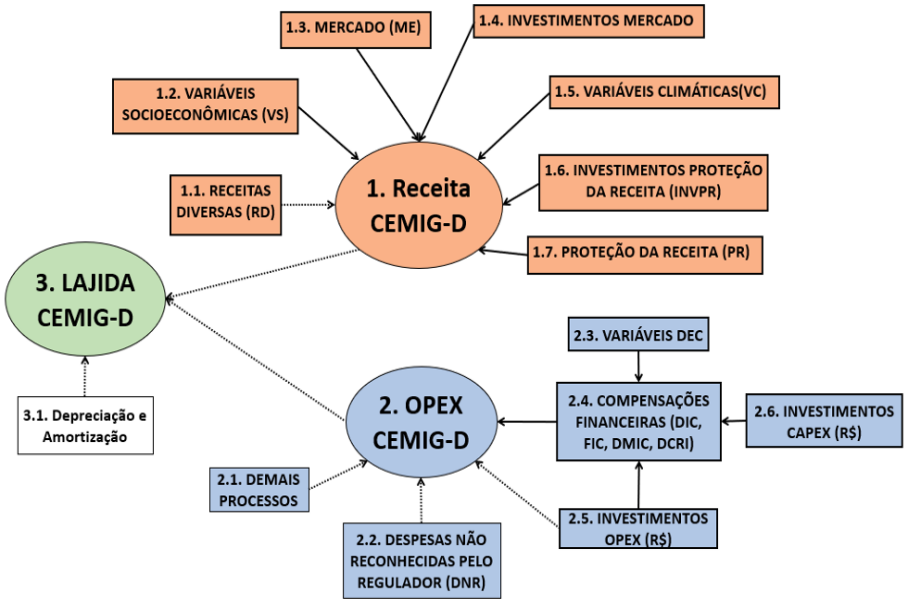


Figura 2.10 – Representação do modelo final.

A Receita é composta por 7 variáveis econômicas, sociais, ambientais e financeiras. Essa representação busca englobar diferentes questões mercadológicas que podem ter influência na geração de receita da CEMIG-D; as demais fontes de receita não relacionadas com a distribuição de energia são agrupadas em Receitas Diversas.

A principal variável na definição do OPEX, por sua vez, é Compensações Financeiras. Essas compensações são decorrentes de falhas na distribuição de energia e correspondem a maior parte do gasto com OPEX pela empresa. As Compensações Financeiras são afetadas diretamente pelos investimentos da empresa em OPEX/CAPEX e pelo indicador DEC — Duração Equivalente de Interrupção por Unidade Consumidora. O DEC indica o tempo médio que um cliente ficou sem fornecimento de energia durante um mês. Isso implica em uma alta relação entre o valor pago em compensações e o valor do DEC, o que denota o motivo de sua grande relação com os valores pagos em Compensações Financeiras. Portanto, considerou-se que as variáveis que impactam no indicador DEC também impactam na variável Compensações Financeiras. Essas variáveis envolvem questões climáticas (e.g. temperatura), financeiras, operacionais (e.g. serviços emergenciais), de infraestrutura (e.g. extensão da rede), mão-de-obra, quantidade de clientes etc. As outras duas variáveis que compõem o OPEX da CEMIG-D são as Despesas não reconhecidas pelo regulador (DNR) e as demais são agrupadas em Demais Processos.

Por fim, o LAJIDA é calculado como a Receita subtraída do OPEX (somado à Depreciação e Amortização, uma correção contábil). Dessa forma, ao invés de tentar modelar diretamente o LAJIDA da empresa, o modelo é capaz de elucidar as diferentes partes que irão interagir para gerar o resultado, e identificar quais são as variáveis operacionais, sociais, econômicas, ambientais, dentre outras, que exercem influência e são capazes de impactar o resultado obtido. Portanto, com esse modelo, torna-se possível compreender as principais dinâmicas existentes dentro da

operação da empresa, e é possível definir e coletar as variáveis necessárias para a execução de análises do modelo de negócio da CEMIG-D através de seu LAJIDA. Ademais, viabiliza-se a elaboração de um modelo de predição das variáveis "Receita" e "OPEX", em função do qual a empresa poderá perceber os impactos e mudanças geradas em função de alterações hipotéticas em investimentos, gestão, fatores socioambientais, alocação de recursos, dentre outros.

2.5 Conclusão

A pesquisa foi conduzida, e posteriormente validada, integralmente em conjunto com membros da CEMIG-D, que participaram, discutiram e opinaram em todas as etapas. Como primeiro resultado desse trabalho, foi desenvolvido um ciclo de eficiência operacional da empresa, que determina de forma simples e direta o que deve ser feito para que as suas operações sejam sustentáveis e prósperas. Outros três ciclos marginais foram desenvolvidos, descrevendo o que a companhia deve fazer para alcançar os resultados desejados.

Em seguida buscou-se criar um *framework Canvas* para o seu modelo de negócio. Essa etapa se mostrou bastante desafiadora, tanto para os pesquisadores quanto para os facilitadores da CEMIG-D, ao questionar padrões que eram considerados óbvios ou que nunca haviam sido discutidos com tanta profundidade. Ao término da sua montagem foi possível obter uma visão objetiva sobre as operações e o negócio da CEMIG-D, como por exemplo a noção de que a empresa é mais do que tudo uma gestora dos ativos da União.

Foi construído um diagrama do modelo de negócio da organização, com base no modelo *framework*, contendo as áreas que constituem o seu negócio: Receita Requerida; Receita; Investimentos; OPEX; Financeiro. Esse diagrama foi consolidado em outro mais detalhado, evidenciando todas as variáveis que integram as suas atividades e as suas relações. Essa representação torna possível, futuramente, a construção de um banco de dados relacional que permita fazer análises estatísticas, simulações, séries temporais, regressões etc.

É importante ressaltar que este modelo proposto no trabalho é puramente teórico, ou seja, ainda não foi realmente testado em nenhuma situação prática. É bastante provável que no futuro a representação do modelo de negócios da CEMIG-D sofra alterações em função da inserção de novas variáveis ou relações não encontradas ainda. Além disso, o setor passa por uma constante revolução tecnológica e regulatória, o que exige que atualizações frequentes sejam executadas no modelo para evitar sua defasagem.

Acredita-se que o resultado encontrado seja aplicável em outras empresas distribuidoras além da CEMIG-D. É provável que as variáveis definidas no modelo tenham importância e impactos diferentes para cada empresa, principalmente em razão das diferenças entre os seus portes. Ainda assim, o modelo se apresenta bastante genérico, e com as devidas adaptações, pode ser visualizado como uma representação verossímil e fidedigna do modelo de negócio de uma gama abrangente de empresas distribuidoras de energia do Brasil.

3 An analysis of the financial-economic sustainability in Brazil's electricity distribution sector

Abstract

Background: The socioeconomic development of Brazil can be compromised by an inefficient and ineffectively arranged energy sector. Improvements in Brazilian energy generation, transmission and distribution systems in a sustainable manner are crucial. The National Electric Energy Agency (ANEEL) developed a 16-indicator based sustainability framework to appraise the financial-economic sustainability status of the Distribution Systems Operators (DSOs). We have developed a broad analysis of the 8 most important of those indicators during 8 years. The goal is to assess how the sector has evolved and what factors might be contributing to performance discrepancies.

Methods: The analysis is carried out using different graphical tools and statistical models. For each indicator, a performance panel containing graphics divided by sector, DSOs, region and control type is presented. This enables analysis regarding how time, control type and regional aspects impact on DSOs' results. Performance evaluations are shown using four groups (quartiles), following ANEEL's method: best performance, good performance, inferior performance and worst performance. Then, an ordinal logistic regression model was developed for each indicator. Last, probability for each quartile, considering each region and control type, was estimated.

Results: A novel graphical analysis for the 8 most important indicators set by ANEEL is presented. Estimated probabilities are calculated, showing which characteristics are more impactful for each indicator. Results show the major challenges facing publicly-controlled DSOs, and DSOs located in specific regions, to achieve good performance. This suggests latent differences between control types and among the five regions of Brazil.

Conclusions: The sustainability framework developed by ANEEL to evaluate the financial-economic sustainability of DSOs is considered adequate. It lacks some adjustments for comparing different control types and regions. Sector results are worrisome and should be closely supervised by ANEEL, especially results of public DSOs and DSOs located in the northern region. The analyses and models presented in the present study can be adapted to evaluate the financial-economic sustainability of new indicators related to sustainability performance.

3.1 Background

The energy sector has long been considered a fundamental part of the development of a country. Agenda 21, an environmental action plan for the 21st century, was issued at the 1992 Rio de Janeiro Earth Summit. It had a goal to end poverty through better management of energy and

natural resources (Meakin, 1992). More recently, in 2015, the United Nations presented the 2030 Agenda for Sustainable Development, with 17 Sustainable Development Goals (SDG) (Cf, 2015). The seventh goal refers to modern, reliable, sustainable and fairly-priced energy distribution for everyone. The economic and social development of a country can be seriously compromised if the population has insufficient access to modern energy services (Kaygusuz, 2012).

Many studies have examined the relationship between energy and (sustainable) development (Oyedepo, 2012; Rösch et al., 2018; Begić; Afgan, 2007; Vera; Langlois, 2007). The methodology commonly used to study that correlation often includes frameworks (Kemmler; Spreng, 2007; Neves; Leal, 2010; Vithayasrichareon; MacGill; Nakawiro, 2012; Sanders et al., 2014) and indicators (Sarangi et al., 2019; Streimikiene; Siksnelyte, 2016; Sharma; Balachandra, 2015; Rösch et al., 2017). Frameworks define important dimensions regarding energy and sustainability. Those dimensions generally include economic, social, environmental and regulatory aspects, among others. Indicators are usually measurable, quantitative variables related to the dimensions chosen previously.

The dimensions in a sustainability framework change according to the research interest. There are some broader dimensions (e.g., economic, social, environmental), while others are more specific. One can include the institutional dimension (Sharma; Balachandra, 2015), technical and institutional dimensions (Iddrisu; Bhattacharyya, 2015), or replace the social aspect with reliability and technical dimensions (Prete et al., 2012).

Even though two frameworks may have similar dimensions, they may comprise different indicators. One may evaluate the economic dimension based on the total installed capacity (Sarangi et al., 2019), while another may focus on the import/export volume of electricity (Streimikiene; Siksnelyte, 2016). The social aspect can be measured by the access-use matrix poverty rate (Kemmler; Spreng, 2007) or by the employment level (Sharma; Balachandra, 2015). Environmental level can be measured by reduction of CO2 emissions (Karger; Hennings, 2009) or the annual emissions of NOx (Prete et al., 2012).

3.1.1 Financial-economic sustainability framework in Brazil

The National Electric Energy Agency (ANEEL) is responsible for regulating the electricity sector in Brazil. In 2016, ANEEL proposed a framework regarding the financial-economic sustainability of the Distribution System Operators (DSO) (ANEEL, 2016b). The proposal includes 16 indicators divided into 7 dimensions: indebtedness, efficiency, investments, profitability, pay-out ratio, operational and renewed concession agreements. Each dimension comprises 1 to 4 indicators; each indicator comprises 1 or more variables. Variables can either be quantitative (absolute or percentage) or binary. Among the 16 indicators, 8 are considered important enough to be assessed quarterly (Table 3.1); the others are measured but not assessed. These 8 indicators are assessed by dividing the DSOs into 4 groups: best performance, good performance, inferior performance and worst performance. The agency uses two methods to evaluate the DSO indicators. The first method is applied only to indicator 1. Indebtedness level from 0 to 7 is considered the best. Between 7 and 14 is considered good. Between 14 and 50 is considered inferior, and values greater than 50 are considered the worst. Results for the 7 remaining indicators are defined by

distributing the DSO sample among four quartiles. Each quartile comprises thirteen to fourteen DSOs (except for years with missing data). The best 25% of DSOs are assigned to the first quartile, the next best 25% are assigned to the second quartile, the next 25% are assigned to the third quartile and the last 25% are assigned to the fourth. This assignment is based on the ordinal performance ranking of the population, i.e., the position in which each DSO is placed when its performance is compared to the others. This method has two main implications. First, unlike what happens to indicator 1, the number of DSOs in each quartile is always similar. Second, there are no constant thresholds between quartiles, as they are set according to all DSO performances in each year. This means that, even if a DSO improves its performance from one year to the next, it is not guaranteed that this DSO will be assigned to a better quartile, as the majority of DSOs might also improve, increasing and shifting the quartile thresholds.

The financial-economic sustainability reports have been published quarterly since 2016, and a database containing data from all DSOs has been published since 2011. The available data allows a deep analysis of DSOs' performances year by year. Even though the reports contain information regarding DSOs' control types (public or private) and locations (by region), there is no control type or regional based evaluations. There is also no application of any statistical methods to assess the impacts of each DSO individualities in the results.

3.2 Literature review

3.2.1 Sustainability frameworks and indicators

Several studies have approached sustainability in the energy sector. Most analyze the energy sector using a framework developed specifically for the respective approach, usually with a single country or a region as the object of study. Brown e Sovacool (2007) developed an Energy Sustainability Index (ESI) based on four dimensions: oil security, electricity reliability, energy efficiency and environmental quality. Those four dimensions comprise twelve indicators which were used in a USA database from 1970 to 2005. The authors emphasize the importance of indexes/indicators and a solid, manageable database.

Ateba e Prinsloo (2019) analyzed South African energy sustainability using a qualitative lens rather than a quantitative method. The authors evaluate regulatory policies and innovate using geographical data to support the results.

Suganthi (2020) performed a literature review about energy and water indicators. A list of 48 indicators is provided and divided into four dimensions: economic, social, environmental and institutional. Results show that each country must have their own indicators, suited to the nature of the available resources. A country's indicators also evolve over time along with its needs. The author used a multi-criteria, fuzzy analytical, hierarchical processing-data envelopment analysis model with data from 48 countries. Thrän et al. (2020) analyzed the sustainability of the biogas sector in Germany. The focus of their study is how regulation, governance and laws impact the energy sector.

Ateba, Prinsloo e Gawlik (2019) used questionnaires to assess the impact of electricity supply sustainability on South African industrial growth. Results reinforce the need for a

Table 3.1 – Dimensions and indicators defined by ANEEL for sustainability analysis of DSOs

Dimension	Number	Indicator	Description
Indebtedness	1	$\frac{RND}{EBITDA - RRQ}$	Ability to fulfill debt commitments by indebtedness (Regulatory Net Debt - RND), cash flow (Earnings Before Interest, Taxes, Depreciation and Amortization - EBITDA) and minimum investments (Regulatory Reintegration Quota - RRQ)
Efficiency	2	$\frac{EBITDA}{Reg\ PBV}$	Margin of resources (EBITDA) remaining over the entire portion of the tariff (Regulatory Parcel B Value - Reg PBV) assigned to the DSO
	3	$\frac{OPEX}{Reg\ OPEX} - 1$	Level of cash generation due to Operational Expenditures (OPEX)
Profitability	4	$\frac{EBIT - Reg\ EBIT}{NPB}$	Profitability through the difference between performed and regulatory Earnings Before Interest and Taxes (EBIT) over the Net Pay Basis (NPB)
Operational	5	$Global\ Continuity\ Performance\ Index\ (GCPI)$	DSO performance in relation to the regulatory continuity level energy supply
	6	$Performed\ Loss\ (\%) - Regulatory\ Loss\ (\%)$	DSO performance in relation to loss management through the difference between performed and regulatory loss
	7	$Market\ (GWh)\ CAGR$	Market growth in Gigawatt-hours (GWh) through Compound Average Growth Rate (CAGR)
	8	$Consumers\ CAGR$	Market growth in consumers through Compound Average Growth Rate (CAGR)

sustainable energy sector to ensure a country's industrial growth.

Prete et al. (2012) analyzed the sustainability of microgrids using a four-dimension framework, considering: environmental, economic, technical and reliability. Each dimension comprises two to four indicators, with a total of 11 indicators. The authors simulate six scenarios with different conditions, assessing the impact of each scenario on the sustainability index.

Farquharson, Jaramillo e Samaras (2018) assessed the impact of a reliable energy sector on sustainability in sub-Saharan countries. Results show that countries with unreliable energy sectors rely on fossil fuel and diesel backup generators. This leads to a struggle to achieve long term sustainability goals, such as seen in the Sustainable Development Goals (SDG) (Cf, 2015).

Özbuğday, Öğünlü e Alma (2016) analyzed the impact of privatization on the sustainability of Turkish electricity distributors and suppliers. Four indicators are examined: transparency, financial soundness, quality and competitive supply market. The authors conclude that political intervention may be harmful to the development and efficiency of the energy sector.

Karger e Hennings (2009) evaluated sustainability in regard to decentralized electricity generation. The authors created a framework with three levels, which they nominated the "value tree". The main goal of the value tree is to achieve sustainability, and its levels contain all the criteria considered important to accomplish the objective. The first level is divided into five areas (or dimensions): environmental protection, health protection, security of supply, economic aspects and social aspects.

May e Brennan (2006) developed a framework for the sustainability of the Australian electricity sector. The model has three dimensions: environmental, economic and social. The focus was to analyze the impact of different fuels on sustainability in the country using 21 indicators.

Millan, Lora e Micco (2001) performed a profound analysis on the impact of regulation and privatization on the electricity sector of Latin America. At the time, Brazil was among the world leaders in private investment in the electricity sector. The sector's three largest firms had a 40% market share (the whole private sector has a 60% market share participation). Privatization led to improvements in companies' efficiency, boosting better prices and quality, and reducing losses. Even so, the authors also highlight the need for regulatory authorities in the electricity sector to ensure the coordination towards an equilibrium between demand and supply capacity.

Kemmler e Spreng (2007) proposed an energy system framework as a proxy to assess sustainable development in India. The framework comprises three dimensions (economy, environment and society), with a total of eight indicators. All indicators are related to the electricity sector (e.g. energy consumption and fuel based emissions) and are used to evaluate the poverty in the country. The authors used data from 1983 to 2000, and developed a projected scenario for 2025.

Neves e Leal (2010) developed a framework for local energy sustainability indicators with three dimensions: environmental, economic and social. Indicators are selected following a seven-step methodology, from a pool of more than 100 energy-based indicators. The authors highlight the importance of using indicators to diagnose the energy companies' current situation. They also emphasize the importance of using indicators as planning tools; i.e., using indicators

as decision criteria.

Vithayasrichareon, MacGill e Nakawiro (2012) developed a sustainability framework that targeted the five largest consumer countries in Asia. The authors use a framework based on the 3A's sustainability objectives created by the World Energy Council: accessibility, availability and acceptability. Each objective unfolds in two dimensions, namely: affordable price and energy services (accessibility); short-term reliability of supply and long-term reliability of supply (availability); safety and greenhouse emissions (acceptability). Each dimension comprises 1 to 4 indicators; 18 in total.

Iddrisu e Bhattacharyya (2015) propose a Sustainable Energy Development Index (SEDI). This index is based on a framework with eleven indicators, divided into five dimensions: technical, economic, social, environmental and institutional. The authors calculated the SEDI for every country that had available data in 2009. The proposed index was then compared with other traditional indexes, such as the Human Development Index (HDI), Energy Development Index (EDI) and Multidimensional Energy Poverty Index (MEPI). The authors found a positive correlation between the Sustainable Energy Development Index and both HDI and EDI, confirming that a country's social development is closely related to its energy system development.

Sharma e Balachandra (2015) developed a multi-hierarchical and multi-dimensional indicator framework for the electricity sector of India. The authors' approach consists of dimensions which are divided into themes, subthemes and composite indicators. Indicator values are calculated for the bottom level; each higher level is calculated as the combination of the respective bottom levels. This approach results in a National Electricity System Sustainability Index (NESSI), a single measure for the overall sustainability performance of the Indian electricity sector.

Streimikiene e Siksnelyte (2016) approached the sustainability aspect of the energy sector through the lens of market characteristics. The authors create a three-dimension framework (economic, environmental and social), with 13 indicators applied to 12 countries. Results indicate that the economic growth of a country is related to opening the market. Also, the authors found a positive correlation among Gross Domestic Product (GDP), sustainability and a diversified electricity balance structure.

Rösch et al. (2017) created a system (framework) of sustainability indicators based on normative values for the German energy system. The system has 4 dimensions: securing human existence, maintaining society's productive potential, preserving society's options for development and action, and instrumental sustainability rules. Comprising a total of 45 indicators, the authors' objective is to encourage deeper discussion about sustainability and broaden the usually chosen set of indicators.

Sarangi et al. (2019) performed a quantitative analysis on the sustainability of the Indian electricity sector. The authors use a three-dimension framework (economic, social and environment), with eleven indicators to represent them. The analysis covers data from twelve states collected over a decade. Statistical methods are used to evaluate the historical performance of each state and the individual result of each dimension. States are also divided into groups according to three different status: reform, human development and economic status. This analysis

allows the authors to identify the impact of social and political decisions on the energy sector.

Ordinal Logistic Regression Model

Methods for evaluating indicators also vary, according to the authors. Studies usually choose quantitative indicators to calculate the mean, standard deviation, median or mode. Another option is to choose qualitative indicators, evaluated through the use of questionnaires. Some authors may choose more complex statistical methods to evaluate sustainability.

For example, Doukas et al. (2012) use Principal Component Analysis (PCA) to assess energy sustainability in rural communities. The authors did not define dimensions, but they did use variables such as population density, energy consumption, per capita GDP and fossil fuel consumption in their analysis. Tavassoli, Ketabi e Ghandehari (2020) developed Network Data Envelopment Analysis (NDEA) to assess the sustainability of Iran's electricity distribution network (generation, transmission and distribution).

Both approaches are interesting and can be properly applied to quantitative variables. The regulatory indicators set by ANEEL are also quantitative, but they have a crucial specificity: seven out of eight indicators are not evaluated based on the absolute value of each observation. Rather, what matters is the order of observations, which is needed to determine to which quartile each observation will be allocated. So, it is possible to consider those indicators as categorical variables, as ANEEL is interested in ordinal data rather than cardinal data.

Ordinal logistic responses are largely used in many situations with categorical responses, such as medical and epidemiologic studies. There are cases where the dependent variable (Y) represents not only a category, but also the levels of a measurement scale, such as Likert scales (Harrel, 2001). Xu et al. (2020) used an ordinal logistic regression model to identify which of 53 variables best represented illness severity level: moderate, severe and critical. Jayawardena, Epps e Ambikairajah (2020) also used an ordinal logistic regression model to represent depression in both the eight-item Patient Health Questionnaire and the Beck Depression Index II scales. Fuks e Salazar (2008) used an ordinal logistic regression model to analyze the household electricity consumption class in Rio de Janeiro. In ANEEL's framework, the dependent variable (indicator result) varies from 1 (first quartile) to 4 (fourth quartile). It means that the dependent variable is not only categorical, but also has a natural ordering in its levels. In this case, it is most appropriate to use the ordinal logistic regression model, also known as the proportional odds model (POM) (Powers; Xie, 2008). This model incorporates ordering through several logit transformations of the response probabilities (Bilder; Loughin, 2014). Let us consider that the cumulative probability for category (quartile) j of Y is $P(Y \leq j) = \pi_1 + \cdots + \pi_j$ for $j = 1, \ldots, J$. The relation among the independent variables $(x_1, \ldots, x_p)$ and the log-odds of cumulative probabilities can be analyzed using a regression model for ordinal multinomial responses

$$\begin{aligned} logit(P(Y \leq j)) &= log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) \\ &= log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right) \end{aligned} \tag{3.1}$$

The POM is a particular case with two peculiarities. First, the logit of these cumulative probabilities changes linearly along with the independent variables. Second, the slope of this relationship does not change regardless of the category j. This results in

$$\begin{aligned} logit(P(Y \leq j)) &= \beta_{j,0} + \beta_1 x_1 + \cdots + \beta_p x_p, \\ j &= 1, \ldots, J-1 \end{aligned} \tag{3.2}$$

The model assumes the same effects β for each category j with respect to the predictor variable, but each category has its own intercept β (Agresti, 2018). Thus, there is no need for subscripts on the β because of the constant relationship between the independent variables and the categories. The name "proportional odds" comes from the fact that each of the odds is a multiple of $\exp(\beta_{j,0})$.

Let us now consider a fixed category j. If we increase the independent variable x_r by c units, it changes all log-odds in Equation 3.2 by $c\beta_r$, holding all other independent variables constant. This also implies that the difference in the log-odds between categories (quartiles) j and j' $(\beta_{j,0} - \beta_{j'0})$ is constant, and is not changed by the independent variables $(x_1 + \cdots + x_p)$ supposing they are held constant. Equation 3.1 shows that the odds for each category increase as category j increases. This happens because there is a greater probability in the numerator $(P(Y \leq j))$, e.g. $\beta_{1,0} < \cdots < \beta_{J-1,0}$.

The probabilities for observing a particular dependent variable j are calculated as

$$\begin{aligned} \pi_j &= P(Y = j) \\ &= P(Y \leq j) - P(Y \leq j-1) \end{aligned} \tag{3.3}$$

where $P(Y \leq 0) = 0$, $P(Y \leq J) = 1$; and

$$P(Y \leq j) = \frac{exp(\beta_{j,0} + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_{j,0} + \beta_1 x_1 + \cdots + \beta_p x_p)} \tag{3.4}$$

For example, the probability for the 2nd quartile is

$$\begin{aligned} \pi_2 &= P(Y \leq 2) - P(Y \leq 1) \\ &= \frac{exp(\beta_{2,0} + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_{2,0} + \beta_1 x_1 + \cdots + \beta_p x_p)} \end{aligned} \tag{3.5}$$

and the probability for category J is

$$\begin{aligned} \pi_J &= P(Y \leq J) - P(Y \leq J-1) \\ &= 1 - \frac{exp(\beta_{J-1,0} + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + exp(\beta_{J-1,0} + \beta_1 x_1 + \cdots + \beta_p x_p)} \end{aligned} \tag{3.6}$$

3.3 Methods

This data and statistical analysis used to prepare the present paper are based on the regulatory financial-economic sustainability framework developed by ANEEL (2016b). The framework has eight indicators that have been divided into four dimensions: indebtedness, efficiency, profitability and operationality. A literature review was carried out to evaluate the framework, and to compare the present analysis with similar frameworks. The literature review identified different approaches to this problem. Examples of these approaches include qualitative, quantitative and geographical analyses; a regulatory approach; and, statistical measures and models. The methods chosen for this paper are: analysis per control type, year, DSO, and region; along with a statistical model for each indicator. Results for the indicators consist of four quartiles, which are defined based on the results obtained by the DSOs. For indicator 1 (indebtedness), the raw value obtained by a given DSO defines to which quartile the company will be designated. For the other seven indicators, a quartile is assigned to a DSO considering its placement on the performance ranking that includes all DSOs. Results are usually presented using a four-color representation, in which each color specifies 1 of the 4 quartiles. From best to worst, they are: green (first quartile), yellow (second quartile), orange (third quartile) and red (fourth quartile).

Data was gathered from reports published quarterly by ANEEL at the Financial-economic sustainability repository (ANEEL, 2016f). Data from 2011 to 2017 comes from the third quarter report of 2018, while data from 2018 comes from the first quarter report of 2019. In addition to the indicators collected for each of the 51 DSOs, there is also information about the control type (private or public) and region (northern, northeastern, central-western, southeastern and southern). The dataset contains data for 51 Brazilian DSOs, of which 36 are privately controlled and 15 are publicly controlled. The DSO segmentation per region is as follows: 7 DSOs are from the northern region, 11 DSOs are from the northeastern region, 5 DSOs are from the central-western region, 14 DSOs are from the southeastern region, and 14 DSOs are from the southern region. The geographical information (shapefile) for each DSO was gathered from the Electric Sector Geographic Information System (SIGEL) (ANEEL, 2017). The map of Brazil, segmented by regions, is shown in Figure 3.1.

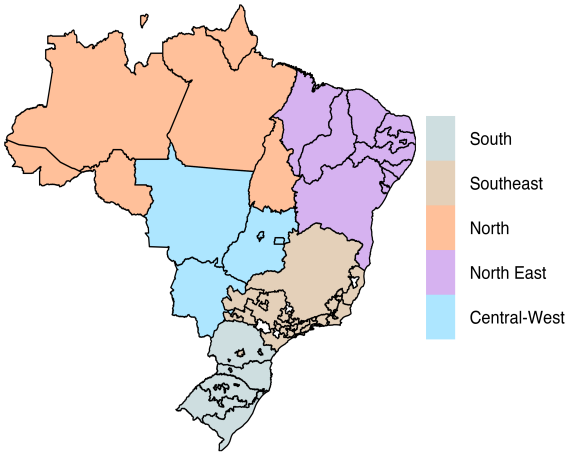


Figure 3.1 – Map of Brazil divided in regions

General results for each control type were calculated using the mean and median of the

data. Results were also evaluated for the sector. Values of the indicators for the sector were calculated as the mean for all DSOs. For each indicator, the quartile for the sector was determined by evaluating which quartile contained the closest value calculated using the mean of all DSOs. Results are shown in a collection of graphics, each more suitable for a specific visualization of the data: lines, maps, jitter plots, box plots, radar plots and bar plots.

Eight ordinal logistic regression models were developed and analyzed in the present project. The models evaluate the impact of region, year and control type on the quartile result for a DSO. The models were validated using statistical inference for each predictor indicator pair.

The R software was used to generate all images for the statistical analysis (Team, 2021). The next section presents the results and discussion for each indicator, as well as for the proposed statistical model.

3.4 Results

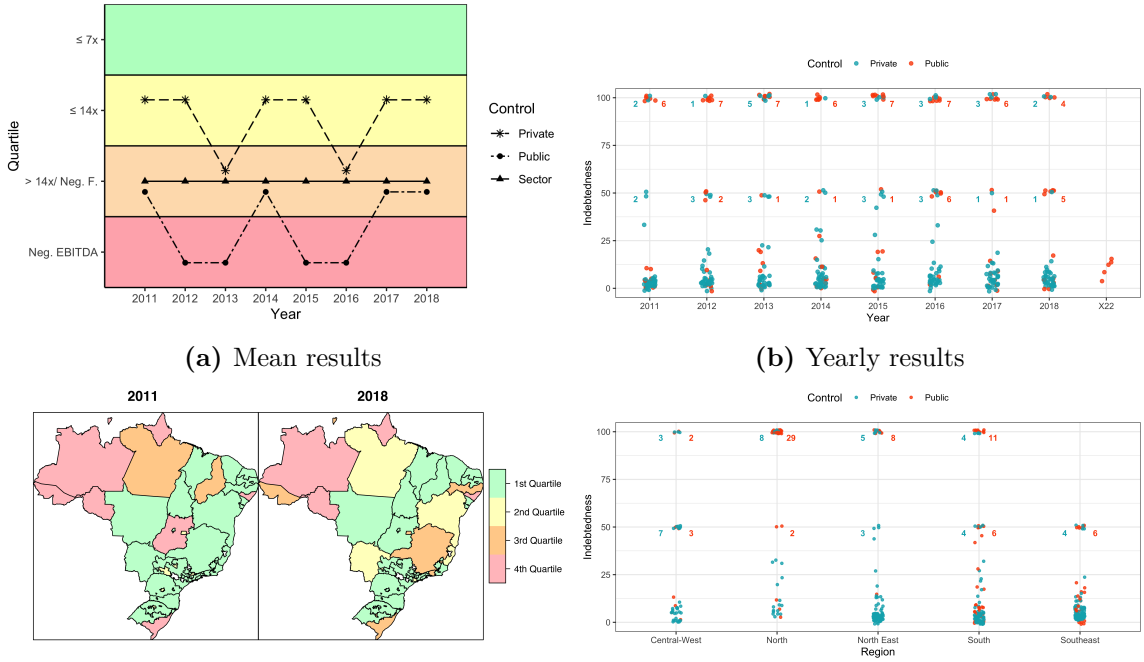
Indicators are divided into four quartiles. Indicator 1 is divided into its quartiles based on the raw calculated values: values from 0 to 7 are assigned to the first quartile (green); values greater than 7 up to 14 are assigned to the second quartile (yellow); values greater than 14 up to 50 are assigned to the third quartile (orange); DSOs with negative EBITDA are assigned to the fourth quartile (red). ANEEL assigns a value equal to 50 to an indicator if a DSO has a negative cash flow (EBITDA - RRQ); and, ANEEL assigns a value of 100 to an indicator if a DSO has a negative EBITDA. Results for the seven other indicators are divided into four quartiles based on the ordinal placement of each DSO. Thresholds between quartiles are calculated considering the data dispersion of each indicator. Each quartile comprises 13 to 14 DSOs. The best 25% of DSOs are assigned to the first quartile (green), the next 25% best of DSOs are assigned to the second quartile (yellow), the next 25% of DSOs are assigned to the third quartile (orange), and the last 25% are assigned to the fourth quartile (red). Some indicators were not available for all DSOs in all years. Implications are mostly seen in the geographical results, with some unfilled (white) regions.

3.4.1 Indicator 1 - Indebtedness

Indicator 1 evaluates the indebtedness level of the DSOs (the lower, the better). Results are presented in Figura 3.2. Figura 3.2a shows the mean results for all private and public DSOs, as well as the general result for the sector. Privately controlled DSOs fluctuate between the second and third quartiles, while publicly controlled DSOs fluctuate between the third and fourth quartiles. Private DSOs present better results than public DSOs throughout the whole period. The sector displays a constant result in the third quarter throughout the whole period. Figura 3.2b presents a jitter plot of the individual results for each DSO from 2011 to 2018. The number of DSOs at levels 50 and 100 are depicted beside the respective points. Private DSOs are concentrated below an indebtedness level of 25. Public DSOs are concentrated at higher levels of indebtedness, with half or more points above an indebtedness level of 25.

Figure 3.2c presents the geographical results for each individual DSO in 2011 and 2018.

This enables the analysis of how geographical results changed from the first observed period to the last observed period. Results in 2011 show that most DSOs assigned to the fourth quartile were located in Brazil's northern region, while all other regions showed results mostly in the first quartile. These patterns partially changed in 2018. The northern region maintains its results in the fourth quartile, but some DSOs improved their results and reached higher quartiles. Other regions, such as the southeastern, show a reduction of DSOs in the first quartile. Results for each region are presented clearly in Figure 3.2d. Figure 3.2d shows the jitter plot, divided by region, for all eight years considered. The northern region shows almost all DSOs assigned to the fourth quartile. The southeastern region does not show any DSOs in the fourth quartile. All regions contain more private than public DSOs in the first and second quartiles.



(a) Mean results

(b) Yearly results

(c) Geographical results

(d) Regional results

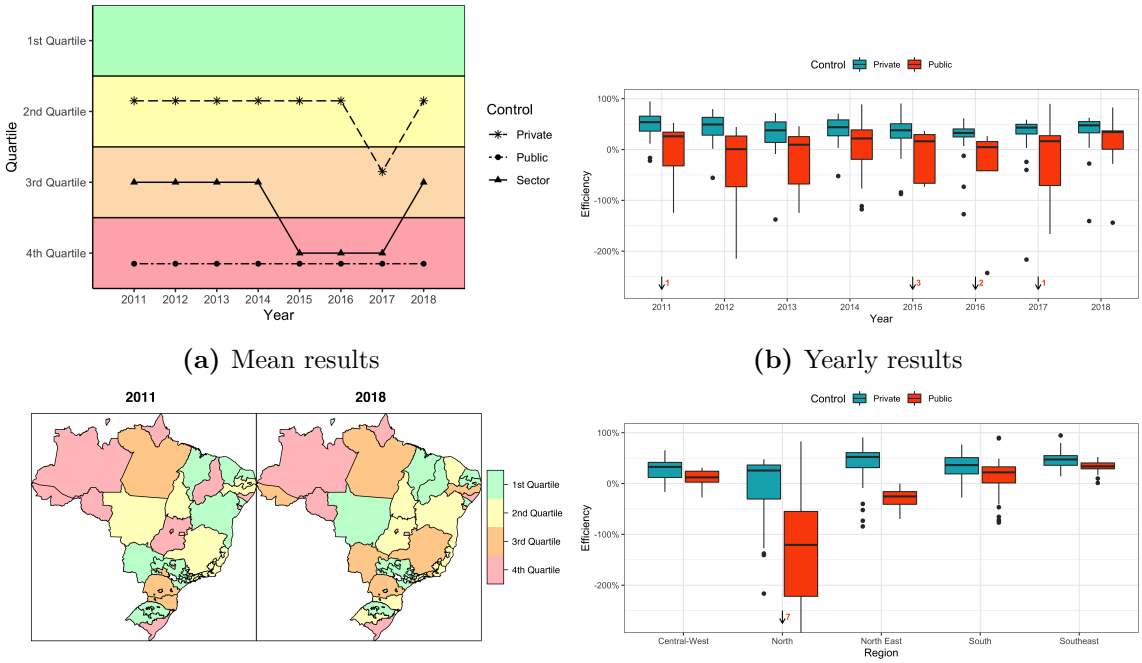
Figure 3.2 – Results for indicator 1

3.4.2 Indicator 2 - Efficiency

Indicator 2 evaluates the efficiency of the DSOs (the higher, the better). Results are presented in Figure 3.3. Figure 3.3a presents the mean results for private and public companies, and the sector. Public DSOs are constantly in the fourth quartile, while private DSOs are in the second quartile for every year but 2017. The sector's result, as expected, is located between private and public results, and is assigned to either the third or fourth quartile. This result shows how impactful public results are. From 2015 to 2017 the sector result was allocated to the worst quartile, meaning that the average of all DSOs is in the worst 25% of performances.

Figure 3.3b presents a boxplot that confirms these results. Most private DSOs show positive efficiency levels, while most public DSOs feature negative efficiency levels. It is also clear that there are impactful negative outliers (observations under -200%) in 2015 (3), 2016 (3), and 2017(2), which presumably were decisive for the sector's awful performance in those years.

Figure 3.3c presents the geographical results for each DSO. In 2011, all regions, except the southeastern region, had at least one DSO in the fourth quartile. This fact remains true in 2018, even though the results for that region did not improve. DSOs in the third and fourth quartiles are mostly in the northern region in 2011 and 2018. The discrepancy between private and public DSOs is shown in Figure 3.3d. While the southeastern, southern and central-western regions show relatively close values for private and public DSOs, the northeastern and northern regions show great differences among the results for each control type. However, it is noticeable that private DSOs have higher medians for every region and year (Figure 3.3b), suggesting better overall performances.



(a) Mean results

(b) Yearly results

(c) Geographical results

(d) Regional results

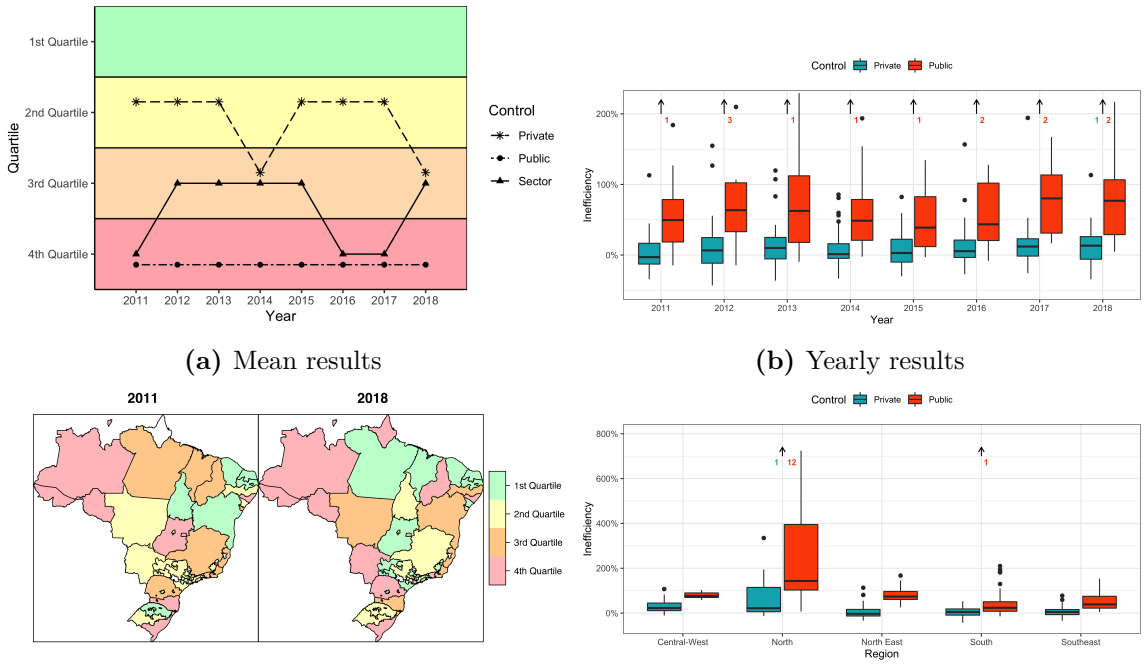
Figure 3.3 – Results for indicator 2

3.4.3 Indicator 3 - Inefficiency

Indicator 3 evaluates the inefficiency of the DSOs (the lower, the better). Results are presented in Figure 3.4. Figure 3.4a presents the results for private and public DSOs, as well as the sector. Public DSOs present a constant result in the fourth quartile for the entire period. Private DSOs present results in the second quartile, with two exceptions, located in the third quartile. The sector fluctuates between the third and fourth quartiles. This difference is shown more clearly in Figure 3.4b. Private DSOs present lower levels of inefficiency, and have no extreme outliers (above 150%) for most years. Public DSOs show higher inefficiency levels, and have the most extreme outlier values for seven consecutive years. Every outlier above 200% but one, in 2018, is an observation from a public DSO.

Figure 3.4c presents the geographic results for all DSOs. The northern region comprises mostly DSOs in the third and fourth quartile for both periods. The northeastern and central-western regions presented decreasing performance from 2011 to 2018.

In 2018, the southern and southeastern regions show slightly better performance as compared to 2011. The bad results for the northern region are shown in Figure 3.4d. The northern region has the biggest box plot, implying higher data dispersion; it also presents a significant number of outliers (13 observations above 700%). All regions show better results for private DSOs. The southern presents the least difference between private and public DSOs. However, it is noticeable that private DSOs have lower dispersion and better overall performances throughout all years and regions.



(a) Mean results

(b) Yearly results

(c) Geographical results

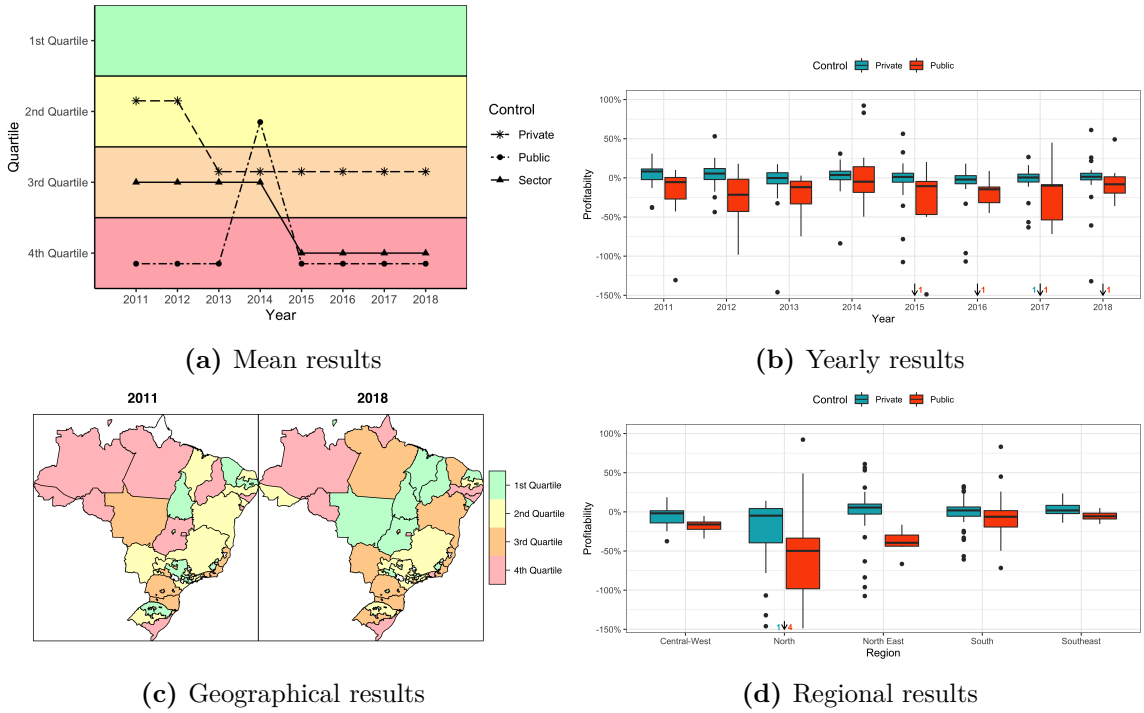
(d) Regional results

Figure 3.4 – Results for indicator 3

3.4.4 Indicator 4 - Profitability

Indicator 4 evaluates the profitability (the higher, the better) of the DSOs. Results are presented in Figure 3.5. Figure 3.5a shows the mean results for private and public DSOs, and the sector. Private DSOs present better results than public ones, every year but 2014. Results for private DSOs are mostly in the third quartile, while results for public DSOs are mostly in the fourth quartile. The sector shows a negative shift during the period. It went from the third quartile during the first four years, to the fourth quartile during the last four years. Figure 3.5b presents box plots of the results for private and public DSOs throughout the years. The special case found in 2014 is presented again. The year 2014 is the only year that the upper limit of the boxplot (third quartile) for the publicly controlled DSOs is higher than that of the boxplot for the privately controlled DSOs. Also, the two farthest outliers are observed for public DSOs in 2014, facts that support and clarify the uncommon results obtained that year. However, it is clear that private results are better in general, with positive values for each year's median, lower dispersion and less negative outliers. The medians for public DSOs are all negative, even in 2014, i.e., the DSO's EBIT is lower than the regulatory EBIT defined by ANEEL. Figure 3.5c shows the geographical results for all DSOs in 2011 and 2018. The northern region stands out negatively

in 2011, with almost all DSOs in the fourth quartile. It gets better in 2018, with DSOs in the first, second and third quartiles. The central-western region also shows a performance boost between those years. According to Figure 3.5d, the regions with greater discrepancy between public and private DSOs are the northern and northeastern. Private DSOs from the northeastern region present the best overall results, and southern and southeastern regions show the smallest differences between public and private DSOs.



(a) Mean results

(b) Yearly results

(c) Geographical results

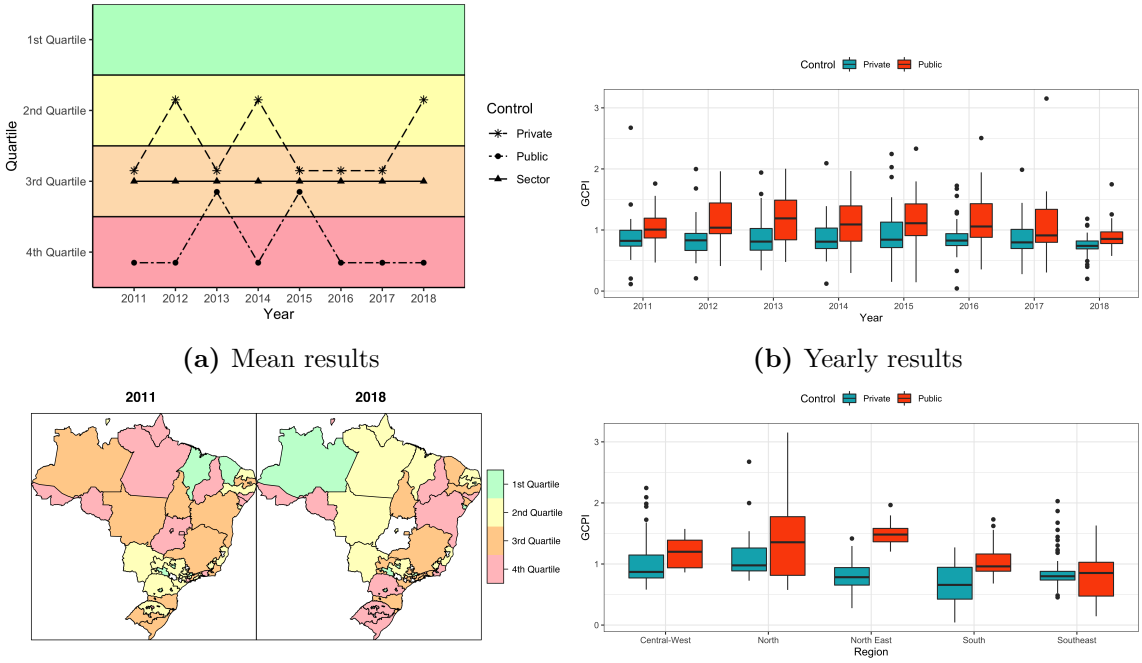
(d) Regional results

Figure 3.5 – Results for indicator 4

3.4.5 Indicator 5 - GCPI

Indicator 5 evaluates GCPI of the DSOs (the lower, the better). Results are presented in Figure 3.6. Figure 3.6a shows the mean results for private and public DSOs, and the sector. The sector presents a constant result in the third quartile. Results for private DSOs fluctuate between the second and third quartiles. Results for public DSOs are worse, in the third and fourth quartiles. Figure 3.6b shows that most private DSOs have a GCPI value below 1 every year, while public DSOs consistently presented a GCPI below 1 (third quartile is below 1) only in 2018.

Figure 3.6c presents the results for each DSO. The northern region has the worst results in 2011, and shows improvement in 2018. Meanwhile, the southern region showed significant worsening for this indicator during these years, obtaining results mostly in the third and fourth quartiles in 2018. Results for each region are presented in Figure 3.6d. Results for private DSOs are better for all regions except, possibly, the southeastern region, in which public DSOs have a lower second quartile, but bigger dispersion and higher second and third quartiles. The northeastern region shows the greatest discrepancy between public and private DSOs. The southern region has the overall best results.



(a) Mean results

(b) Yearly results

(c) Geographical results

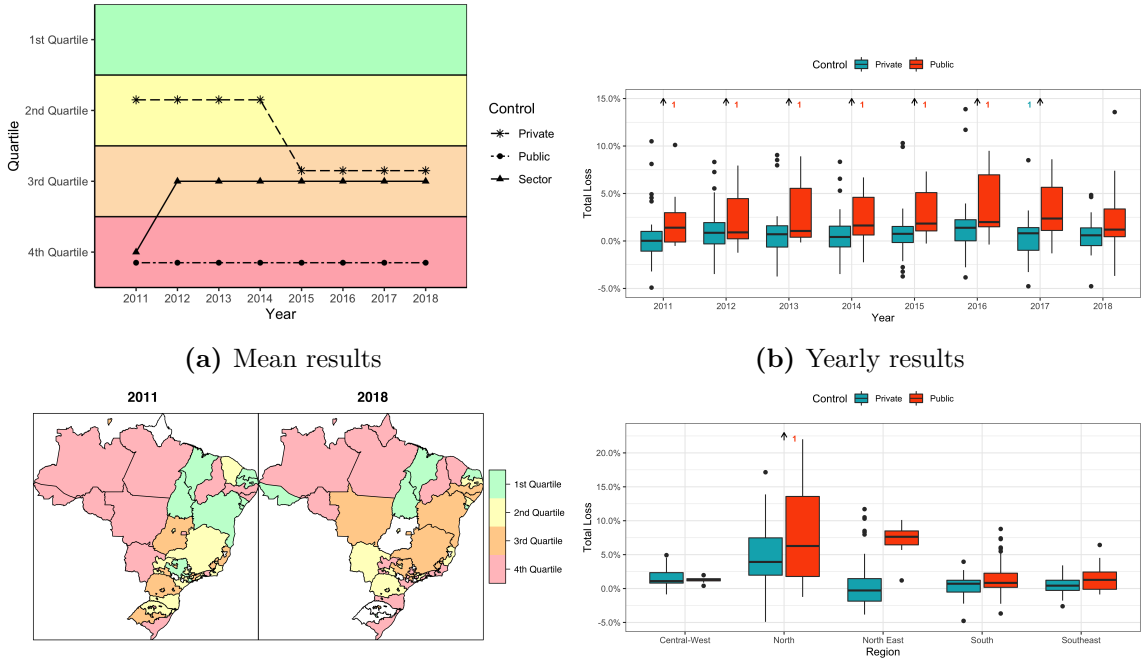
(d) Regional results

Figure 3.6 – Results for indicator 5

3.4.6 Indicator 6 - Total Loss

Indicator 6 evaluates the total loss of the DSOs (the lower, the better). Results are presented in Figure 3.7. Figure 3.7a shows the mean results for private and public DSOs, and the sector. Public DSOs present constant results in the fourth quartile, for all years. Private DSOs show good results (second quartile) from 2011 to 2014, but fall to the third quartile from 2015 to 2018. The sector shows results consistently in the third quartile, except in 2011 when it had a fourth quartile result. Figure 3.7b shows the results for private and public DSOs from 2011 to 2018. Public DSOs present higher losses than private DSOs every year, and the dispersion differences based on control type are particularly noticeable for this indicator. On the one hand, private DSOs have greater dispersion on the lower part of the boxplots, and small dispersion between the first and third quartiles (the colored part of the boxplots). On the other hand, public DSOs have greater dispersion on the higher part of the boxplots, with discrepant outliers and a noticeable dispersion between the first and third quartiles. The best result for the public DSOs is in 2011, followed closely by the results of 2018. Losses increased from 2012 to 2017. Results for the private DSOs' did not feature any major differences during the analyzed period.

Figure 3.7c shows the geographical results for all DSOs in 2011 and 2018. In 2011, the northern and central-western regions present almost exclusively fourth quartile results. The northeastern region stands out in 2011 with several first quartile results. In 2018, the central-western and northern regions show slight improvements, while the northeastern region presents worse results. Figure 3.7d presents the results for each region. Results for public DSOs from the central-western, southern and southeastern regions are better than results for private DSOs from the northern region, and similar to private results in the northeastern region. The northern (private and public) and northeastern (public) regions stand out negatively with the worst results.



(a) Mean results

(b) Yearly results

(c) Geographical results

(d) Regional results

Figure 3.7 – Results for indicator 6

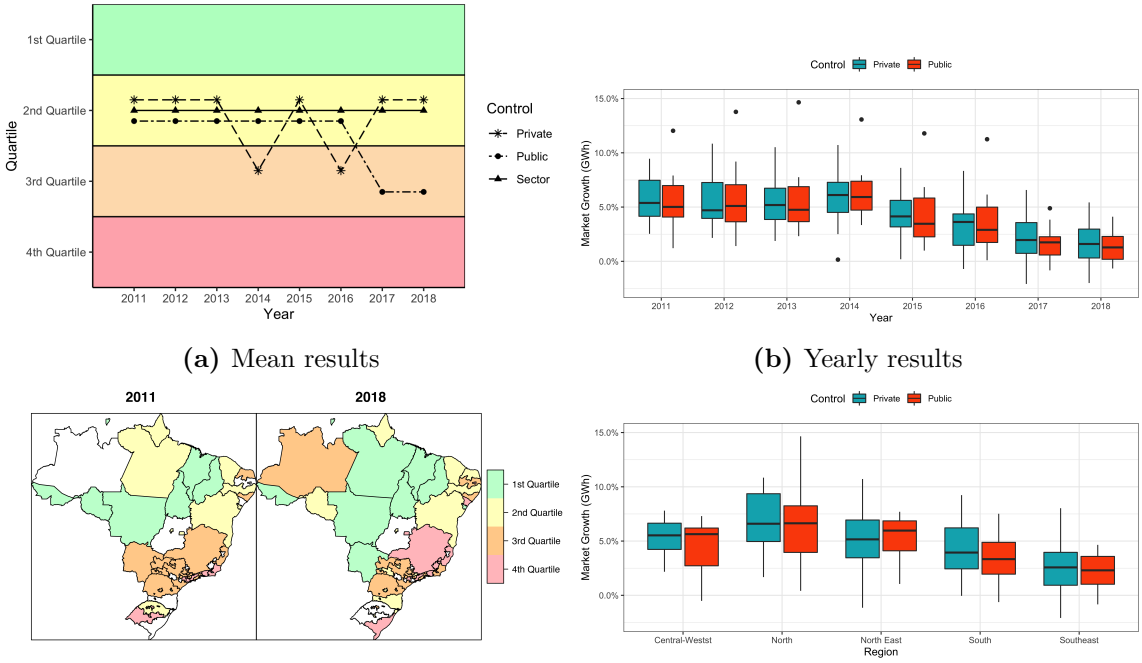
3.4.7 Indicator 7 - Market Growth (GWh)

Indicator 7 evaluates the market growth, in GWh (the higher, the better), of the DSOs. Results are presented in Figure 3.8. Figure 3.8a shows the mean results for private and public DSOs, and the sector. Results for the sector are consistently in the second quartile for all years. Both private and public DSO groups reached a second quartile result for almost all years. The exceptions are 2014 and 2016 for public DSOs, and 2017 and 2018 for private DSOs. Figure 3.8b presents results for all private and public DSOs from 2011 to 2018. Results for public and private DSOs are similar every year; even the downtrend that started in 2015 affects both similarly. The market growth median per year in the first four years fluctuates around 5%, and this number starts gradually to reduce between 2014 and 2018, reaching a median of 2% in 2018.

Figure 3.8c presents the results for each DSO in 2011 and 2018. The northern region stands out positively in both periods, with most results in the first and second quartiles. The southern and southeastern regions present bad results in 2011, and even worse results in 2018. Figure 3.8d presents the results for all DSOs, divided by regions. The northern region shows the best results for both private and public DSOs. The public result of the northern region is even better than all private results from other regions. The southeastern region is the worst for both private and public DSOs.

3.4.8 Indicator 8 - Market Growth (Consumers)

Indicator 8 evaluates the market growth of the DSOs in number of consumers (the higher, the better). Results are presented in Figure 3.9. Figure 3.9a shows the mean results for private and public DSOs, and the sector. Private DSOs show a consistent result in the second quartile from 2011 to 2018. The sector also shows a consistent result in the second quartile, except for



(a) Mean results

(b) Yearly results

(c) Geographical results

(d) Regional results

Figure 3.8 – Results for indicator 7

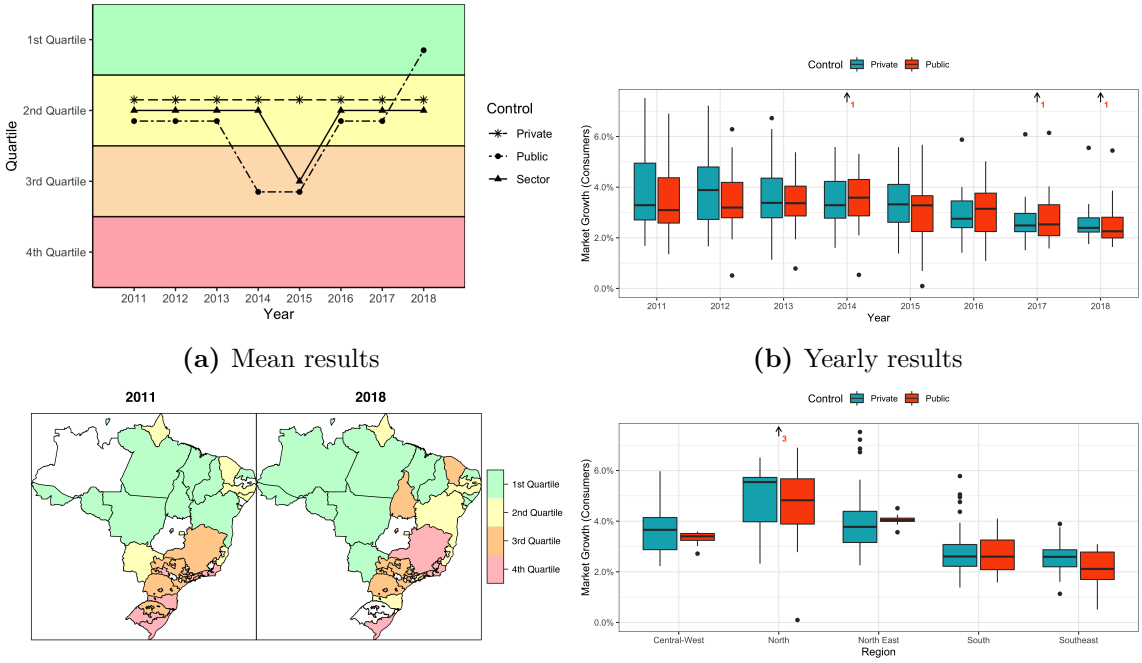
2015. Public DSOs fluctuate more, among the first, second and third quartiles. It is noteworthy that this is the first mean result in the first quartile for all indicators. Figure 3.9b presents the results for all DSOs from 2011 to 2018. Private DSOs have higher market growth than public DSOs in the first five years. Both private and public DSOs present a decrease in market growth in the last three years. The decrease is sharper for the private DSOs, resulting in better results for public DSOs from 2016 to 2018.

Figure 3.9c shows the results for all DSOs in 2011 and 2018. The northern region stands out positively in both years with almost all results in the first quartile. Meanwhile, the southern and southeastern regions present almost exclusively third and fourth quartile results for both years. This tendency is demonstrated in Figure 3.9d. It shows that the southern and southeastern results, for both private and public DSOs, are worse than in all the other regions. The northern region shows the best results, followed by the northeastern and central-western regions, respectively.

3.4.9 Ordinal Logistic Regression Models

In this section we present the ordinal logistic regression results for each indicator. The dependent variable (quartile result) is ordered so that 4th quartile (worst) $<$ 3rd quartile $<$ 2nd quartile $<$ 1st quartile (best). We chose the control type and region as the independent variables of the model. Ordinal logistic models usually present results in terms of odds, or chances. Odds, or chances, are the increased or decreased probabilities of a quartile change when altering one or more independent variables. Besides presenting the odds, or chances, we also chose to include the predictive probability for each quartile, using interpretable and concise charts. Probabilities for each indicator and each quartile vary depending on the region and the control type.

The exponential estimates (and the respective $p-value$) for the proposed ordinal logistic



(a) Mean results

(b) Yearly results

(c) Geographical results

(d) Regional results

Figure 3.9 – Results for indicator 8

regression models are presented in Table 3.2. The reference for the region variable is the northern region, and the reference for the control type variable is public control. Significant estimates ($p - value \leq 0.05$) are highlighted with an asterisk and bold font.

Table 3.2 – Exponential estimates for the ordinal logistic regression model

Variable	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6	Ind. 7	Ind. 8
Central – West Region	**6.651***	**3.459***	0.671	**0.579***	1.222	**1.698***	**0.373***	**0.177***
Northeast Region	**8.744***	**13.111***	**3.965***	**2.178***	**2.728***	**5.525***	**0.358***	**0.333***
Southeast Region	**18.279***	**17.027***	**3.263***	**2.013***	**3.745***	**3.931***	**0.016***	**0.011***
South Region	**15.489***	**6.492***	**3.982***	**1.822***	**3.919***	**3.541***	**0.091***	**0.018***
Private Control	**7.556***	**6.076***	**6.974***	**4.607***	**3.977***	**1.798***	**1.623***	1.342
Year	**0.970***	1.000	**0.996***	1.000	**1.036***	**0.979***	**0.936***	**0.937***

The estimated coefficients present the impact of the region, control type and year for each indicator. At least 5 out of 6 coefficients are significant for each indicator, demonstrating the consistency of the model. The reference is the northern region, and the exponential estimated coefficients of the other regions for indicators 1 to 6 are greater than 1. Thus, we verify that simply changing the DSO region from the northern to any other region has a positive impact on achieving a better quartile result. This extent of this impact changes according to the indicator. The southeastern and southern regions are more favored for indicator 1, while the northeastern

and southeastern regions are more favored for indicator 2. The behavior shifts when analyzing indicators 7 and 8. The exponential estimated coefficients for the other regions is less than 1, indicating that the northern region performs better for these indicators.

The same can be said about control type. The reference is public control, and the exponential estimated coefficients of private control for all indicators is greater than 1. This implies that changing the control type from public to private is enough to increase the chances of achieving better results for any indicator. This is more relevant for indicators 1, 2 and 3, and less relevant for indicators 6, 7 and 8. The significant estimates for the variable year show a decreasing tendency for all indicators except indicator 5, which implies that the GCPI is increasing each year.

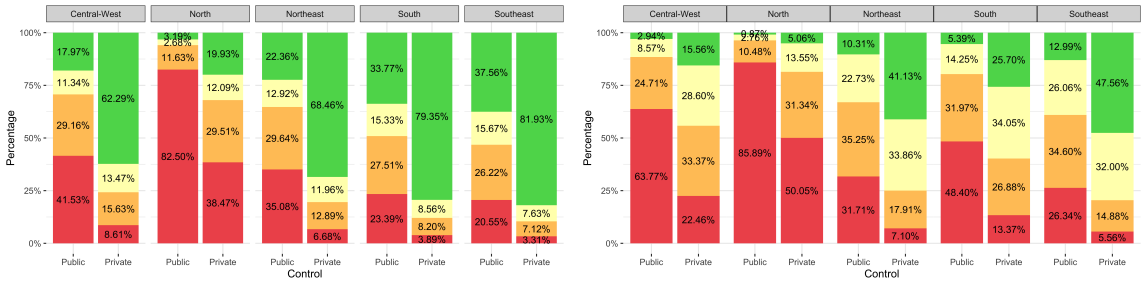
Predictive results for all indicators are presented using bar plots in Figure 3.10 and Figure 3.11. Figure 3.10a presents the results for indicator 1 (indebtedness). Results from the northern region clearly show the struggle of its DSOs to achieve a first quartile result. A private DSO in the northern region has a 19.94% chance of being in the first quartile, while a public DSO has only a 3.19% chance of being in the first quartile. In comparison, a public DSO in the southeastern region has 37.52% of being in the first quartile, and a private DSO in the southeastern region has a surprising 81.92% chance of being in the first quartile. Public DSOs in the central-western, northern and northeastern regions are more likely to be in the last two quartiles. Public DSOs in the southeastern region stand out positively. They are the only public DSOs with higher probabilities of being in the first two quartiles than in the last two quartiles.

Figure 3.10b presents the probabilities for indicator 2 (efficiency). Almost every DSO struggles to achieve a good quartile result. The exceptions are private DSOs in the northeastern, southern and southeastern regions. Public DSOs from all regions have less than a 40% chance of achieving the first two quartiles, with those in the northern region having less than a 5% chance. Unfortunately, a public DSO located in the northern region has an 86% chance of being among the worst 25% DSOs.

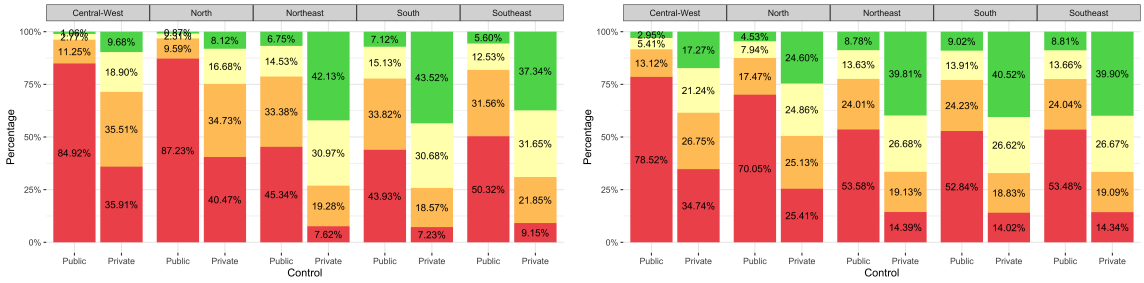
Figure 3.10c presents the results for indicator 3 (inefficiency). Results for the central-western and northern regions are similar for both public and private DSOs. Results obtained for the northeastern, southern and southeastern regions are also similar for both public and private DSOs. Private results for the northern and central-western regions are similar to the public results for the other three regions. Private results for the northeastern, southern and southeastern regions outperform all results from the other two regions. Northeastern private DSOs have the best performance regarding indicator 3.

Figure 3.10d presents the results for indicator 4 (profitability). Private DSOs located on the northeastern, southern and southeastern regions achieved the best performances. Private DSOs in the northern region have equal probabilities of being in any of the four quartiles. Private DSOs in the central-western region have higher probabilities of being in the last two quartiles. Lastly, public DSOs in all regions have at least a 75% chance of being in the last two quartiles.

Figure 3.11a presents the results for indicator 5 (GCPI). Private DSOs from the central-western and northern regions performance similarly to public DSOs from the northeastern,



(a) Indicator 1 - Probabilities for each quartile **(b)** Indicator 2 - Probabilities for each quartile



(c) Indicator 3 - Probabilities for each quartile **(d)** Indicator 4 - Probabilities for each quartile

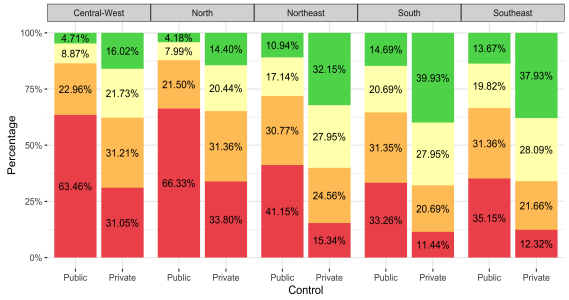
Figure 3.10 – Results for indicators 1 to 4

southern and southeastern regions. Private DSOs from the northeastern, southern and southeastern regions stand out positively. They have less than a 40% chance of being in the last two quartiles, and less than a 16% chance of being in the last quartile. DSOs from the northern and central-western regions have at least a 31% chance of being in the last quartile, regardless of the control type.

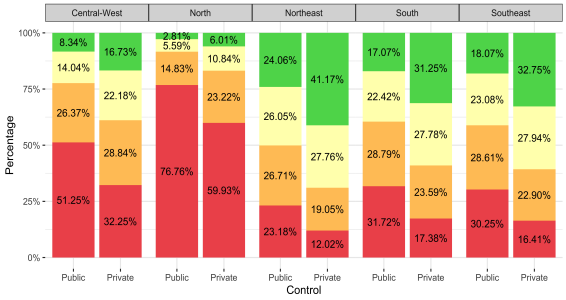
Figure 3.11b presents the results for indicator 6 (total loss). The northern region stands out negatively, with at least a 60% chance of being in the last quartile, regardless of the control type. The northeastern region stands out positively, with at least a 50% probability of being in the first two quartiles, regardless of the control type, which is the best overall result. The southern and southeastern regions present better results in comparison to the central-western and northern regions.

Figure 3.11c presents the results for indicator 7 (market growth (GWh)). The northern, central-western and northeastern regions, in this order, show the best results regardless of the control type. The southern region presents worse results than these three regions. The negative highlight is the southeastern region, with at least an 88% of being in the last two quartiles. The northern region is the positive highlight, with at least a 92% of being in the first two quartiles, regardless of control type.

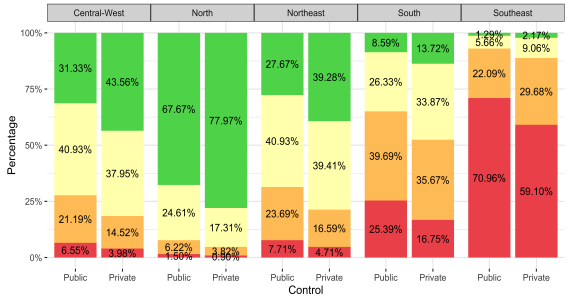
Figure 3.11d presents the results for indicator 8 (market growth (consumers)). The northern region is the best region. This region has at least a 97% chance of being in the first two quartiles, and at least an 81% chance of being in the first quartile. The central-western and northeastern regions present similar, good results. The southern and southeastern regions stand out negatively, with at least a 73% chance of being in the last two quartiles, regardless of control type.



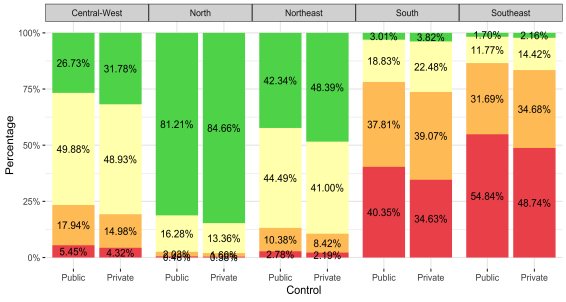
(a) Indicator 5 - Probabilities for each quartile



(b) Indicator 6 - Probabilities for each quartile



(c) Indicator 7 - Probabilities for each quartile



(d) Indicator 8 - Probabilities for each quartile

Figure 3.11 – Results for indicators 5 to 8

3.5 Discussion

3.5.1 Sector analysis

Sector results are calculated using the means of all public and private DSOs. Results show that the sector is not performing well according to the set of indicators defined by ANEEL. The sector presents results in the third and fourth quartiles for indicators 1 to 6. Results for private DSOs are often in the second and third quartiles. Public DSOs present results in the third and fourth quartiles, thus contributing largely to the bad performance of the sector. Public DSOs attained better results for the last two indicators, improving the sector's overall performance. This explains the sector improvement for the last two indicators, with results mostly in the second quartile. Private DSOs perform better for indicators 1 to 6, but do not attain the best possible result (first quartile). Figure 3.12 presents a radar plot, showing the median of each indicator for both public and private DSOs. The median is defined considering data for all indicators, for all 8 years. The median results are similar to those calculated using the means. They emphasize the relationship between private and public DSOs, in which the former performs better for indicators 1 to 6, and the latter performs better for indicators 7 and 8.

The indebtedness level (indicator 1), among the indicators investigated in this work, is arguably the most critical indicator evaluated by ANEEL, thus having specific quartile thresholds. The observed indebtedness level of the sector is worrisome, as it does not even each a second quartile result in any year. Performances of public DSOs are critical for this result. Indicators 2 to 6 also show the same pattern: private DSOs in the second and third quartiles, public DSOs in the third and fourth quartiles. Indicators 5 (GCPI) and 6 (total loss) are the best performing indicators of the sector, showing a consistent result in the third quartile for the last seven years.

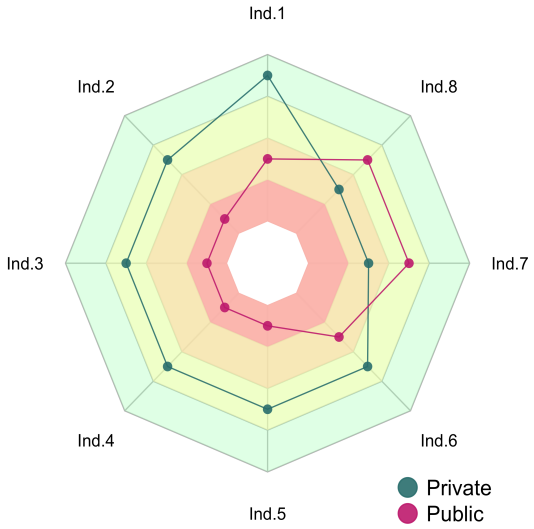


Figure 3.12 – Radar plot of the median for all indicators for public and private DSOs

In contrast, indicators 2 (efficiency), 3 (inefficiency) and 4 (profitability) present the worst results for the last four years, with at least two consecutive years in the fourth quartile.

The pattern of the previous indicators changes for indicators 7 and 8. For these indicators, public DSOs perform similarly to private DSOs. Therefore, they do not impose an outstanding negative impact on the mean result of the sector. For both indicators, this culminates in a second quartile result for public and private DSOs, almost every year. However, the bad results shown for indicators 1 to 6 is a concern, if the sector is going to be able seize this market growth opportunity. High indebtedness, inefficiency and loss levels can jeopardize the survival of the DSOs in the future, mainly the publicly controlled DSOs.

3.5.2 Control type analysis

Results shows that control type is a determining variable of the outcome of a DSO. It is clear by now that, in regard to indicators 1 to 6, public DSOs present worse results than private DSOs. It is not even possible to calculate the indebtedness level of most public DSOs, due to the fact that they have attained negative cash flows (EBITDA - RRQ) or negative EBITDA in several observations. These results, as explained, result in them being assigned to the third and fourth quartiles, respectively. As shown in the jitter plot in Figure 3.2b, 2011 is the year with the best results as only 6 (40%) of public DSOs were assigned to the two last quartiles. The scenario gets worse in the following years, and the worst performance occurs in 2016, with 13 (86.67%) public DSOs assigned to the last two quartiles. The number of private DSOs is almost double that of public DSOs (36×15). Even so, the number of private DSOs with negative cash flow (EBITDA - RRQ) or negative EBITDA is less in both absolute value and in percentage. The worst year for private DSOs is 2013, with 8 (22.22%) operators assigned to the third and fourth quartiles. The best years for private DSOs are 2014 and 2018, both with only 3 (8.33%) operators assigned to the third and fourth quartiles.

The box plots for indicators 2 to 6 show a similar pattern of private DSOs performing better than public DSOs. It is important to highlight that, while private DSOs present the best

results, these results are not necessarily good. The methodology chosen by ANEEL, for indicators 2 to 8, is to compare the result of a DSO to the results of all other DSOs in a certain year and then rank them ordinally based on this comparison. The threshold limits for each quartile are volatile, and change according to the data every year. This can lead to unusual results, as they are basically comparisons of a DSO and its peers, without a component reflecting real, cardinal (absolute) improvements. For example, in one year a DSO can achieve an efficiency level of 5% and be assigned to the first quartile, and in the following year achieve an efficiency level of 5.5% and be assigned to the second quartile. While it is important to assess quartile results for the indicators, it should also be important to determine specific threshold limits for the indicators, e.g., a 7% efficiency level. ANEEL should at least evaluate improvements obtained by each DSO without comparing them to their peers, i.e., observe if a DSO improves its own performance, individually. That said, the greatest contrast between public and private DSOs implies that the former control type has management issues, suggesting a lack of crucial attributes that would enable success in maintaining sustainability. Most public companies perform badly. Likely causes of this problem may be related to problems in how Brazilian public organizations function, and the political influences on them. A future study should consider and analyze these factors as possible causes of this generally poor performance. It is not the objective of the present work to explain the differences found, only to demonstrate them and encourage a deeper discussion of this topic.

Indicators 7 and 8 clearly diverge from the previous ones, with public DSOs performing equal to or better than private DSOs. Indicator 8 had the only first quartile result of all the indicators, and that was obtained by public DSOs (2018). Private DSO results are found mostly in the second quartile. As the sector result is just the mean of all DSOs, it had good results as well. Both indicators 7 and 8 measure market growth, using different parameters. Market growth is a relevant variable, but it is important to highlight that DSOs have little control over it since it is mainly dependent on the social, economic and demographic characteristics of the concession area. It is different from indebtedness or efficiency, which depend almost completely on the actions of the DSOs, themselves, in order to improve.

3.5.3 Region analysis

Results divided by regions show that the northern region struggles to achieve good performance. For indicators 1 to 6, the northern region is largely dominated by third and fourth quartile results, and it has not exhibited much improvement in the last 8 years. It is noteworthy that Brazil is a country of continental size, and each region has a distinct geography, climate and vegetation. These variables have direct impact on the operation of a DSO. For example, a large part of the northern region is covered by the Amazon rainforest. This implies dense vegetation, heavy rainfall and logistical difficulties. It can be considered unrealistic to compare the performance of two DSOs, if one were in the northern region and other were in the southern region. Several studies have already come to this realization, when defining the efficient operational costs of DSOs (Andrade et al., 2014; Silva et al., 2019c; Ganhadeiro et al., 2018; Silva et al., 2019b; Gil et al., 2017; Andrade et al., 2008; Silva et al., 2019a). Instead of using a simple Data Envelopment Analysis (DEA), these studies have proposed that ANEEL use a second stage DEA

model that includes environmental variables to assess the disparities among regions. It is not the objective of the present work to propose a fairer comparative methodology, only to expose the differences among regions and encourage a deeper discussion of the matter.

The disparity between public and private DSOs is also present in indicators 1 to 6, but it varies according to the region. The northern and northeastern regions present considerable differences between the results of public and private DSOs. The central-western, southern and southeastern regions present small differences between public and private DSOs, with the latter presenting the best results. Results for indicators 7 and 8 present less disparity between public and private DSOs. The northern region stands out positively with the highest market growth in recent years, for both public and private DSOs, in terms of GWh and number of consumers.

3.5.4 Ordinal logistic regression model analysis

The ordinal logistic regression models for the eight indicators verify some assumptions created by the authors throughout this study. Public control has negative influence on the probability of achieving a good quartile for indicators 1 to 6. The effect is the opposite (positive) for indicators 7 and 8.

The region where a DSO is located also has an important impact on its result. For indicators 1 to 6, any DSO in the northern region, independent of the control type, has at least a 50.54% chance of being assigned to the last two quartiles. In comparison, in the southern and southeastern regions, this probability falls to 12.08% and 10.44%, respectively. The probability of a DSO in the northern region achieving a result in the first two quartiles varies from 3.18% (indicator 3) to 49.46% (indicator 4). For the southern region, these numbers rise to 22.25% (indicator 3) and 87.92% (indicator 1), respectively.

Again, results for indicators 7 and 8 differ from the rest. The impact of public control on these indicators is positive, regardless of the region. The effect of the region aspect is also switched. Being a DSO in the northern, central-western and northeastern regions almost assures a good result (first two quartiles). However, being a DSO in the southern and southeastern regions almost assures a result in the last two quartiles. This implies that, while the southern and southeastern regions perform well in the other 6 indicators, they do not show potential for market growth in the future, which can compromise expanding on these regions.

3.6 Conclusions

The data analysis and statistical models carried out in the present study demonstrate the importance of control type and region in the result of a DSO. These conclusions are reflected in the different graphical representations used throughout the present article. For example, the northern region presents worrisome results, while the southern and southeastern regions consistently present good results, regardless of control type.

Results suggest that regional differences play an important role in the performance of a DSO. Good and bad performance patterns are seen consistently in the same regions, for multiple

indicators. These characteristics are not yet assimilated into the current methodology used by ANEEL.

Results also demonstrate a significant discrepancy between the results from public and private DSOs. We cannot confirm that public DSOs have management problems. However, the clear difference in the results between public and private DSOs suggests the existence of issues in how public companies function. The exchange of information and experiences between public and private DSOs should be stimulated by ANEEL to level and boost sector results.

The ordinal logistic regression model is an important tool for confirming assumptions and exploratory results. Without the model, the study would lean only on data visualization. The statistical model allowed not only confirmation of the results found, but also measurement of them. The model proved successful in quantifying the struggle of public DSOs, and of the northern and central-western regions, to achieve good results.

The methodology proposed by ANEEL is considered adequate to measure the financial-economic sustainability of DSOs in Brazil. However, it can be considered superficial: it does not evaluate the individual improvements of a DSO over the years, neither does it evaluate regional differences nor control type. We propose that ANEEL adapt its methodology to assess these questions. The kind of analysis and models implemented in the present study can be adapted to evaluate the financial-economic sustainability of new indicators related to sustainability performance. It is also important that ANEEL carefully evaluate the situation of public DSOs in the coming years. The difference between public and private DSOs is significant, and it does not seem to be narrowing in the near future. The northern region, particularly, has the largest growth potential (among all DSOs). However, according to the results for indicators 1 to 6, it lacks the organizational structure to seize this opportunity.

3.7 Abbreviations

ANEEL: National Electric Energy Agency; CAGR: Compound Average Growth Rate; DEA: Data Envelopment Analysis; DSO: Distribution Systems Operators; EBIT: Earnings Before Interest and Taxes; EBITDA: Earnings Before Interest, Taxes, Depreciation and Amortization; EDI: Energy Development Index; ESI: Energy Sustainability Index; GCPI: Global Continuity Performance Index; GDP: Gross Domestic Product; GWh: Gigawatt-hour; HDI: Human Development Index; MEPI: Multidimensional Energy Poverty Index; NESSI: National Electricity System Sustainability Index; NDEA: Network Data Envelopment Analysis; NPB: Net Pay Basis; OPEX: Operational Expenditures; PCA: Principal Component Analysis; PBV: Parcel B Value; Reg: Regulatory; RND: Regulatory Net Debt; RRQ: Regulatory Reintegration Quota; SDG: Sustainable Development Goals; SIGEL: Electric Sector Geographic Information System.

4 Multilayer Hybrid Models applied to the power outage index

Abstract

The National Electric Energy Agency (ANEEL) is responsible for the regulation of the electric sector in Brazil. The Index of Equivalent Duration of Interruption per Consumer Unit (DEC) is used by ANEEL as the main indicator of the quality level in the operations of distribution companies. Using data from the operations of Companhia Energética de Minas Gerais (CEMIG) from 2019, it was carried out a study on the DEC index using Structural Equation Models (SEM) and Hybrid multilayer models. The objective in this paper was twofold: 1) identify the main variables that impact the DEC; 2) find the best possible predictive model for DEC. The results found allowed the identification of the most relevant variables when analyzing the DEC. Also, a model with considerable predictive power ($R^2_{pred} = 70\%$) was achieved using accounting, operational, climatic and geographic variables as input.

Keywords: Power Outage Index; Hybrid Model; Structural Equation Model; CART; Random Forests.

4.1 Introduction

In Brazil, as in most countries, the energy distribution sector can be defined as a natural monopoly (Sioshansi; Pfaffenberger, 2006). Therefore, there should be a regulatory agent to guarantee a healthy economic relationship between the distribution system operators (DSO) and the customers. In Brazil, this agent is Agência Nacional de Energia Elétrica (ANEEL).

The absence of such an agent in a natural monopoly can result in an unbalanced relationship between supply and demand (Posner, 1999). Thus, direct market controls are necessary to ensure a satisfactory performance by the companies. Some examples include control of profit, specific interest rate, permission to enter the market and minimum quality requirements.

In regard to quality, ANEEL defines it mainly as interruptions longer than three minutes in the electric energy supply (ANEEL, 2016c). To analyze the quality of the DSO's, ANEEL has developed an index called *Duração Equivalente de Interrupção por Unidade Consumidora* (DEC). This index defines the average time a customer remained without energy.

The DEC index has a maximum threshold defined by ANEEL for each DSO. Moreover, each DSO must reach a different maximum threshold for each of its electrical groups. Electrical group is a geographical division of the concession area of the DSO, defined by ANEEL. The quantities and areas, as well as the number of customers, of the electrical groups are varied.

By defining different thresholds for different electrical groups, ANEEL takes into consideration the intrinsic differences between the regions. For example, one electrical group may contain

more rural customers, while another may contain more industrial customers. It is also possible to find distinct climate conditions in two groups from the same DSO, as shown in Queiroz, Costa e Lopes (2017).

Therefore, the DSOs must compute their DEC indices and submit them to ANEEL, which then evaluates the quality of the DSOs. The DEC index can be defined as shown in Equation 4.1 (ANEEL, 2016c):

$$DEC = \frac{\sum_{j=1}^{Cc} DIC(j)}{Cc} \tag{4.1}$$

where DEC is the average interruption duration per customer, $DIC(j)$ is the individual interruption duration for each consumer j, Cc is the total consumer units and j is the consumer units index.

Even though the DEC index can be easily computed, as shown, there is not a simple way to keep it below the regulatory threshold. The DEC index represents the summary of the quality of the DSO operation, which is dependent on managerial decisions and countless variables. Hence, the main challenge is to define which variables are the most important and impactful in such an analysis.

Some variables have a direct relation with the DEC index, e.g., the frequency of interruptions in the electrical grid. However, there are many other variables that, at first, do not appear to have direct correlation with the DEC. Some examples are climate variables (e.g., temperature and pluviometric index) and the DSO investment profile.

The DEC evaluation is important for the DSOs, as results above the maximum threshold for two consecutive years, or in 2020, can result in the opening of a concession expiry process by ANEEL (ANEEL, 2016e). Ergo, the main goal of this paper is to use Structural Equation Models (SEM) along with hybrid models in order to clarify the behavior of the DEC index and its relationship with operational, climate and accounting variables.

4.2 Material and Methods

4.2.1 Structural Equation Models (SEM)

Pugesek, Tomer e Eye (2003) define Structural Equation Models (SEM) as a group of techniques that combine factorial analysis and multiple regression analysis. SEM is widely used in several areas such as biology, economy, marketing and health. SEM tries to explain the relations among multiple dependent and independent variables, continuous, discrete or latent, like a series of multiple regression equations, where the variables are non-observable or latent factors (Moeinaddini et al., 2020).

For Schumacker e Lomax (1996), SEM can use several types of models in order to represent different relations among the observed variables. The main objective is to provide a statistical quantitative measurement of a theoretical model hypothesized by a researcher. Using SEM, the researcher can test several hypotheses of how variables can define a construct and

how these constructs relate to each other. SEM can also test various types of theoretical models: linear regression, path and confirmatory factor models.

According to Grace (2006), SEM can be defined as simply as possible using two or more structural equations to model multivariate relations. Multivariate relations are those that involve influences (independent variables) and results (dependent variables) simultaneously. A generic representation of a SEM is shown in Figure 4.1 and Equations 4.2, 4.3 and 4.4.

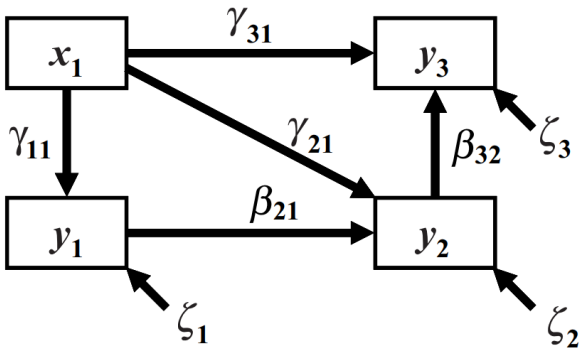


Figure 4.1 – Generic representation of Structural Equation Model.

$$y_1 = \alpha_1 + \gamma_{11} x_1 + \zeta_1 \tag{4.2}$$

$$y_2 = \alpha_2 + \beta_{21} y_1 + \gamma_{21} x_1 + \zeta_2 \tag{4.3}$$

$$y_3 = \alpha_3 + \beta_{32} y_2 + \gamma_{31} x_1 + \zeta_3 \tag{4.4}$$

Figure 4.1 shows one independent variable (x_1) used to model three dependent variables (y_1, y_2 and y_3). The dependent variables can be either quantified, such as temperature, or non-quantified (non-observable), such as quality. This flexibility, to include variables that cannot be quantified, is one of the most important advantages of the SEM. Furthermore, the dependent variables can also be used as independent variables in other regressions. For example, y_3 is dependent on both x_1 and y_2, while y_2 is dependent on both x_1 and y_1. Moreover, each estimated variable (y) is also associated with its own error measure (ζ).

Generally, SEM translate several cause and effect relations that the researcher suspects are true. Such relations are described by parameters that indicate the magnitude of the influence (direct or indirect) of the independent variables (observed or latent) on the dependent variables (observed or latent).

SEM can be divided into confirmatory or exploratory: the former tries to prove a set of proposed relations; the latter tries to develop a theory by repeated applications in the same database.

The variables used in the SEM are also divided into two groups: the observable variables and the latent variables. Observable variables can be quantified by research or data collection. Latent variables, or constructs, are hypothetical or theoretical variables that cannot be directly observed, making it difficult to quantify their existence or influence (Pugesek; Tomer; Eye, 2003).

The variables can also be defined as exogenous, endogenous or mediator. Exogenous variables are those that are only independent, regardless of the regression. Endogenous variables are those that are only dependent, regardless of the regression. Mediator variables are those that can be both independent and dependent, according to the regression. Figure 4.2 shows the most common representation of a SEM using diagrams, symbols and arrows.

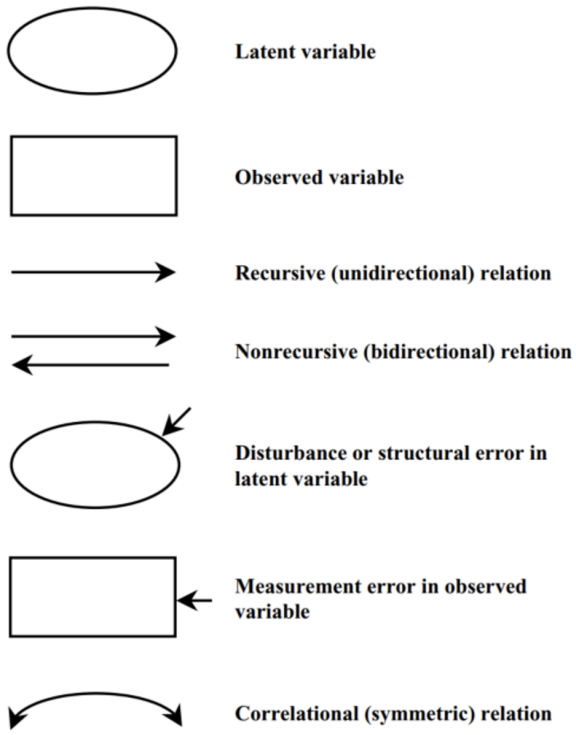


Figure 4.2 – Standard symbology for representing Structural Equation Models.

4.3 Hybrid models

Costa et al. (2019a) present a unique and innovative approach over white and black box models based on the work of Breiman (2001). For Costa et al. (2019a), the best solution to a problem can be a white box model, a black box model or the combination of both, depending on the problem and the available database.

White box models, associated with data modeling, are parametric statistical models in which the data serves as the base to estimate the parameters of the model and to do statistical inferences. White box models are usually based on patterns and rules and they provide a definition closer to the human language (Loyola-Gonzalez, 2019). White box models have the advantage that the estimated parameters generally have real-world significance, which allows for credible interpretations.

Black box models, associated with algorithms, are used to find the best predictive model. According to Wang e Lin (2021), black box models are in definition hard for humans to comprehend because they have a complex and opaque decision-making process. They use several variables and have a more complex structure with non-linear functions, transformations, and

learning with trials and errors. Therefore, the parameters estimated using a black box model are usually not interpretable. Depending on the objective, this may not be a relevant problem, e.g., when the interpretation of the parameters has low priority. Some examples of black box models are neural networks and fuzzy systems.

According to Costa et al. (2019a) the best way to evaluate and compare white and black box models is through cross validation. To perform cross validation, the database is divided into two subsets: training and testing. The training subset is used to create the model, i.e., to estimate the parameters. The testing subset is used subsequently to evaluate the error prediction of the model adjusted from the training subset.

4.3.1 Lasso and Ridge regularization

Costa et al. (2019a) states that the model with the best statistical response/result is not necessarily the one with the best predictive power. An example is a model constructed with several variables. Also, the inference power of the predictors is sensitive to the multicollinearity, due to the excessive number of variables (Nelder; Baker, 1972).

In such cases, regularization methods can be helpful. Regularization methods can increase the predictive power of a model without removing any variables, even those that are not significant. Regularization incorporates additional information to the problem under consideration in order to smooth noisy data by adding an extra term (a penalty) to the usual least square regression (Saccoccio et al., 2014). A simple regression model defined by the minimization of the sum of squared residuals is shown in Equation 4.5:

$$\hat{\beta} = \arg\min_{\beta}\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\} \tag{4.5}$$

where $\hat{\beta}$ is the least squares estimator, $\beta = (\beta_0,\beta_1,\ldots,\beta_k)$ is the vector of estimated parameters, $X = (x_1, x_2, \ldots, x_k)$ is the set of column vector predictors and $Y = (y_1, y_2, \ldots, y_k)$ is the vector of response variables.

Adding the restriction helps to mitigate the effect of data overfitting. Data overfitting happens when the model is oversized, i.e., when it uses more variables than necessary. In this case, the model presents an excellent adjustment (overfit) for the data it was built upon but does not present satisfactory performance when using new data, thus showing a low predictive power.

One way to mitigate these problems is to use the Least Absolute Shrinkage and Selection Operator (LASSO) method, proposed by Tibshirani (1996), also known as the l1-norm. The LASSO model is a good option to handle overfitting when working with huge covariates (Nandagopal et al., 2019). The LASSO is applied by adding a penalty to Equation 4.5 that restricts the sum of the absolute value of the weights (β) to a pre-determined value, as shown in Equation 4.6:

$$\begin{aligned} &\tilde{\beta} = \arg\min_{\beta}\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\} \\ &subject\ to: \sum_{i=1}^{k} |\beta_i| \leq c \end{aligned} \tag{4.6}$$

where $\tilde{\beta}$ is the least square estimator of the LASSO and c is a pre-determined value. This restriction allows the tuning of the model, which can also incur in some of the coefficient estimators being equal to zero, thus filtering only the most important independent variables (Zhang et al., 2021). Examples of LASSO models can be found in Cao et al. (2020), Zhang et al. (2021) and Nandagopal et al. (2019).

Another alternative is to use the Ridge regression method, also known as the l2-norm, proposed by Hoerl e Kennard (1970). The Ridge estimator adds a penalty which limits the sum of squared values of the estimated parameters (β) to a pre-determined value (Reineking et al., 2006), as shown in Equation 4.7:

$$\begin{aligned} &\tilde{\beta} = \arg\min_{\beta}\{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)\} \\ &subject\ to: \sum_{i=1}^{k} \beta_i^2 \leq c \quad ou \quad \beta^T\beta \leq c \end{aligned} \tag{4.7}$$

where $\tilde{\beta}$ is the least square estimator of the Ridge penalty and c is a pre-determined value (Hoerl; Kennard, 1970). Ridge models can be found in Jiang e Li (2016) and Zhou et al. (2021). The main difference between the two methods is that in LASSO, some estimated parameters can assume null values; while in Ridge, the values of the parameters are reduced but usually do not reach null values.

4.3.2 Classification Trees and Random Forests

Classification Trees, also known as Decision Trees or Classification and Regression Trees (CART), is a way to represent differences among the data using hierarchical relations to divide and explore. This method describes, classifies or generalizes the data (Murthy, 1998). CART models are popular within machine learning works, with various authors using it successfully, such as Jia, Li e Yu (2003), Costa et al. (2017), Zhang et al. (2021) and Li et al. (2014).

Choubin et al. (2018) define CART as a recursive algorithm used to explore the structure of a dataset. The CART creates decision rules that are easy to visualize in order to predict a categorical variable, through the classification tree, and a continuous variable, through the regression tree. The main difference of classification and regression trees is that the latter does not create classes of dependent variables.

The CART model can be represented as a binary tree with branches from which new branches or leaves are created. Each branch of the tree is a ramification created from the division of the data using one of the available predictors. A leaf indicates a final branch from which no new branches are created. Each predictor variable is evaluated individually, generating new branches or leaves, as necessary. The predictor variable and the threshold value to split the data

can be chosen using error minimization, maximum likelihood estimation, or any other criteria. To determine when to stop the growth of the tree, it is recommended to use a rule based on a statistic such as a Qui-Squared test (Fisher, 1922) or a mean t-test (Snedecor; Cochran, 1989).

Breiman et al. (1984) define Random Forests as the combination of CART models. Random Forest models are an efficient technique for supervised classification or regressions (Dong et al., 2021) which are robust against overfitting (Ngouna et al., 2020). They are user-friendly since it only has two necessary parameters: the number of trees in the forest and the number of variables in each node (Keramat-Jahromi et al., 2021). Each tree, or CART model, is adjusted using a random sample of the predictor variables. If we consider a model with m predictor variables, from which k ($k << m$) variables are randomly chosen, these k variables can be used to adjust a CART model. This procedure is repeated n times. At the end of this procedure, there is a total of n trees (or CART models) from which the results are combined to form the Random Forest. The mean or the median of the n models is usually chosen as the main aggregation method. Thus, it is also possible to calculate the mean contribution of each predictor in the maximization of the likelihood of the model. This is known as feature importance, i.e., it identifies the most important predictors. Random Forests models are a very flexible technique. Some examples of use include classification and survival analysis problems (Voronov; Jung; Frisk, 2021), genetic epidemiology and microbiology (Uimonen et al., 2020), protein structure prediction (Kalaiselvi; Thangamani, 2020).

4.3.3 Hybrid Gradient Boosting (HGB)

When creating hybrid models, the goal is to improve the predictive power by combining linear and non-linear models. The idea of Ensemble Models or Boosting (Friedman; Hastie; Tibshirani, 2001) is well presented in the literature. Recently, (Costa et al., 2019a) proposed a generic structure using the Gradient Boosting (Friedman, 2002) concept to improve a base model, like a simple statistical model (white box).

To improve the base model, non-linear models can be applied, such as Random Forest (Zhang; Ma, 2012), XGBoost (Chen et al., 2015), CART (Aguiar et al., 2012) or the combination of models. In the case of regression models, it is possible to combine the residuals derived from the current model into a new response variable for the next model, thus creating new model layers. For example, a Random Forest model can be applied over a multiple linear regression model to identify the most important predictors through the use of feature importance.

HGB is based on exponential family distribution, used in generalized linear models (Nelder; Baker, 1972). A random variable will belong to the exponential family distribution if its density function can be described as shown in Equation 4.8:

$$\log f_Y(y|\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \tag{4.8}$$

where θ is the canonical parameter, ϕ is the dispersion parameter, and $b(\theta)$, $c(y,\phi)$ and $a(\phi)$ are functions related to the density distribution of Y.

It is usually possible to associate the θ parameter with the mean and variance of the response variable. The ϕ parameter is exclusively associated with the variance of the response variable. From Equation 4.8 it is possible to identify that $\theta = \mu$, $\phi = \sigma^2$ and $a(\phi) = \phi$ when Y follows a Gaussian distribution, $Y \sim Normal(\mu, \sigma^2)$. These relations result in Equations 4.9 and 4.10:

$$b(\theta) = \frac{\theta^2}{2} \tag{4.9}$$

$$c(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + ln(2\pi\phi)\right\} \tag{4.10}$$

It is also possible to define the likelihood function from Equation 4.8, as shown in Equation 4.11:

$$\log Lik = \sum_{i=1}^{N} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \tag{4.11}$$

Among the properties of the exponential family, some important ones are the fact that $E(Y) = b(\theta)$ and $Var(Y) = a(\phi)b''(\theta)$. It is easy to incorporate a simple linear regression equation into the exponential family model by making $\theta(X) = X^T\beta$, where $X^T\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is the linear predictor variable. Several distributions can be written with the same form presented in Equation 4.11, like the Normal, Poisson, Binomial, Bernoulli, Multinomial and Gamma (Costa, 2019).

Following Friedman (2001), Costa et al. (2019a) apply an additive expansion to the canonical parameter θ_X presented in Equation 4.12:

$$\theta_X = \sum_{m=0}^{M} \theta_m(X;\beta_m) \tag{4.12}$$

Consider that $\theta_0(X;\beta_0)$ represents any base model, such as a multiple linear regression. Therefore, $\theta_1(X;\beta_1),\ldots,\theta_M(X;\beta_M)$ can represent different models, such as CART or XGBoost. The term β_m represents the vector of parameters associated with each model, like the base model or any later implementation. The HGB algorithm was proposed based on the exponential family representation shown in Equation 4.8 and the additive expansion from Equation 4.12, as shown in Algorithm 1.

$\theta = argmax_\theta \log Lik(\mathbf{y},\theta_{(\mathbf{x})})$;
for $m = 1$ **to** M **do**
  $\tilde{y}_i = -\left[\frac{\partial \log Lik(y_i,\theta_{(x_i)})}{\partial\theta_{(x_i)}}\right]_{\theta_{(\mathbf{x})}=\theta_{m-1}(\mathbf{x})}, i = 1,\ldots,N$;
  $\beta_m = argmin_\beta \sum_{i=1}^{N}[\tilde{y}_i - h_m(\mathbf{x}_i;\beta)]^2$;
  $\theta_{m(\mathbf{x})} = \theta_{m-1(\mathbf{x})} + h_m(\mathbf{x};\beta_m)$;
end

Algoritmo 1: Hybrid gradient boosting.

The base model, preferably a simpler linear model adjusted using maximum likelihood, is presented in Line 1. Thereafter, the HGB algorithm creates new pseudo-answers ($\tilde{y}_i$) that are used as dependent variables in subsequent non-linear models (Line 3).

HGB then estimates the parameters from the M black box models as shown in Line 4. The canonical parameter value of the base model (θ_m) is updated after each iteration m according to the results of the black box models (Line 5).

New pseudo-answers are obtained at each subsequent step of the algorithm by calculating the differences between the observed results and those obtained from previous layers of the algorithm. Each new pseudo-answer tries to aggregate new information that was not captured by the previous layers. The new black box models added to the base model at each step m are adjusted again to the new pseudo-answers by the least square. This adjustment leads to new estimated parameters $\tilde{\beta}_m$ that are used to update the value of the canonical parameter $\theta_{m(\mathbf{x})}$. This procedure occurs iteratively until the addition of new layers no longer present any statistical gain to the model.

The idea behind the above is similar to that of the XGBoost; however, it is more flexible as it permits the use of a generalized statistical structure. The HGB algorithm creates several layers of models for regression problems by using the residuals from previous layers as new pseudo-answers to the subsequent layers.

4.4 Results and Discussion

4.4.1 The proposed model

The Structural Equation Model (SEM) was used on a database from CEMIG-D. The data is from 2018, when the concession area from CEMIG-D was divided among 271 electrical groups. The global DEC index is calculated as the average of the DEC index from each electrical group.

Thus, the data comprises 271 samples for each variable. The variables were chosen after several meetings between the researchers and a technical team from CEMIG-D. In the end, the database was constructed with 23 variables from different areas such as operations, accounting, finance and climate. The correlations among all variables are shown in Figure 4.3. The variables with brief explanations include:

1. DL.km: length of distribution lines (km);
2. Network.km: length of distribution network (km);
3. Total.Customers: total number of customers;
4. SS.number: number of substations;
5. Protective.Equip: amount of protective equipment;
6. Automated.Equip: amount of automated equipment;

7. Area.km2: concession area (km^2);
8. Roads.km: road length in the concession area (km);
9. Cities: number of cities;
10. Places: number of places;
11. Maintenance.number: number of maintenance teams;
12. Commercial.Services: number of commercial services performed;
13. Emergency.Services: number of emergency services performed;
14. DL.FSS: number of interruptions caused by trees on distribution lines;
15. SS.FSS: number of interruptions caused by trees on substations;
16. Network.FSS: number of interruptions caused by trees on the distribution network;
17. OPEX.Resources: investments in OPEX (Operational Expenditures - R$);
18. CAPEX.Resources: investments in CAPEX (Capital Expenditures - R$);
19. Rain.Volume: rain volume (mm);
20. Humidity: humidity level (%);
21. Temperature: average temperature (^oC);
22. DEC.2017: DEC index of 2017;
23. DEC: DEC index of 2018 (minutes).

We first developed a SEM model with a technical team from CEMIG-D. The goal was to predict the "Quality Performance" (the DEC index variation) from the latent variables "Electrical Assets", "Logistic Assets" and "Service Demand". This first configuration is shown in Figure 4.4. It was referred to as the empirical model because it was created solely by the CEMIG-D team.

Several variations and variable combinations were tested next. One example is using factorial analysis and principal component analysis to find the best variable distribution among the latent variables. This model was referred to as the statistical model. This model provided stronger quality indexes than the empirical model provided. However, the latent variables created by the statistical model were not identified as real-world objects. This happened because the statistical model totally ignored the company business model.

Thus, we opted to build a hybrid model which combines both the technical contributions from the empirical model and the techniques from the statistical model. This approach resulted in the best configuration possible, as shown in Figure 4.5. The hybrid model reached significant quality indexes and maintained the interpretability of the empirical model.

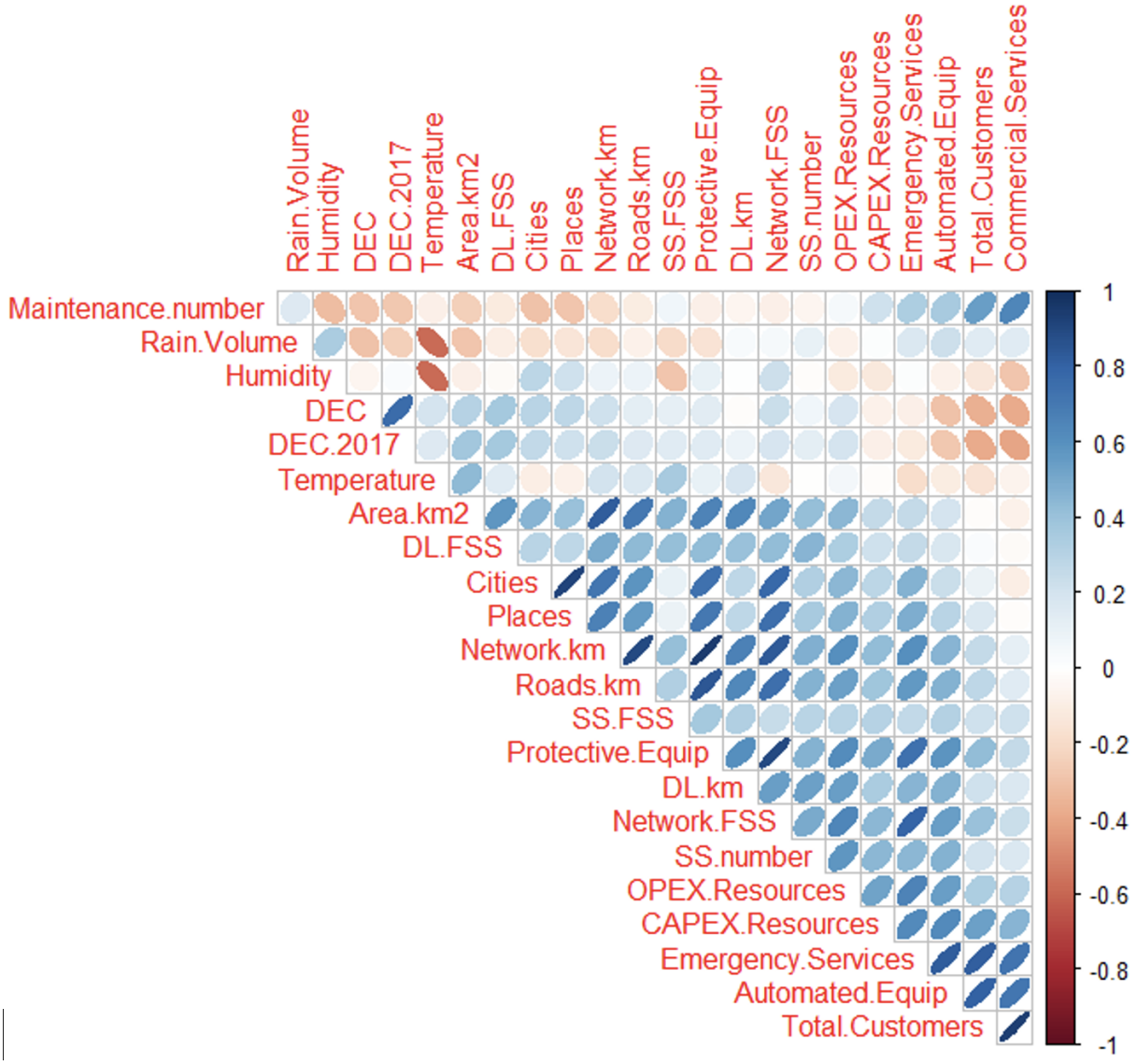


Figure 4.3 – Correlations among all 23 variables from the database.

While the empirical model had four latent variables ("Quality Performance", "Electrical Assets", "Logistic Assets" and "Service Demand"), the hybrid model has nine latent variables ("Electrical Assets 1", "Electrical Assets 2", "Geographical Assets 1", "Climate Variables 1", "Service Demand 1", "Service Demand 2", "Funding Application 1", "DEC 2017" and "Quality Performance 1").

Guidance in choosing and grouping those latent variables was a mix of technical knowledge of the sector and statistical indicators. It is worth noting that the names of the latent variables satisfactorily explain which elements constitute those variables.

The DEC index (in minutes) was chosen as the model's response variable. However, the data show large asymmetry, with distinct values from different electrical groups. Tests using the logarithm of the DEC index and from some predictors indicated more consistent and accurate results. Thus, it was decided that the model would be constructed with the logarithm of the

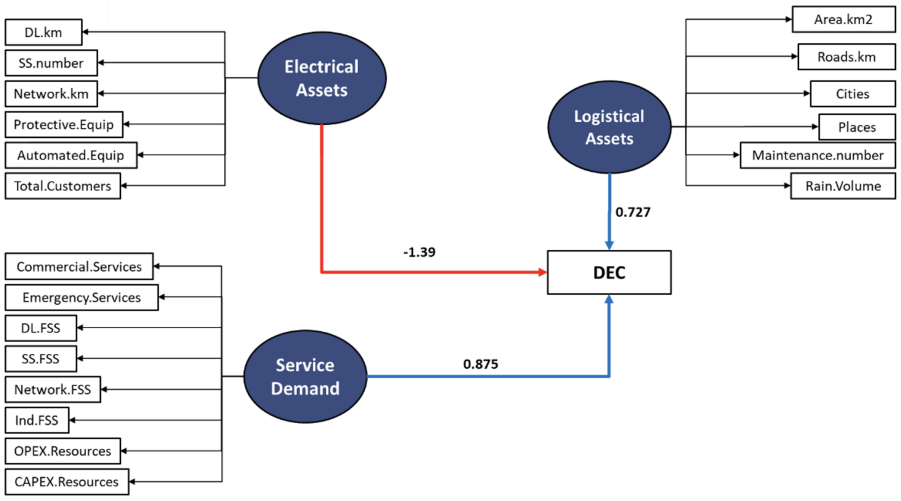


Figure 4.4 – Empirical Structural Equation Model.

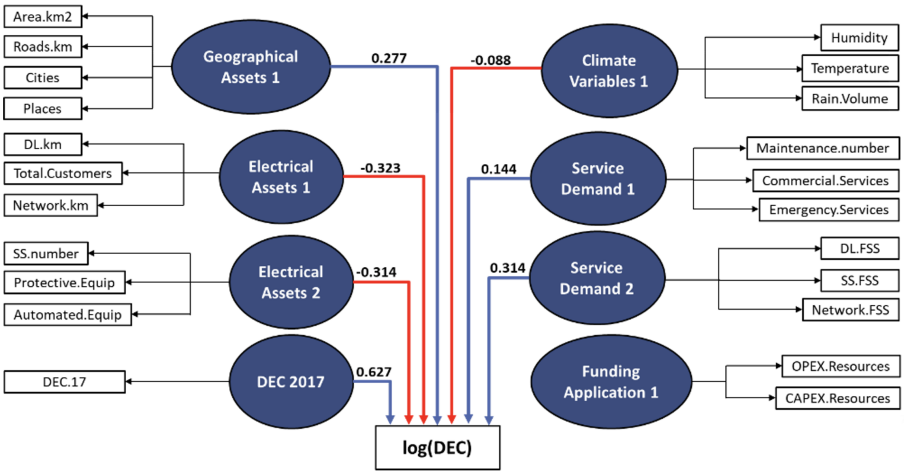


Figure 4.5 – Final Structural Equation Model for the DEC index.

DEC index as the response variable the log of some of the predictors.

The nine latent variables defined in the first layer, i.e., base model, of the hybrid model comprise:

- **Geographical Assets 1 (GA1)**: log(Area.km2), log(Roads.km), Cities, Places;
- **Electrical Assets 1 (EA1)**: log(DL.km), log(Network.km), log(Total.Customers);
- **Electrical Assets 2 (EA2)**: SS.number, log(Protective.Equip), log(Automated.Equip);
- **Climate Variables 1 (CV1)**: log(Humidity), log(Temperature), log(Rain.Volume);
- **Service Demand 1 (SD1)**: log(Maintenance.number), log(Commercial.Services), log(Emergency.Services);
- **Service Demand 2 (SD2)**: DL.FSS, SS.FSS, log(Network.FSS);

- **Funding Application 1 (FA1)**: log(OPEX.Resources), log(CAPEX.Resources);
- **DEC 2017 (DQ17)**: log(DEC.2017);
- **Quality Performance 1 (QP1)**: log(DEC).

Several quality indices such as Cronbach's alpha, loadings, crossloadings, Average Variance Extracted and Goodness-of-fit were used to validate the base model. All of them showed significant results, which validated the model. The results of the regression model for the logarithm of the DEC index, without the latent variable FA1, is shown in Table 4.1. The reason, already shown in Figure 4.5, is that latent variable FA1 is the only one that does not have a statistically significant impact on the DEC index. The others have both positive or negative significant influence on the index.

Table 4.1 – Regression model for the DEC index only using significant latent variables.

Predictors	Estimate	Std. Error	P-value
Intercept	0.0000	0.0311	1.0000
GA1	0.2766	0.0654	0.0000
EA1	-0.3226	0.0750	0.0000
EA2	-0.3144	0.0794	0.0000
CV1	-0.0880	0.0342	0.0107
SD1	0.1444	0.0687	0.0367
SD2	0.3143	0.0633	0.0000
DQ17	0.6273	0.0461	0.0000
Adjusted R^2 : 74.34%			

Among the significant latent variables, the one with the largest absolute estimator, therefore is most impactful, is DQ17. This aligns with the strong correlation between the DEC index from 2017 and 2018 shown in Figure 4.3. The results indicate that the quality performance from an electrical group is highly dependent on and associated with the results from the previous year.

The EA1 variable has the largest negative coefficient, indicating good influence on the DEC index. This implies that the longer the distribution lines and network, the lower the DEC index. The same relation applies to the number of customers; more customers means lower DEC, which is counterintuitive.

The variable EA2 also has a significant negative coefficient. This variable comprises mainly protection and automated equipment, to protect the network when unforeseen events

happen. Investments in more equipment of such a nature reduce the number of interruptions and improve the quality of the service.

The SD2 variable has the second largest positive coefficient, so the higher the value of this variable, the higher the DEC index. It comprises the FSS indices, which are related to interruptions in the energy supply. Ergo, a reduction in the FSS indices causes a substantial reduction in the DEC index.

The variable SD1 also presents a positive coefficient with lower value. Since it comprises services related to the energy supply failure, such as emergency and commercial services, it is as expected.

The GA1 variable showed a positive coefficient, as expected. It comprises the geographical variables which the DSO has no influence on. The results indicate that the larger the concession area and the higher the number of places served, the harder it will be to maintain the DEC index at a satisfactory level.

The variable CV1 has a negative coefficient, indicating that the more it rains, the lower the DEC index. In this case, however, the expected relation is just the opposite: more rain should result in a higher DEC index. This result was analyzed further and subsequently explained, when the CART will be applied in the model.

The result showing that the variable FA1 is not significant requires further clarification. It is reasonable to assume that the more a DSO invests in CAPEX and OPEX, the lower the DEC index would become. Even though it seems counterintuitive, it is possible to understand the causes by analyzing the investment profile of the company in recent years.

In recent years, most of the investments made by CEMIG-D were focused on network expansion, with a much lower amount invested in improving the energy supply quality, as shown in Figure 6. This investment profile is more likely to have no influence on the DEC index since there is no significant investment in quality measures, such as protection equipment. In conclusion, investments in recent years, for the most part, were neither helpful for the DEC index nor were they intended to be.

4.4.2 Multilayer Hybrid Models applied to the DEC index

Figure 4.7 shows the histogram and boxplot for the DEC index. The data show asymmetry to the right, which means few electrical groups had a high DEC index. The DEC index has a mean value of 17.39 minutes, median value of 14.98, minimum value of 2.15 and maximum value of 66.18. Figure 4.8 shows the spatial distribution of the DEC index for each electrical group.

It was in our interest to investigate the effect of potential variables to predict the DEC index. We modeled the DEC index using only the SEM, combining all 22 potential predictors into 8 latent variables.

The final regression model (inner-model), using SEM, was shown in Table 4.1. This model, without the latent variable FA1, explains 74.34% of the variance of the logarithm of the DEC index (R^2_{adj}). The SEM reduces the size of the problem by creating the latent variables.

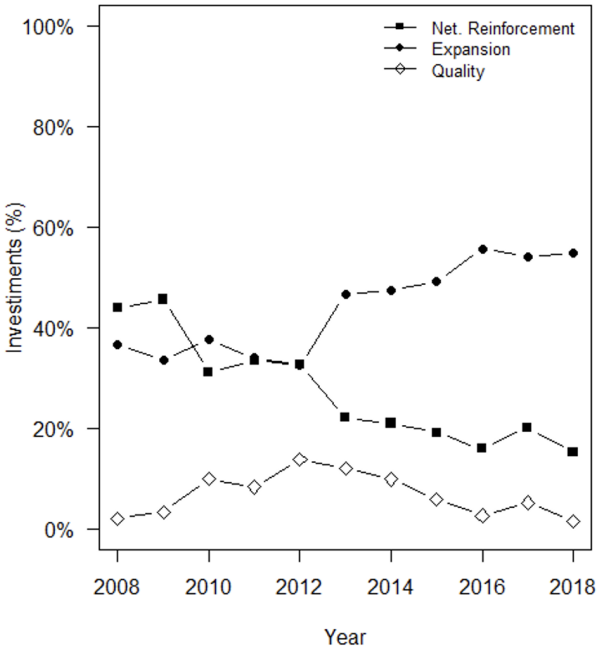
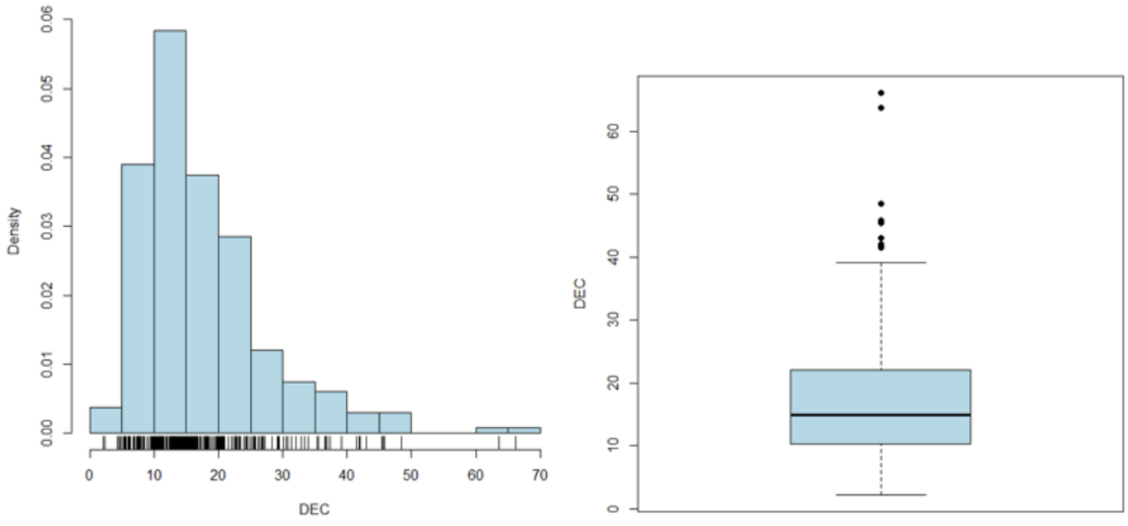


Figure 4.6 – CEMIG-D's investments in network reinforcement, expansion and quality between 2008 and 2018.



(a) **(b)**

Figure 4.7 – Histogram (a) and boxplot (b) for the DEC index.

Ergo, it is correct to assume there has been an information loss in the process.

To further explore the relation between the response variable ($log(DEC)$) and the predictors, Table 4.2 presents the regression for the univariate models for each predictor. The predictors were ordered according to their respective coefficient of determination (R^2). The variable DEC.2017 has the highest R^2 ($R^2 = 68.76\%$), followed by the variables Area.km2 and Commercial.Services. It is also worth noting the results referring to the signs of the coefficients. For example, an increase in the area increases the mean DEC value. Some results are counterintuitive, like the increase of the DEC value with the increase of the protective equipment.

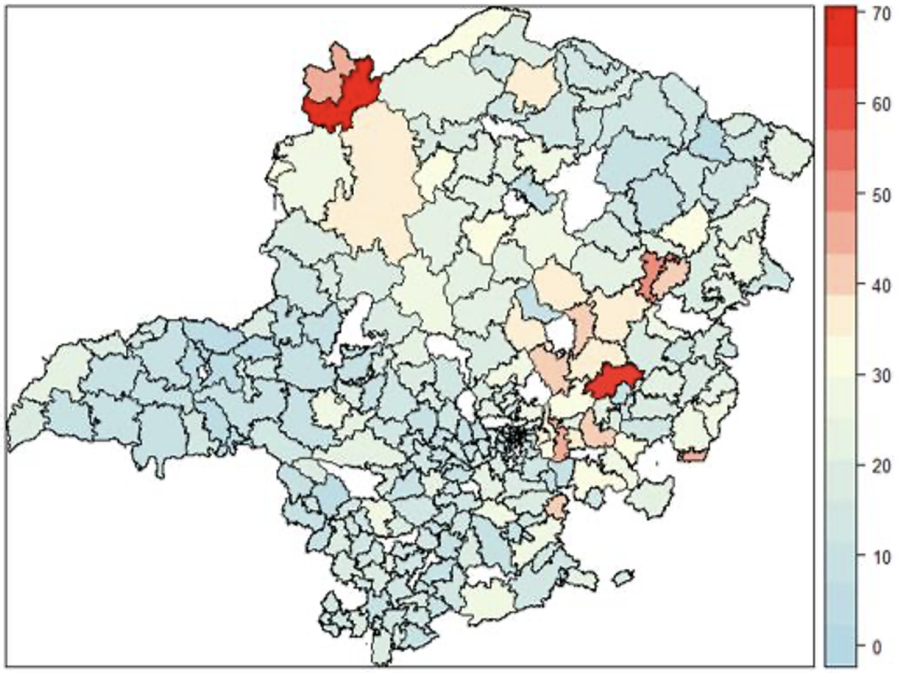


Figure 4.8 – DEC index spatial distribution.

Another univariate analysis using the latent variables was conducted. The results are shown in Table 4.3. The results indicate that the latent variable FA1 is not statistically significant. Among those that are statistically significant, the variables DQ17 and GA1 show the highest explanatory power. Moreover, these variables also have the largest absolute coefficient, indicating they are the most impactful in the DEC index.

To demonstrate the usefulness of a multi-layer model, consider the SEM using log(DEC) as the response variable and all latent variables (GA1, EA1, EA2, CV1, SD1, SD2, FA1 and DQ17) as predictors. The results for this model, including the latent variable FA1, are shown in Table 4.4. This model resulted in an adjusted coefficient of determination (R^2_{adj}) of 74.35%. This shows that adding the latent variable FA1 increases the R^2_{adj} of 0.01%, i.e., it has no significant impact. Beyond R^2_{adj}, it is also interesting to evaluate the predictive coefficient of determination for the model.

To evaluate the predictive power of such models, an alternative is to use predictive statistics, e.g., the predictive coefficient of determination, or R^2_{pred}. The predictive response of a model can be calculated using cross-validation procedures.

For example, using a leave-one-out cross-validation, each observation is temporarily removed from the database. The model is then adjusted with the remaining observations and the forecast error is defined by the difference between the observed response of the element removed from the database, and the predictive response from the model without that observation as shown in Equation 4.13:

$$error_i = y_i - \hat{y}_i \tag{4.13}$$

Table 4.2 – Univariate regression models for the DEC index using all predictor variables.

Predictors	Estimate	P-value	R^2
DEC.17	0.8695	0.0000	0.6876
Area.km2	0.1821	0.0000	0.2470
Commercial.Services	-0.2499	0.0000	0.2119
Total.Customers	-0.2884	0.0000	0.1853
Automated.Equip	-0.2798	0.0000	0.1322
Cities	0.0707	0.0000	0.1267
DL.FSS	0.0837	0.0000	0.1192
Roads.km	0.1392	0.0000	0.1083
Rain.Volume	-0.7795	0.0000	0.0989
Network.km	0.2320	0.0000	0.0888
Places	0.0563	0.0000	0.0888
Network.FSS	0.2259	0.0000	0.0712
OPEX.Resources	0.1380	0.0014	0.0378
Temperature	2.2730	0.0020	0.0355
Emergency.Services	-0.1286	0.0131	0.0230
Protective.Equip	0.1351	0.0143	0.0224
Maintenance.number	-0.0536	0.0211	0.0199
CAPEX.Resources	-0.1029	0.0282	0.0180
DL.km	0.0406	0.1785	0.0068
SS.FSS	0.0168	0.2686	0.0046
SS.number	0.0252	0.4115	0.0025
Humidity	0.0991	0.8531	0.0001

where y_i is the observed value of the i-th observation and $\hat{y}_i$ is the predictive response for the i-th observation from the model without the i-th observation. Using the predictive errors from

Table 4.3 – Univariate regression models for the DEC index using latent variables as predictors.

Predictors	Estimate	P-value	R^2
DQ17	0.8292	0.0000	0.6876
GA1	0.4286	0.0000	0.1837
SD2	0.3425	0.0000	0.1173
SD1	-0.3121	0.0000	0.0974
CV1	-0.3075	0.0000	0.0946
EA2	-0.1972	0.0012	0.0389
EA1	0.1248	0.0416	0.0156
FA1	0.0786	0.2007	0.0062

Table 4.4 – Regression model for the DEC index using all latent variables.

Predictors	Estimate	Std. Error	P-value
Intercept	0.0000	0.0311	1.0000
GA1	0.2701	0.0657	0.0001
EA1	-0.3217	0.0750	0.0000
CV1	-0.0836	0.0345	0.0160
EA1	-0.3337	0.0815	0.0001
SD1	0.1130	0.0749	0.1330
SD2	0.3152	0.0633	0.0000
DQ17	0.6111	0.0486	0.0000
AR1	0.0587	0.0559	0.2940
Adjusted R^2 : 74.35%			

all observations of the database, it is possible to calculate the R^2_{pred} as shown in Equation 4.14:

$$R^2_{pred} = 1 - \frac{\sum error_i^2}{\sum (y_i - \bar{y})^2} \tag{4.14}$$

where $\bar{y}$ is the sample mean of the response variable.

The adjusted coefficient of determination (R^2_{adj}) and the predictive coefficient of determination (R^2_{pred}) can present different values/results for the same model, which they usually do. The latter is preferable to select models when the main interest/goal is to evaluate the predictive capacity of a model. Therefore, the predictive coefficient of determination will be used as the model selection criteria to predict the DEC index.

Consider two examples of the R^2_{pred} statistic. The first is the model shown in Table 4.4; even though it has an $R^2_{adj} = 74.35\%$, the $R^2_{pred} = 73.50\%$. Second, consider the multiple regression model using all 22 variables as predictors shown in Table 4.5. The results show an $R^2_{adj} = 80.02\%$ and an $R^2_{pred} = 77.84\%$, indicating that the adjusted statistic wrongly overestimated the predictive power of the model.

This is relevant when comparing different coefficients of determination. In general, several variables in a regression model often result in a high coefficient of determination but, taking into account the predictive capacity of the model, the statistical value may be lower.

The regularized estimator l2-norm was used to improve this result. Applying the regularized estimator in the regression model, with all 22 predictor variables, resulted in a $R^2_{pred} = 77.92\%$ and $\lambda = 0.0799$. This R^2_{pred} result is better than that achieved by the SEM model, showing that linear models can achieve, at best, a predictive power of almost 78%.

4.4.3 Structural Equation Models (SEM) and Classification and Regression Trees (CART)

Classification and Regression Trees (CART) are local mean models that create partitions in the original data and estimate an answer using the local mean y (Breiman et al., 1984). Though it is a non-linear model, the regression tree has an intuitive hierarchical representation, which favors the visual interpretation of the model.

The first hybrid model for the DEC index prediction is constructed by combining the SEM and CART model. The SEM is adjusted and then the CART model is also adjusted using the residuals from the SEM as the response variable. All 22 available variables in the database are used as the predictors.

The R^2_{pred} coefficient, in a leave-one-out cross-validation, was chosen to select the number of branches of the CART model and to measure the predictive power of this first hybrid model. Using all 22 variables as predictors, an $R^2_{pred} = 74.56\%$ for the hybrid model is achieved. The result also indicates that the variables Humidity, Network.km and Rain.Volume are the most important.

Figure 4.9 shows the classification tree using only these three variables. In cases where the log of the humidity from the electrical group is greater than 4.3 and the log of the length of the network is less than 6.5, the $\log(DEC)$ must be reduced by -0.37. This means that the DEC estimated by the SEM has to be multiplied by $\exp(-0.37) = 0.6907$, representing a significant decrease of 30.93%. However, if the log of the humidity from the electrical group is greater than 4.3 but the log of the length of the network is greater than 6.5, then the DEC estimated by the SEM has to be multiplied by $\exp(-0.096) = 0.9085$, representing a decrease of 9.15%.

Table 4.5 – Multiple linear regression model using all 22 predictor variables.

Predictors	Estimate	Std. Error	P-value
Intercept	4.6845	3.7736	0.2157
DL.km	-0.0672	0.0204	0.0011
SS.number	-0.0262	0.0177	0.1395
Network.km	-0.2324	0.1532	0.1307
Protective.Equip	-0.2581	0.1362	0.0594
Automated.Equip	-0.0763	0.0587	0.1950
Total.Customers	-0.2201	0.0870	0.0121
Area.km2	0.0810	0.0484	0.0956
Roads.km	-0.0480	0.0428	0.2636
Cities	0.0017	0.0162	0.9141
Places	0.0223	0.0146	0.1275
Maintenance.number	0.0202	0.0135	0.1351
Rain.Volume	-0.3579	0.1266	0.0051
Commercial.Services	-0.1634	0.0666	0.0148
Emergency.Services	0.4099	0.1365	0.0030
DL.FSS	0.0201	0.0087	0.0219
SS.FSS	0.0010	0.0087	0.9114
Network.FSS	0.5875	0.1097	0.0000
OPEX.Resources	-0.0021	0.0331	0.9506
CAPEX.Resources	-0.0346	0.0326	0.2904
Humidity	-1.3149	0.4160	0.0018
Temperature	1.5139	0.6824	0.0275
DEC.17	0.4171	0.0546	0.0000

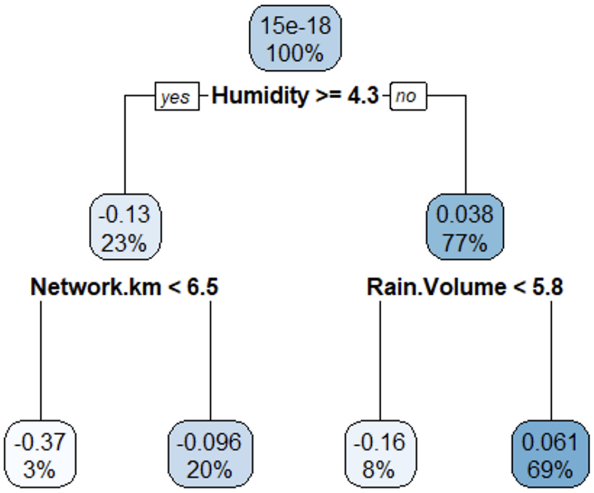


Figure 4.9 – Classification and Regression Tree for the DEC model.

In cases where the log of the humidity is lower than 4.3 and the log of the rain volume is less than 5.8, the DEC estimated by the SEM must be multiplied by $\exp(-0.16) = 0.8521$ (a 14.79% decrease). On the other hand, if the log of the humidity is lower than 4.3 but the log of the rain volume is greater than 5.8, the DEC estimated by the SEM must be multiplied by $\exp(0.061) = 1.0629$ (a 6.29% increase).

The CART model adjusts the SEM result in one of two possible non-linear ways. It increases the 2018 DEC index by 6.29% in electrical groups that present a low humidity and high rain volume. It decreases the 2018 DEC index varying from 9% to 30% in electrical groups that either present a low humidity and a low rain volume or those that present high humidity. This corrects the counterintuitive results regarding the CV1 latent variable using only the SEM, shown previously.

4.4.4 Structural Equation Models (SEM) and Random Forest models

The second hybrid model to predict the DEC index combines the SEM and the Random Forest model. The SEM is adjusted and then the Random Forest model is also adjusted using the residuals from the SEM as the response variable. All 22 variables available from the database are used as predictors. Even though the results from the Random Forest model are not as visible and interpretable as those from the CART model, it is possible to estimate the feature importance for each predictor variable used. The feature importance statistic measures the mean error decrease caused by each predictor variable. This statistic indicates the influence order of the variables in reducing the error of the Random Forest model.

The hybrid SEM + Random Forest model achieved an $R^2_{pred} = 77.65\%$. Figure 4.10 shows the feature importance statistic for each predictor variable, sequenced in decreasing order. The results indicate the most impactful variables in reducing the model error is Humidity, followed by Rain.Volume, and Temperature.

A comparison with the univariate model shown in Table 4.2 has interesting results. These

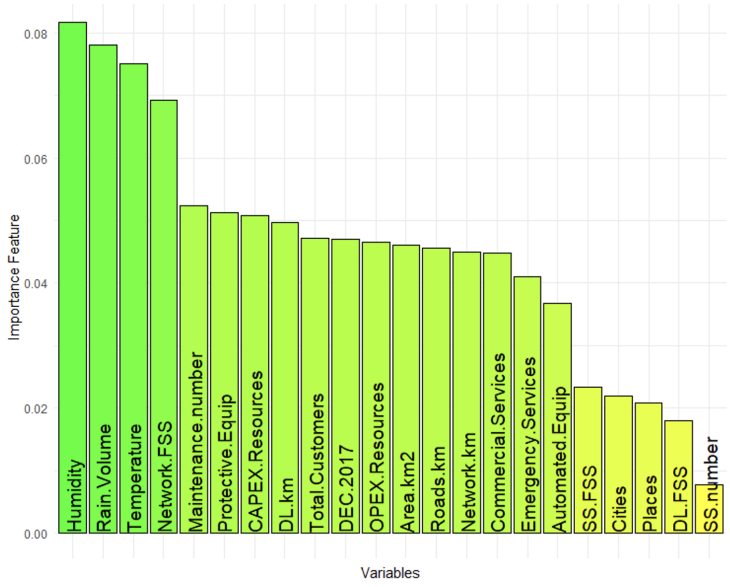


Figure 4.10 – Feature Importance from the Random Forest Hybrid model.

three variables show some of the lowest values for the R^2 when analyzed separately. Humidity, for example, has the lowest R^2 and it is not even significant. However, when analyzed together in the non-linear Random Forest model, they show the largest values in the feature importance chart.

The first 3 variables in the ranking correspond to the latent variable CV1. The latent variable CV1 is not much influent when analyzing exclusively the SEM, but the variables that comprise it are significant in the feature importance chart. Such behavior is not unexpected when we consider the results from the CART hybrid model.

This result is also on par with the results from multiple linear regression model shown in Table 4.5. Humidity and Temperature have the first and second largest absolute coefficient in the model. This implies that the importance of the climatic variables is not captured by the SEM model, which is dominated by the DEC.17 latent variable, but is captured in the linear model and in the non-linear models.

The Random Forest model showed the best predictive power among the non-linear models. The results are close (<1%) to the best model achieved in the research, which was the l2-norm model. Even though the results are close, the hybrid model has the advantage of interpretability, which is not the case for the l2-norm model. This shows the potential of this new method, combining both the high predictive power and the interpretations.

4.5 Conclusion

This paper used a hybrid multi-layer model to evaluate the DEC index for CEMIG-D. A multiple linear regression model with SEM was chosen as the base model. This model was used along with a dataset of 22 predictor variables. After several tests, it showed considerable predictive power ($R^2_{pred} = 74.34\%$). The model also proved interpretable, allowing for acceptance

or rejection of assumptions about the relations among the variables that it comprises.

Next, different non-linear models (black box) were applied on top of the base model in order to increase its predictive power. Among the models tested, such as l2-norm, CART and Random Forest, the former ($R^2_{pred} \approx 77.92\%$) and the latter ($R^2_{pred} \approx 77.65\%$) presented the best results.

Despite all the innovative statistical application, which proved effective, the main goal of the present work is not to simply develop a statistical model to explain the DEC index. The main goal is to make technical and objective analysis of the electrical groups from CEMIG-D, from a statistically validated model.

As such, it is possible to use the model as the foundation for simulating investment decisions. The use of a white box model as the base layer makes such analysis easier. It could then be possible to predict what the impact on the DEC index would be in a specific electrical group when the number of substations or the grid length is increased.

Moreover, it becomes possible to explore different strategies and scenarios referring to the DEC index. Currently unanswered questions include "What is the DEC index impact caused by the rain level?" or "Given a fixed amount of resources, is it best to hire more workforce or acquire more protection equipment?" These questions can be discussed in a more objective and strategic manner.

The division of the database into electrical groups and the geographical components can also be used to determine which are the most critical. The company must comply with a regulatory limit of the global DEC index (the mean of the DEC indices from all electrical groups) and has finite resources. This makes it impractical to invest in all electrical groups equally. It is essential to develop a way to sort the most important and impactful electrical groups.

Future work will focus on developing an optimizer based on the aforementioned simulator. The underlying idea is to provide different investment options to CEMIG-D from constrained financial resources and an operational/regulatory set of restrictions. The goal of the optimizer is not to show the best strategical path, but rather to show different paths that can lead to the same outcome, thereby enriching discussions on this topic.

Parte III

Artigo submetido em congresso

5 Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico

Lista de Siglas

AGV: Agência Virtual; ANEEL: Agência Nacional de Energia Elétrica; BAR: Base de Anuidade Regulatória; BRL: Base de Remuneração Líquida; BRR: Base de Remuneração Regulatória; CAIMI: Custo Anual das Instalações Móveis e Imóveis; CAOM: Custo de Administração, Operação e Manutenção; CAPEX: Capital Expenditure; CART: Classification And Regression Trees; CEMIG-D: Companhia Energética de Minas Gerais - Distribuição; DEC: Duração Equivalente de Interrupção por Unidade Consumidora; DNR: Despesas Não Reconhecidas pelo regulador; EBITDA: Earnings Before Interest, Taxes, Depreciation and Amortization; FEC: Frequência Equivalente de Interrupção por Unidade Consumidora; GBDCD: Gaussian Bayesian Detection of Cluster and Discontinuities; GD: Geração Distribuída; GESEL: Grupo de Estudo do Setor Elétrico do Instituto de Economia; IPCA: Índice Nacional de Preços ao Consumidor Amplo; LAJIDA: Lucros antes de juros, impostos, depreciação e amortização; MEE: Modelos de Equações Estruturais; OPEX: Operational Expenditure; O&M: Operação e Manutenção; PMSO: Pessoal, Material, Serviço de Terceiro e Outros; QRR: Quota de Reintegração Regulatória; RC: Remuneraçãod de Capital; RCOV: Resources-Capabilities; RECOMP: Recursos e competências; RI: Receita Irrecuperável; SEM: Structural Equation Models; SEP: Sistema Elétrico de Potência; TUSD: Tarifa de Uso do Sistema de Distribuição; URA: Unidade de Resposta Audível; VA: Valor Arrecadado; xgboost: eXtreme Gradient Boosting.

Resumo

O presente estudo tem por objetivo realizar a Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando conhecimento técnico e bases de dados no período de 2018 a 2020. Este trabalho está vinculado ao P&D-636, de mesmo nome, realizado em parceria entre a Companhia Energética de Minas Gerais e a Universidade Federal de Minas Gerais (UFMG). A modelagem ocorreu em três etapas: modelagem do framework de negócio da CEMIG; desenvolvimento de um modelo explicativo/preditivo para o DEC/Compensações Financeiras e modelo explicativo/preditivo para o valor arrecadado. Para tanto, foram utilizados os modelos Canvas, Modelos de Equações Estruturais, Modelos Lineares, Modelos de Machine Learning e Modelos Híbridos. Os resultados permitem compreender melhor o funcionamento da empresa e as variáveis que mais impactam no seu negócio e os modelos estatísticos e computacionais implementados podem ser utilizados para o auxílio a tomada de decisões operacionais.

Palavras-chave: Modelo de Negócio, Canvas, Modelos Híbridos.

5.1 Introdução

A Companhia Energética de Minas Gerais realizou o projeto de P&D intitulado "Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico" (Projeto P&D-636), que está sendo executado em conjunto com a Universidade Federal de Minas Gerais (UFMG). O presente trabalho apresenta os resultados do projeto que teve início em agosto de 2018 e término em julho de 2021.

A relevância para o desenvolvimento do projeto foi a necessidade de implementar ferramentas estatístico-computacionais no contexto da regulação das empresas distribuidoras, especificamente para a CEMIG-D. O ambiente regulado no qual a empresa atua impõe obstáculos complexos de serem superados sem o suporte tecnológico atualmente disponível. Assim, o principal objetivo do projeto é a modelagem do negócio da CEMIG-D utilizando diversas ferramentas estatístico-computacionais e com o suporte de dados da empresa.

O ponto de partida foi a elaboração de uma representação gráfica do modelo de negócio da empresa e do setor de distribuição. As representações foram desenvolvidas em reuniões com a equipe da CEMIG-D e da UFMG. Na análise, foram utilizados os frameworks de Ciclos Virtuosos/Marginais, Causa e Consequência (nível macro e micro) e Canvas. Os resultados permitiram uma melhor compreensão de todas as variáveis que interferem nas operações da empresa assim como direcionar a coleta de dados para os modelos estatísticos.

Outra contribuição da aplicação da metodologia de modelos de negócio foi a definição da principal variável para a CEMIG-D. Com base nos resultados encontrados no estudo realizado pelo Grupo de Estudo do Setor Elétrico do Instituto de Economia (GESEL) da UFRJ (Castro et al., 2017), definiu-se o LAJIDA como a variável resposta e mais importante para representar o desempenho da CEMIG-D. O LAJIDA é composto, de forma simples, pela diferença entre as Receitas e o OPEX. Assim, foram definidos dois modelos, um para cada variável.

Duas variáveis foram selecionadas como as mais representativas para modelar o OPEX: o indicador DEC e as Compensações Financeiras. Para realizar essa modelagem foram coletadas variáveis operacionais, regionais, geográficas, financeiras e climáticas. Os resultados se mostraram promissores ao permitir a construção de um modelo altamente interpretável e com boa capacidade preditiva.

Para o modelo da Receita optou-se pelo valor arrecadado como variável mais representativa. Para a sua modelagem, buscou-se coletar variáveis climáticas, de investimento, mercado e geração distribuída. Os resultados também foram considerados relevantes, mesmo verificando o fato de a Receita ter sido afetada de forma significativa pela pandemia de COVID-19. Esse fato inviabilizou a criação de um modelo preditivo consistente ao longo dos anos, mas indicou resultados promissores na criação de modelos explicativos para cada ano presente na base de dados.

Este trabalho está estruturado, além da introdução, em uma explanação, sobre Modelos de Negócios (item 5.2), Modelos de Equações Estruturais (item 5.3) e Modelos Híbridos (item

5.4). No item 5.5 são apresentados e discutidos os resultados dos modelos. No item 5.6 tem-se as considerações finais.

5.2 Modelo de negócio

Segundo Wirtz et al. (2016), modelo de negócios é a representação simplificada e agregada das atividades relevantes de uma empresa, descrevendo como ela captura valor no mercado. Neste contexto, é importante destacar a estrutura conceitual de um negócio que é definido a partir de três componentes principais: (a) a estratégia de negócio, que ocupa o topo da estrutura, (b) o(s) modelo(s) de negócio(s) e (c) o plano de negócio. A estratégia de negócio envolve uma missão, um posicionamento da empresa em relação ao cenário de negócios. Como concepção, a estratégia aponta para o futuro. O modelo de negócio estrutura a lógica de captura de valor da empresa, oferecendo maneiras coerentes para implantar a sua estratégia. O plano de negócio delineia as ações concretas que colocarão o modelo de negócio em prática. Em suma, há uma ordem hierárquica para a estratégia de negócio, que ocupa o topo da estrutura, os modelos de negócio e os planos de negócio, conforme ilustra a Figura 5.1.

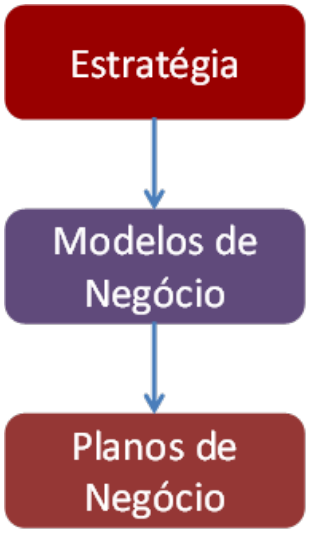


Figura 5.1 – Estrutura hierárquica conceitual de um negócio definida pela estratégia do negócio, o(s) modelo(s) de negócio e o plano de negócio.

Existem, na literatura especializada, diferentes estruturas ou frameworks para estruturar, definir ou desenhar um modelo de negócio. Neste sentido, frameworks são estruturas genéricas que apresentam os componentes centrais de um modelo de negócio.

Casadesus-Masanell e Ricart (2011) apresentam o framework de escolha e consequência. Este framework tem como foco a rede de atividades da empresa e permite identificar os círculos virtuosos que otimizam a captura de valor. Também permite identificar os círculos viciosos que comprometem a captura de valor. Esta ferramenta requer um maior tempo de elaboração, pois é necessário identificar as atividades principais e seus interrelacionamentos.

Demil e Lecocq (2010) apresentam o framework RCOV (Resources-Capabilities) ou recursos e competências (RECOMP). Este framework tem como foco os recursos e competências

internos da empresa e apresenta poucos componentes, em sua totalidade, componentes de operacionalização. A sua principal vantagem é a simplicidade de elaboração. Por outro lado, esta ferramenta é direcionada para o ambiente interno da empresa, não permitindo uma análise mais completa do modelo de negócio com relação ao mercado.

Osterwalder e Pigneur (2010) apresentam o framework Business Model Canvas que tem como foco a criação de novos modelos de negócio ou a avaliação de modelos existentes. O Canvas apresenta nove componentes bem intuitivos. São esses (i) segmento de clientes, (ii) proposição de valor, (iii) canais, (iv) relacionamento com clientes, (v) fluxo de receitas, (vi) recursos chave, (vii) atividades chave, (viii) parcerias chave e (ix) fluxo de receitas. Basicamente os itens (i) a (v) representam o fluxo de receitas do Business Model Canvas, ao passo que os itens (vi) a (ix) representam a estrutura de custos. Destaca-se que a utilização desse framework pode se tornar complexa caso a equipe de trabalho seja extremamente detalhista. É importante destacar que um modelo de negócio deve ser parcimonioso.

5.3 Modelo De Equações Estruturais

De acordo com Pugesek, Tomer e Eye (2003), Modelos de Equações Estruturais (MEE) – ou em inglês Structural Equation Models (SEM) – são um conjunto de técnicas, que combinam análise fatorial e análise de regressão múltipla, muito utilizada em diversas áreas do conhecimento, com destaque especial para biologia, economia, marketing e medicina. SEM é uma família de modelos estatísticos que procura explicar as relações entre múltiplas variáveis, semelhante a uma série de equações de regressão múltipla, entretanto, geralmente as variáveis são fatores não observáveis ou latentes.

De forma geral, os SEM representam as traduções de várias relações de causa e efeito que o pesquisador suspeita que sejam verdadeiras. Tais relações são descritas por parâmetros que indicam a magnitude do efeito (direto ou indireto) que as variáveis independentes (sejam elas observadas ou latentes) causam nas variáveis dependentes (observadas ou latentes) (Pugesek; Tomer; Eye, 2003).

Os modelos podem ser divididos em confirmatórios e exploratórios: no primeiro caso, deseja-se comprovar um conjunto proposto de relações; no segundo, deseja-se desenvolver uma teoria através de repetidas aplicações em um mesmo conjunto de dados. As variáveis que compõem os modelos também são divididas em dois grupos: as variáveis observáveis e as variáveis latentes. As variáveis observáveis são aquelas que possuem valores medidos, seja através de pesquisas ou coleta de dados. As variáveis latentes, ou constructos, são variáveis hipotéticas ou teóricas que não podem ser observadas diretamente, tornando difícil medir a sua existência ou a sua influência.

As variáveis também podem ser divididas em variáveis exógenas, ou seja, que causam, mas não são causadas; variáveis endógenas, isto é, que são causadas, ou variáveis mediadoras, que causam e são causadas. A representação mais comum para esse modelo é uma forma de diagrama com símbolos e setas.

5.4 Modelos Híbridos

O modelo híbrido proposto por Costa et al. (2019a) objetiva criar uma estrutura genérica para melhorar um modelo base, como um modelo estatístico simples (caixa branca). Para melhorar o modelo base, podem ser aplicados modelos não-lineares como Árvores de Regressão, Florestas Aleatórias, xgboost, redes neurais ou combinações de modelos. Por exemplo, um modelo de Florestas Aleatórias pode ser aplicado sobre um modelo de regressão linear múltipla para identificar as variáveis preditoras mais importantes do modelo através do feature importance e melhorar as previsões. O HGB é baseado na distribuição da família exponencial utilizada em modelos lineares generalizados (Nelder; Baker, 1972) e no algoritmo de Gradient Boosting, proposto por Friedman (2001).

O objetivo de criar modelos híbridos consiste em melhorar o desempenho preditivo a partir da combinação de modelos lineares e não lineares. O conceito de Ensemble models ou Boosting (Friedman; Hastie; Tibshirani, 2001) é bem conhecido na literatura. Recentemente, Costa et al. (2019a) propôs o uso de modelos híbridos utilizando o conceito de Gradient Boosting (Friedman, 2001). No caso dos modelos de regressão, é possível combinar diferentes modelos a partir da realimentação dos resíduos de um modelo criando uma nova variável resposta para o modelo seguinte. Formando assim, camadas de modelos. Costa et al. (2019a) propõem utilizar um modelo estatístico linear na primeira camada e modelos não-lineares do tipo machine learning (CART, Random Forests, xgboost, redes neurais artificiais) nas camadas seguintes. A seguir diferentes combinações de modelos são apresentados e avaliados.

5.5 Apresentação e Análise dos Resultados

5.5.1 Desenvolvimento do Modelo de Negócios da CEMIG-D

Para o desenho do framework Canvas do modelo de negócio da CEMIG-D, foi organizada uma pesquisa qualitativa abrangendo um grupo focal de técnicos e gerentes das áreas de regulação e operação da CEMIG-D, totalizando 12 indivíduos. O grupo focal participou de um treinamento de 16 horas divididos em quatro módulos de 4 horas diárias. Durante o treinamento, foram apresentados os conceitos introdutórios sobre Modelos de Negócios e o framework Canvas. Entrevistas semi-estruturadas foram conduzidas utilizando formulários on-line para o preenchimento dos nove campos do framerwork Canvas.

Os formulários foram preenchidos em duas etapas, sendo a primeira referente aos campos de fluxo de receitas e a segunda etapa abrangendo os campos da estrutura de custos. Os formulários foram preenchidos após o segundo e o terceiro módulos do treinamento, respectivamente. Após o preenchimento dos formulários, as respostas foram discutidas e sintetizadas pelo grupo focal ainda nos módulos 3 e 4 do treinamento.

A Figura 2.5 apresenta o círculo virtuoso do modelo de negócio da CEMIG-D, procurando demonstrar quais os ganhos que a empresa de distribuição pode obter ao realizar mais e melhores investimentos. Idealmente, um maior investimento permite: alcançar um maior fornecimento de energia; melhorar a qualidade dos indicadores de Duração Equivalente de Interrupção por Unidade

Consumidora (DEC) e Frequência Equivalente de Interrupção por Unidade Consumidora (FEC); aumentar a Base de Remuneração Regulatória (BRR) (ANEEL, 2016d) e Base de Remuneração Líquida (BRL); aumentar a Quota de Reintegração Regulatória (QRR) (ANEEL, 2016d) e a remuneração do capital; aumentar a receita; reduzir as despesas operacionais (OPEX); maior Lucros Antes de Juros, Impostos, Depreciação e Amortização (LAJIDA) ou EBITDA. Esse incremento no LAJIDA, que ocorre em função do aumento da receita e redução dos custos, melhora a capacidade de reinvestir da empresa, fechando dessa forma o ciclo virtuoso.

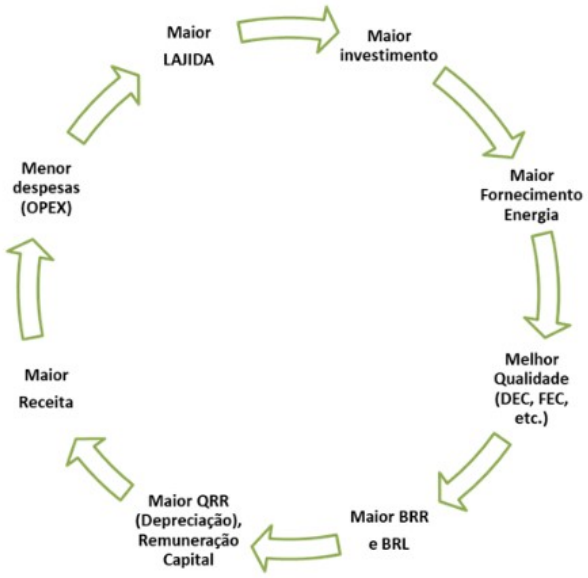


Figura 5.2 – Modelo ótimo de funcionamento de uma concessionária distribuidora de energia elétrica.

A partir desse ciclo eficiente de operação (Figura 5.2), podem ser criados ciclos complementares, como demonstrado na Figura 5.3. Esses ciclos possuem a finalidade de detalhar os ganhos decorrentes em algumas componentes do ciclo principal. Nesse caso, foram escolhidas as três componentes consideradas mais importantes do modelo de negócio da distribuidora: maior CAPEX, maior receita e menor OPEX.

Com relação ao entendimento de cada um dos nove campos do framework Canvas, uma interpretação, segundo o grupo focal de técnicos e gerentes, foi definida e apresentada na Figura 5.4. Com base nesse framework Canvas foi desenvolvido um diagrama do modelo de negócio relacionando as principais atividades e variáveis da CEMIG-D. O principal objetivo deste diagrama é apresentar como se relacionam as diversas receitas, despesas e variáveis presentes nas suas operações.

De forma geral, pode-se dividir as atividades e variáveis da CEMIG-D em cinco grandes áreas, apresentadas na Figura 5.5. São essas: Receita Requerida; Investimentos; Variáveis vinculadas à Receita; Variáveis vinculadas ao OPEX; Variáveis vinculadas ao setor financeiro. Todas as cinco áreas são relacionadas entre si direta ou indiretamente, ou seja, impactos em um campo reverberam em todos os demais. Dessa representação, depreende-se que a divisão da Receita Requerida entre os investimentos, OPEX e Receitas (CAPEX) é a principal variável que impacta nos resultados financeiros da empresa.

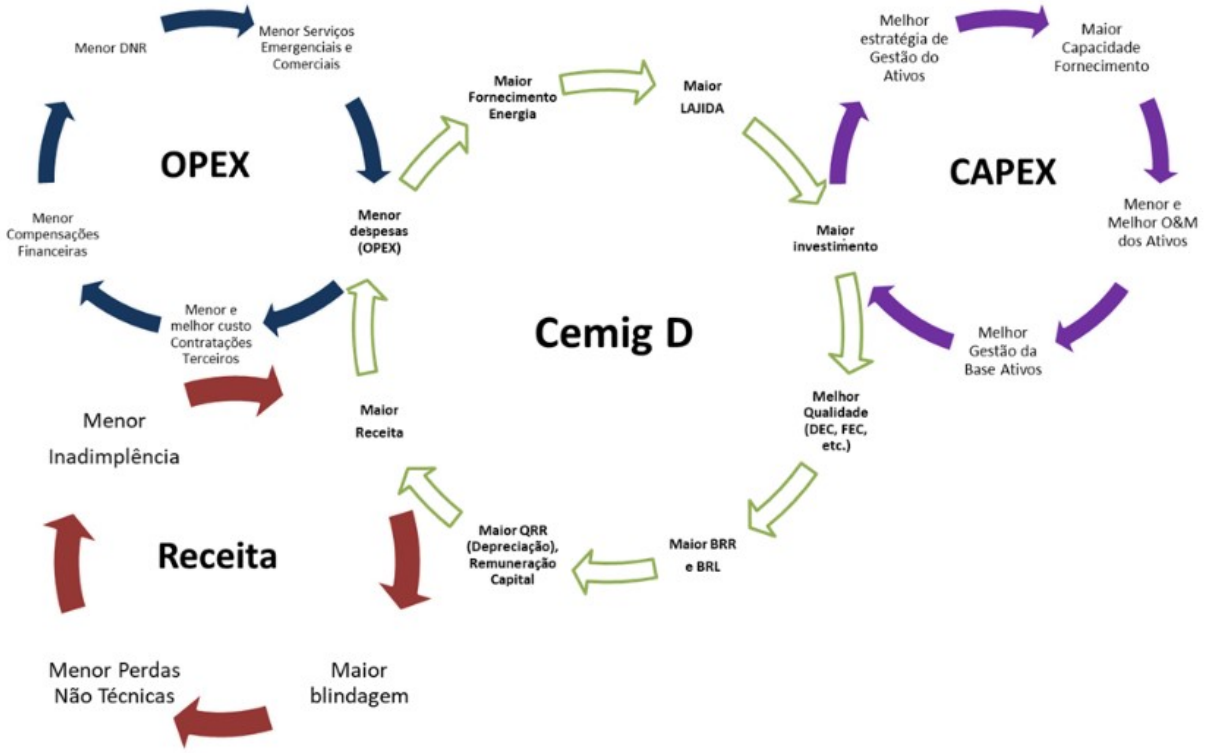


Figura 5.3 – Ciclo marginal do OPEX, CAPEX e Receita do modelo de negócio da CEMIG-D.

A Figura 5.6 expande essa representação para todas as relações entre esses blocos e as variáveis relevantes. O bloco da Receita Requerida reúne as 12 variáveis que formam a receita que a ANEEL permite que as distribuidoras arrecadem, calculado de forma que as permita cobrir seus custos e realizar os investimentos necessários. Impactam diretamente: Compra de Energia; Gastos com Transporte/Encargos; Perdas regulatórias; Receita Irrecuperável (RI); Outras Receitas; Ajustes Financeiros; PMSO Regulatório (CAOM); Remuneração de Capital (RC); QRR; Custo Anual das Instalações Móveis e Imóveis (CAIMI); Ultrapassagem de Demanda; Excedente Reativo.

O bloco de investimentos é formado pelas nove opções que compõem o portfólio de CAPEX da empresa. A receita obtida pela RC e pela QRR devem ser reinvestidas na empresa para renovar os seus ativos, garantir a sua expansão e implementar novas tecnologias em suas operações. Esse portfólio é composto por: Reforço; Mercado/Expansão; Perdas; Medição; Automação; Telecom; Melhoria da Qualidade; Segurança; O&M. O grupo das variáveis vinculadas à receita envolve todas as etapas do processo que impactam diretamente na receita obtida pela distribuidora. Ela aborda desde a compra de energia no mercado, o fornecimento/faturamento das contas, as tarifas, a geração distribuída, a arrecadação e por fim o cálculo da receita real obtida.

O grupo das variáveis vinculadas ao OPEX envolve os próprios consumidores, os serviços emergenciais, comerciais, de perda e de inadimplência, a qualidade do serviço/produto e comercial, compensações financeiras, Fator X, as DNR e o cálculo do custo operacional real.

Por fim, o último bloco envolve as variáveis vinculadas ao setor financeiro e que tem como principal indicador o LAJIDA. Com base nesse lucro calculado, a CEMIG-D consegue calcular as suas receitas e despesas financeiras e o nível de endividamento desejado.

Figura 5.4 – O Modelo de Negócio da CEMIG-D.

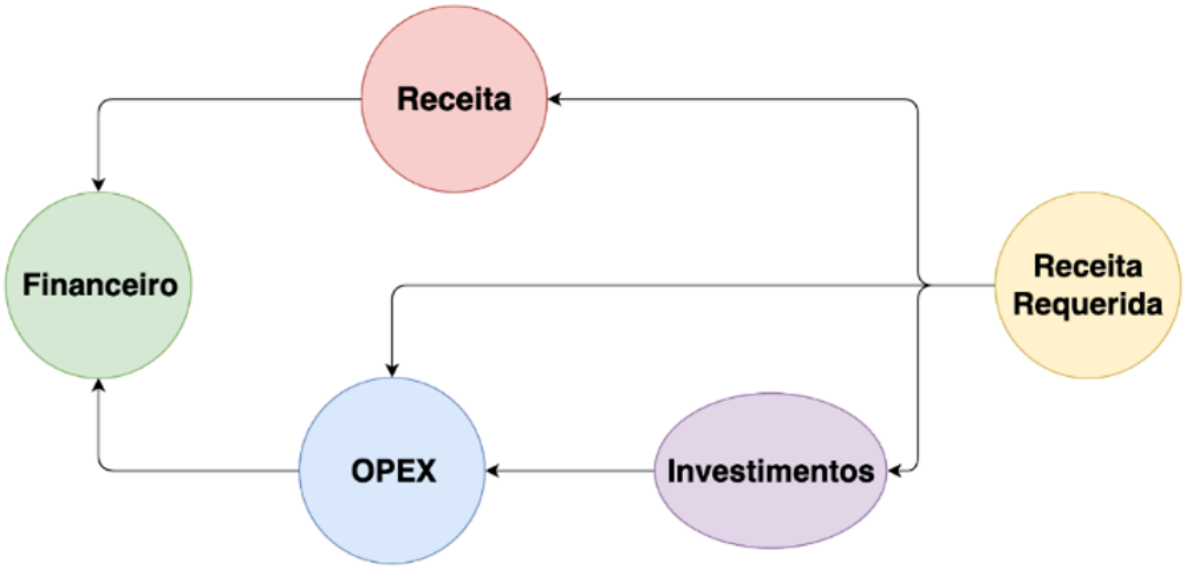


Figura 5.5 – Representação macro das variáveis do modelo de negócio da CEMIG-D.

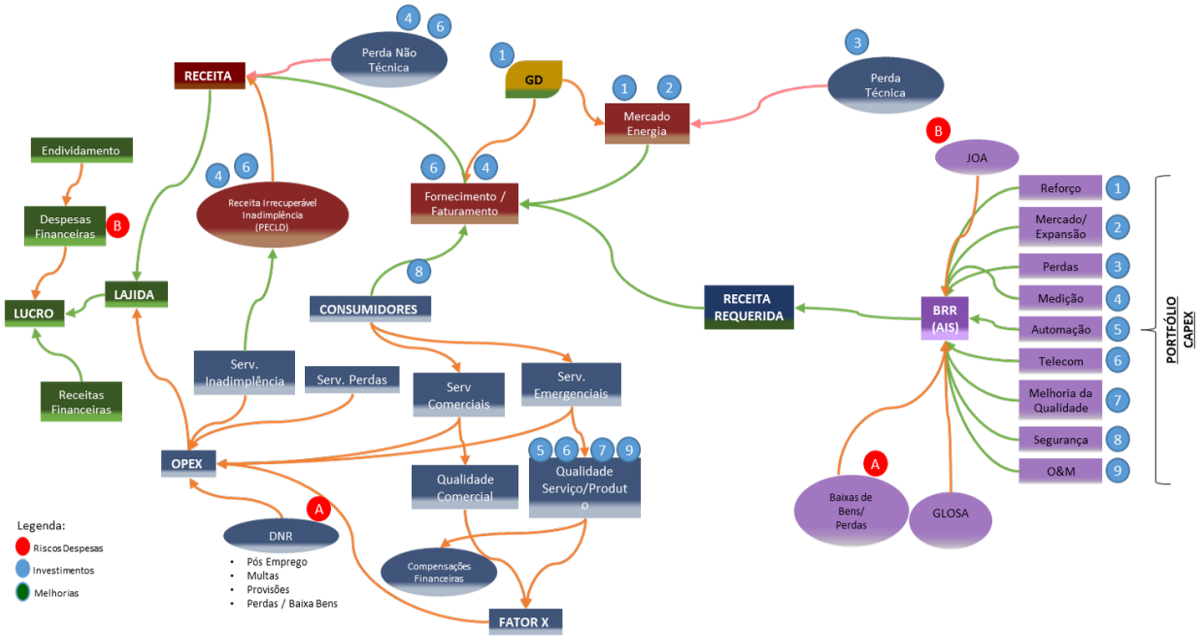


Figura 5.6 – Representação das principais variáveis do modelo de negócio da CEMIG-D.

Fica claro após essa explanação, que a gestão de um negócio de distribuição de energia elétrica é extremamente complexa. A empresa precisa lidar com diversas variáveis, algumas controláveis e outras não, atender aos objetivos regulatórios definidos pela ANEEL e fornecer um serviço de qualidade para a sociedade. Como consequência, o uso de ferramentas estatísticas e computacionais é imprescindível para o desenvolvimento de uma gestão eficiente. Por outro lado, o uso de frameworks qualitativos caracterizam um ponto de partida para a seleção de bancos de dados e para o desenvolvimento de ferramentas que deem suporte a empresa nessa missão.

5.5.2 Desenvolvimento do Modelo DEC/Compensações Financeiras

A metodologia de MEE foi aplicada à uma base de dados disponibilizada pela CEMIG-D no intuito de identificar o comportamento do indicador DEC. Dessa forma, o DEC global da CEMIG-D é calculado a partir da média dos DECs para cada Conjunto Elétrico. No caso da CEMIG-D, até o ano de 2018, a sua área de concessão era dividida em 271 conjuntos elétricos, que representam áreas geográficas disjuntas. Para realizar a análise aqui proposta, foram levantados dados segmentados por conjunto elétrico. A seleção das variáveis que compõem o modelo ocorreu através de reuniões com colaboradores da empresa que definiram um conjunto de variáveis, tecnicamente, as mais impactantes no DEC. A base de dados possui 26 variáveis operacionais, contábeis, financeiras e climáticas referentes ao ano de 2018. Segue a lista de variáveis utilizadas:

1. km.LD.s: extensão das linhas de distribuição (em km);
2. km.Rede: extensão da rede de distribuição (em km);
3. Total.Clientes: número total de clientes atendidos;
4. Quant.SE.s: quantidade (número) de subestações;
5. Equip.Protecao: quantidade (número) de equipamentos de proteção;
6. Equip.Automatizados: quantidade (número) de equipamentos automatizados;
7. Area.km.quad: área de atendimento (em km2);
8. Estradas.km: extensão de estradas na área de atendimento (em km);
9. Municipios: quantidade (número) de municípios atendidos;
10. Locais: quantidade (número) de locais;
11. Forca.de.Trabalho: quantidade (número) de força de trabalho utilizada;
12. Servicos.Comerciais: quantidade (número) de serviços comerciais realizados;
13. Servicos.Emergenciais: quantidade (número) de serviços emergenciais realizados;
14. FSS.Ind: quantidade (número) de interrupções na distribuição de energia por causa de árvores nas linhas individuais;

15. FSS.LD.s: quantidade (número) de interrupções na distribuição de energia por causa de árvores nas linhas de distribuição;
16. FSS.SE.s: quantidade (número) de interrupções na distribuição de energia por causa de árvores nas subestações;
17. FSS.Redes: quantidade (número) de interrupções na distribuição de energia por causa de árvores nas redes;
18. R..OPEX: capital gasto com OPEX (Operational Expenditures - R$);
19. R..CAPEX: capital gasto com CAPEX (Capital Expenditures - R$);
20. Volume.chuva: quantidade de chuva (mm);
21. Vegetacao.km: quantidade de vegetação;
22. Descargas.atm: quantidade (densidade) de descargas elétricas;
23. Vento: velocidade do vento (m/s);
24. umidade: índice de umidade no período (%);
25. temperatura: temperatura média no período ($^{\circ}C$);
26. DEC: indicador DEC (minutos).

O melhor modelo de Equações Estruturais desenvolvido é apresentado na Figura 5.7, alcançando valores significativos nos indicadores de qualidade e mantendo a interpretabilidade do modelo. Esse modelo possui 8 variáveis latentes (Ativos Elétricos 1, Ativos Elétricos 2, Ativos Logísticos, Variáveis Climáticas, Demanda de Serviços 1, Demanda de Serviços 2, Aplicação de Recursos e Desempenho Qualidade). Os dois direcionadores para realizar tal divisão e agrupamentos foram: (i) conhecimento do setor e (ii) indicadores estatísticos. Ressalta-se que os nomes atribuídos à cada variável latente refletem de forma satisfatória as variáveis que o compõem.

Inicialmente, optou-se por utilizar como variável resposta do modelo o valor absoluto do DEC em minutos. Entretanto, os dados apresentaram grande assimetria, com valores muito distintos entre os conjuntos elétricos. Testes utilizando o logaritmo do DEC indicaram resultados mais precisos e consistentes. Assim, optou-se por utilizar o logaritmo do DEC como variável resposta do modelo. As oito variáveis latentes, ou constructos, são compostas pelas seguintes variáveis:

- Ativos Geográficos 1 (AG1): Area.km.quad, Estradas.km, Municipios, Locais;
- Ativos Elétricos 1 (AE1): km.LD.s, km.Rede, Total.Clientes;
- Ativos Elétricos 2 (AE2): Quant.SE.s, Equip.Protecao, Equip.Automatizados;
- Variáveis Climáticas 1 (VC1): umidade, temperatura, Volume.chuva;

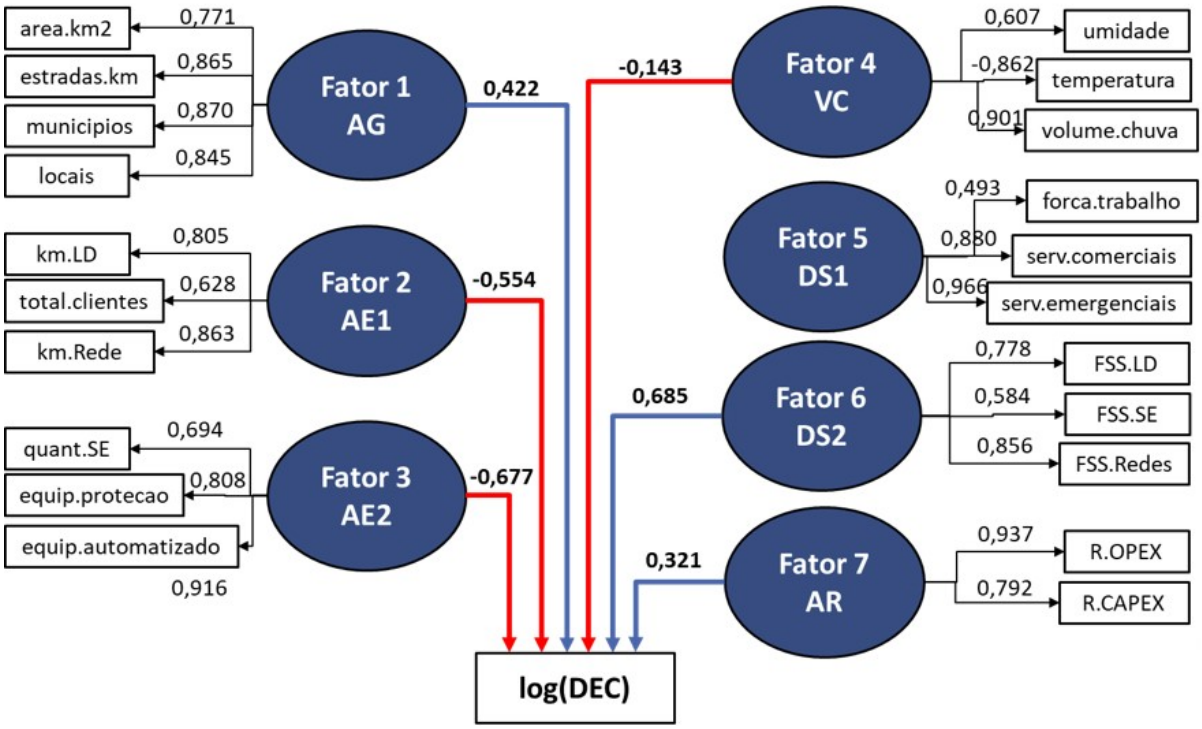


Figura 5.7 – Modelo final de Equações Estruturais para o índice DEC.

- Demanda de Serviços 1 (DS1): Forca.de.Trabalho, Servicos.Comerciais, Servicos.Emergenciais;
- Demanda de Serviços 2 (DS2): FSS.LD.s, FSS.SE.s, FSS.Redes;
- Aplicação de Recursos 1 (AR1): R..OPEX, R..CAPEX;
- Desempenho Qualidade 1 (DQ1): log do DEC.

De posse do modelo validado, iniciou-se a interpretação dos resultados do modelo de regressão para o DEC sem a variável DS1, apresentado na Tabela 5.1. Ressalta-se, conforme percebido na Figura 5.7, que a única variável latente que não possui impacto estatisticamente significativo no DEC é DS1, formada pela força de trabalho e serviços comerciais e emergenciais, enquanto as demais possuem impactos diretos significativos positivos e negativos.

Dentre as variáveis, a que possui o maior valor absoluto no estimador, e por conseguinte, tem o maior impacto no DEC é a variável latente DS2, composta por variáveis relacionadas a falhas no fornecimento de energia, portanto, condizentes com o esperado. Vale destacar que quanto maior o valor do DEC, pior é a qualidade das operações da empresa. Assim, devido ao sinal positivo do estimador, conclui-se que quanto maior a quantidade de interrupções, e consequentemente o valor da variável DS2, maior será o DEC. Logo, demonstra-se que uma redução dos indicadores FSS causa uma redução mais expressiva no indicador.

Considerar um grande conjunto de variáveis preditoras em um único modelo de regressão apresenta inúmeros desafios estatísticos. Devido à presença de colinearidade e multicolinearidade entre as variáveis preditoras, a inferência estatística com relação aos parâmetros do modelo é

Tabela 5.1 – Modelo de regressão linear para o índice DEC utilizando as variáveis latentes significativas (sem a variável DS1).

Variável Preditora	Coeficiente	Erro Padrão	Valor-P
Intercepto	0.0000	0.0401	1.0000
Ativos Geográficos	0.4270	0.0726	0.0000
Ativos Elétricos 1	-0.5832	0.1000	0.0000
Ativos Elétricos 2	-0.7102	0.1086	0.0000
Variáveis Climáticas	-0.1922	0.0424	0.0000
Demanda de Serviços 2	0.8532	0.0825	0.0000
Aplicação de Recursos	0.2215	0.0651	0.0008
R^2 Ajustado: 0.5558			

comprometida. Em geral, maiores serão os valores-P dificultando a identificação das variáveis mais pertinentes para o problema. Por outro lado, a exclusão de variáveis pode implicar na perda de informação preditiva. Caso o interesse do ajuste do modelo seja definir a melhor opção preditiva em detrimento das propriedades estatísticas, pode ser utilizado um estimador viesado, como um estimador do tipo *Ridge Regression* (Hoerl; Kennard, 1970), também conhecido como regularização do tipo L_2. O parâmetro de regularização pode ser selecionado de modo a maximizar a capacidade preditiva do modelo, ou R^2_{pred}. Aplicando o estimador regularizado ao modelo de regressão utilizando as 25 variáveis preditoras obtém-se $R^2_{pred} = 0.5674$ e $\lambda = 92.57$. Este é um resultado interessante pois o valor preditivo alcançado é muito próximo ao modelo sem regularização e ao MEE. Em suma, o uso de modelos lineares permite obter um valor máximo preditivo de 56,74%.

Modelos de árvores de regressão ou classificação (*CART - Classification And Regression Tree*) são modelos de média local que criam partições nos dados originais e estimam uma resposta usando a média local, $\bar{y}$ (Breiman et al., 1984). O modelo CART pode ser representado por uma árvore binária. Cada ramificação da árvore é gerada particionando os dados usando um dos preditores disponíveis. Cada preditor é avaliado separadamente. O preditor e o limiar de corte são escolhidos com base na minimização do erro ou na maximização da função verossimilhança.

Como um primeiro modelo híbrido para predizer o índice DEC, propõe-se combinar o modelo MEE e o CART da seguinte forma. Inicialmente o modelo MEE é ajustado e na sequência o modelo CART é ajustado utilizando como variável resposta os resíduos do modelo MEE e como variáveis preditoras todas as 25 variáveis disponíveis no banco de dados. Para medir a capacidade preditiva desse modelo, e para realizar a seleção do número de ramos do modelo CART foi utilizado o coeficiente de determinação preditivo (R^2_{pred}) em um procedimento de

validação cruzada do tipo *leave-one-out*. Utilizando todas as 25 variáveis preditoras, foi obtido o valor $R^2_{pred} = 0.5714$ para o modelo híbrido. Ao avaliar as variáveis selecionadas no modelo CART na validação cruzada, percebeu-se que a área do conjunto elétrico (Area.km.quad) e o volume de chuva (Volume.chuva) foram predominantes. Dessa forma, foi estruturado um modelo híbrido utilizando somente as variáveis área do conjunto elétrico e volume de chuva no modelo CART. Este novo modelo híbrido obteve $R^2_{pred} = 0.5911$. A Figura 5.8 mostra a árvore de classificação resultante. Considerando que o modelo CART utiliza os resíduos do modelo MEE como variável resposta, caso a área do grupo elétrico seja menor que 20 km2 então o log(DEC) deve ser reduzido em -0.81. Isso significa que, nesses grupos elétricos, o DEC estimado pelo modelo MEE deve ser multiplicado por $\exp(-0.81) = 0.4449$, o que representa uma redução de 0.5331 (53,31%). Caso a área do grupo elétrico seja maior que 20 km2, então caso o volume de chuva seja menor que 327mm então o DEC estimado pelo modelo MEE deve ser multiplicado por $\exp(-0.36) = 0.6977$ (redução de 20,21%); por outro lado, caso o volume de chuva seja maior que 327 mm então o DEC estimado pelo modelo MEE deve ser multiplicado por $\exp(+0.057) = 1.0587$ (aumento de 5,87%). O modelo CART está realizando um ajuste não-linear na resposta do MEE ao aumentar o DEC nos grupos elétricos que apresentam grandes áreas e precipitação.

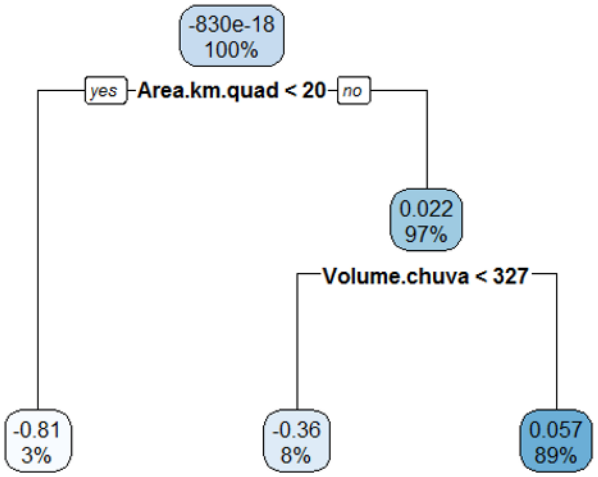


Figura 5.8 – Árvore de Classificação aplicado ao modelo DEC.

Os modelos de Florestas Aleatórias, ou *Random Forests* (Breiman, 2001), são combinações de modelos CART. Cada árvore ou modelo CART é ajustado usando uma amostra aleatória dos preditores e amostras aleatórias das observações. Por exemplo, suponha que m preditores estejam disponíveis, então k preditores ($k << m$) são escolhidos aleatoriamente. Esses k preditores são usados para ajustar um modelo CART. Em sequência, este procedimento é repetido n vezes. Portanto, um total de n modelos CART (ou árvores) são ajustados. O resultado final do modelo de Florestas Aleatórias é uma combinação dos resultados de cada modelo CART. O valor médio ou mediano dos n modelos CART são as estatísticas de agregação mais comuns. O segundo modelo híbrido para predizer o índice DEC combina o modelo MEE e o Random Forests. Inicialmente o modelo MEE é ajustado e logo após o modelo Random Forests é ajustado utilizando como variável resposta os resíduos do modelo MEE e como variáveis preditoras todas as 25 variáveis disponíveis no banco de dados. Embora o resultado do modelo *Random Forests* não seja visualizável como é

o modelo CART, é possível estimar as características de importância (feature importance) para cada variável preditora utilizada. A estatística feature importance mede a redução média do erro atribuída a cada variável preditora. Dessa forma, esta estatística indica uma ordem da influência das variáveis na redução do erro para o modelo *Random Forests*. O modelo híbrido MEE + *Random Forests* obteve $R^2_{pred} = 0.6465$. A Figura 5.9 mostra a estatística feature importance para cada uma das variáveis preditoras. As variáveis foram ordenadas em ordem decrescente do índice (*feature importance*). Os resultados mostram que a variável com maior impacto na redução do erro é representada pelos Serviços Comerciais, seguida pela FSS Redes e pelo Total de Clientes. Por outro lado, as variáveis FSS Redes, Serviços Emergenciais e Equipamentos de Proteção apresentam valores baixos para o R^2, mas apresentaram grandes valores de *feature importance* no modelo de *Random Forests*.

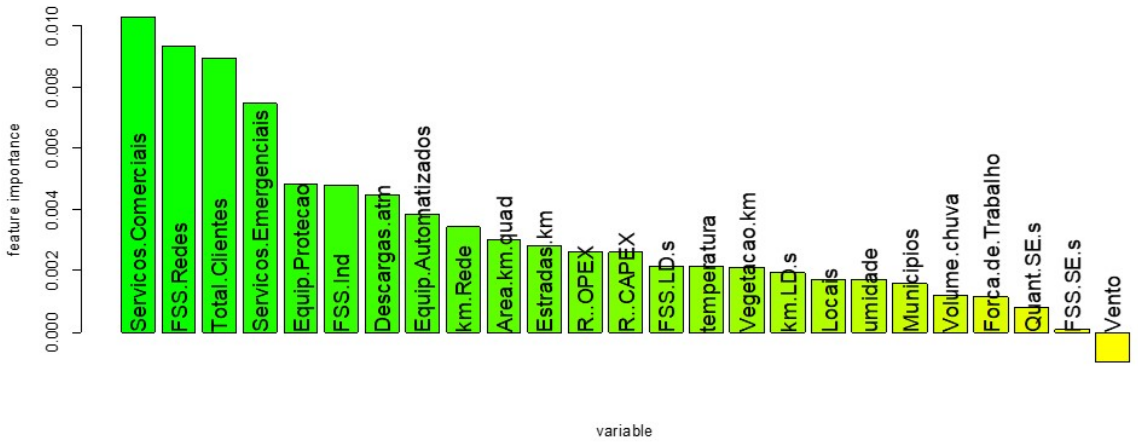


Figura 5.9 – Características de importância (*feature importance*) obtidas utilizando o modelo de Florestas Aleatórias (*Random Forests*).

O *eXtreme Gradient Boosting*, ou xgboost (Chen et al., 2015), é uma implementação rápida do algoritmo *gradient boosting decision tree* (Friedman; Hastie; Tibshirani, 2001). O xgboost usa uma formalização regularizada para controlar o *overfitting* (super ajuste aos dados), oferecendo melhor desempenho. A ideia básica do algoritmo xgboost consiste em realizar a expansão em Série de Taylor da função de perda até a segunda ordem e, portanto, aproximar a minimização da função de perda como um problema de minimização do erro quadrático. Além disso, a função objetivo inclui um termo de regularização ou um termo de complexidade do modelo. Assim, o algoritmo agrega a minimização do erro e da complexidade do modelo. O modelo híbrido MEE + xgboost obteve $R^2_{pred} = 0.6344$, sendo ligeiramente inferior ao modelo híbrido utilizando *Random Forests*.

Os resultados anteriores indicam que o modelo híbrido combinando MEE e *Random Forests* resultaram no maior valor do coeficiente de determinação preditivo. Há ainda de considerar a existência da componente geográfica dos conjuntos elétricos. Para isso, uma terceira camada é proposta utilizando o método Bayesiano de regionalização espacial (Costa et al., 2019b). O método conhecido como Método Bayesiano Gaussiano para Detecção de Conglomerados e Discontinuidades (GBDCD - *Gaussian Bayesian Detection of Cluster and Discontinuities*) tem como objetivo estimar a distribuição a posteriori do número de conglomerados (*clusters*) e das

respectivas partições em um mapa formado por regiões contíguas. O método utiliza o algoritmo de Cadeias de Markov Monte Carlo com saltos reversíveis (*Reversible Jump Markov Chain Monte Carlo*) para varrer o espaço paramétrico e gerar amostras das distribuições a posteriori. Esta metodologia foi inicialmente proposta por Knorr-Held e Raßer (2000) e adaptada por Costa et al. (2019b).

A análise identificou que existe alta probabilidade da existência de dois conglomerados (*clusters*) no mapa, conforme ilustrado na Figura 5.10. A localização geográfica dos *clusters* é apresentada na Figura 5.11a. A Figura 5.11b também mostra os boxplots da distribuição dos resíduos do modelo MEE + *Random Forests* para os dois *clusters* encontrados.

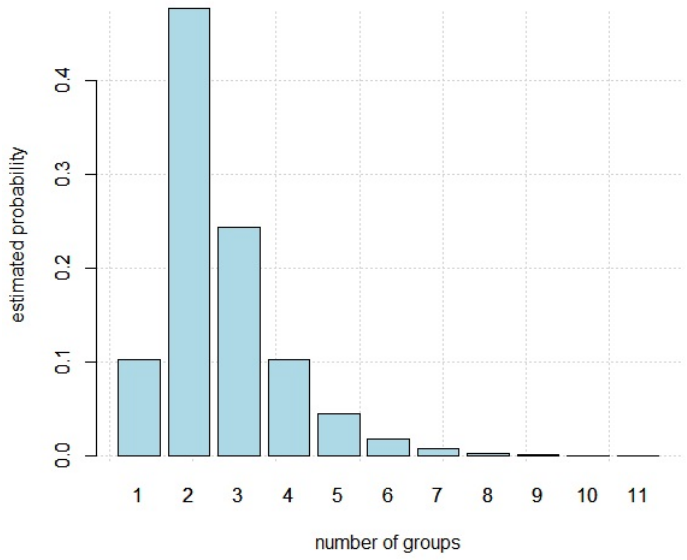
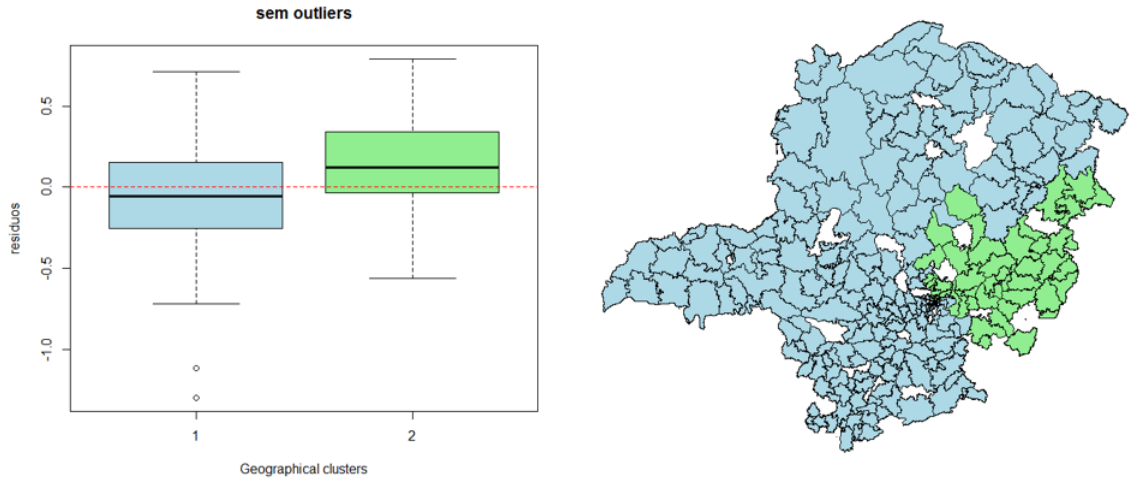


Figura 5.10 – Distribuição a posteriori do número de *clusters*.

Em suma, existe um *cluster* onde a média do índice DEC está acima da média do outro *cluster*. Este *cluster* está localizado à direita do mapa. Na prática, mesmo após as estimativas do índice DEC utilizando os modelos MEE e *Random Forests*, caso o conjunto elétrico esteja localizado no *cluster* 2 (verde) é necessário considerar um aumento médio de $\exp(+0.185) = 1.2032$, ou seja, 20,32% do índice DEC. O modelo híbrido multi-camadas obteve um valor de $R^2_{pred} = 0.6736$. No caso da análise de *clusters*, não foi realizada uma análise preditiva do modelo. Isso porque a estimativa do *cluster* é similar à criação de uma nova variável preditora para o mesmo. Para criação dessa variável, todas as informações dos resíduos, para todos os conjuntos elétricos, foram utilizadas, caracterizando um fenômeno de *data leakage* (vazamento de dados). O coeficiente de determinação ajustado preditivo foi calculado utilizando a resposta preditiva dos modelos MEE e *Random Forests* ao qual foi adicionado os valores médios referentes a cada *cluster*. Finalmente, uma síntese da capacidade preditiva do modelo híbrido multicamadas é apresentada na Figura 5.12. A Figura 5.12a mostra a distribuição do índice DEC no mapa dos conjuntos elétricos. A Figura 5.12b mostra o resultado do modelo híbrido multi-camadas ajustado. A Figura 5.12c mostra o mapa dos resíduos do modelo híbrido multi-camadas e a



(a) Boxplot dos resíduos para os diferentes *clusters*. **(b)** Localização geográfica dos *clusters* detectados.

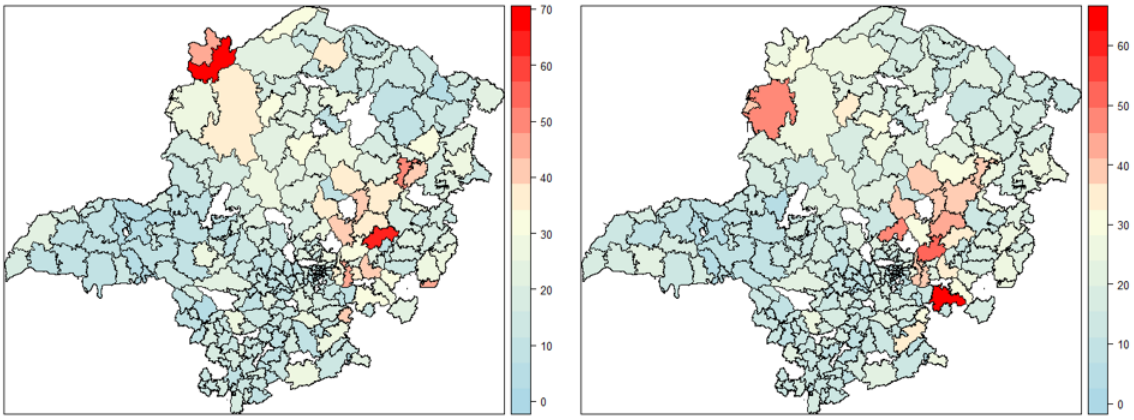
Figura 5.11 – Resultado da análise de *clusters* utilizando o método GBDCD. A distribuição a posteriori indica que o mapa dos resíduos apresenta dois *clusters* com alta probabilidade.

Figura 5.12d mostra os boxplots do índice DEC e dos resíduos do modelo, isto é, a diferença entre o índice DEC observado e o índice DEC estimado pelo modelo. É importante destacar que esta análise foi realizada utilizando o índice DEC e não o logaritmo do mesmo. Destaca-se que, no mapa dos resíduos, valores positivos indicam que o valor observado foi maior que o valor estimado pelo modelo, valores negativos indicam conjuntos elétricos cujo valor observado foi menor que o valor estimado pelo modelo. A Figura 5.12c evidencia que houve uma captura do valor médio do índice DEC pelo modelo híbrido multicamadas. É importante ressaltar que o modelo obteve um coeficiente de determinação próximo de 70% (67,36%) para o logaritmo do DEC. Ou seja, não é possível explicar 30% da variabilidade do índice log(DEC) utilizando as atuais variáveis preditoras.

5.5.3 Desenvolvimento do Modelo da Receita

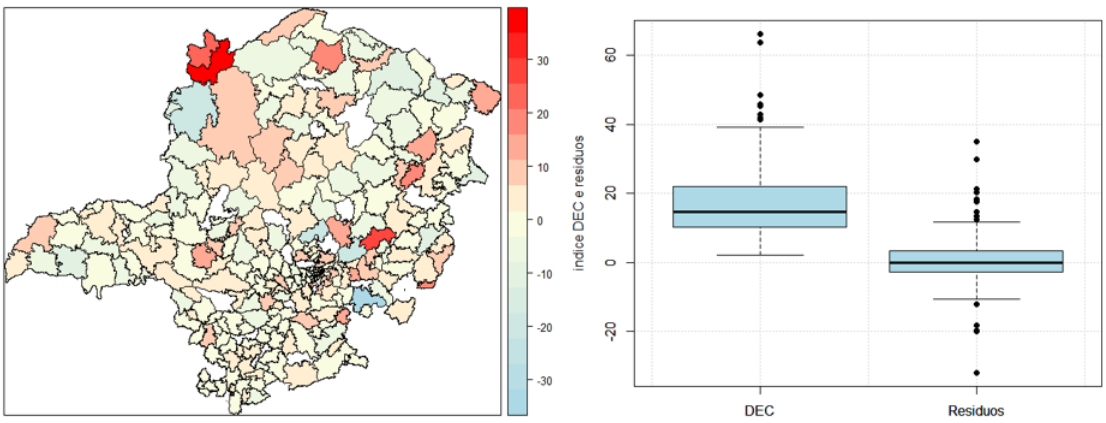
Os modelos para o Valor Arrecadado foram construídos utilizando os aplicativos desenvolvidos no P&D 636. A base de dados utilizada contém variáveis de consumo, quantidade de clientes, investimentos em mercado e proteção de receita, e variáveis climáticas (velocidade do vento, temperatura, umidade e precipitação) para os anos de 2018 e 2019, além do valor arrecadado em 2020.

O primeiro modelo foi denominado Modelo de Atualização. Ele é um modelo de regressão linear simples sem regionalização que utiliza como variável resposta o valor arrecadado no ano t_1 e como variável preditora o valor arrecadado no ano t_0 (sem o intercepto). O objetivo desse modelo é de identificar a tendência (crescimento ou queda) do valor arrecadado ao longo dos anos. Ele simplifica a análise ao identificar uma taxa de atualização para o valor arrecadado que não necessariamente está associada à uma taxa de juros, como o IPCA por exemplo. A representação matemática do modelo é:



(a) Valor do índice DEC observado nos conjuntos elétricos da CEMIG-D.

(b) Valor do índice DEC estimado pelo modelo híbrido multi-camadas (MEE + *Random Forest* + regionalização).



(c) Mapa dos resíduos do modelo híbrido multi-camadas.

(d) Boxplot do índice DEC e dos resíduos do modelo híbrido multi-camadas.

Figura 5.12 – Síntese do ajuste dos modelos híbridos multi-camadas para o índice DEC. Mapa da resposta (a), mapa dos valores estimados (b), mapa dos resíduos (c) e boxplots do índice DEC e dos resíduos do modelo híbrido multi-camadas com regionalização geográfica.

$$V.A._{t_1} = V.A._{t_0} \times \theta \tag{5.1}$$

Onde: $V.A._{t_1}$ = valor arrecadado no ano t_1; $V.A._{t_0}$ = valor arrecadado no ano t_0; θ = taxa de atualização.

Esse modelo foi aplicado para os pares de ano 2018-2019 e 2019-2020. Para o par de anos 2018-2019 a taxa de atualização foi igual a 1,13:

$$V.A._{2019} = V.A._{2018} \times 1{,}13 \tag{5.2}$$

Esse modelo apresentou um poder preditivo alto (R^2_{pred} = 99,56%) e um erro de R$ 28.632.755 (0,13%). Já para o par de anos 2019-2020, a taxa de atualização foi de 0,978:

$$V.A._{2020} = V.A._{2019} \times 0{,}978 \tag{5.3}$$

O modelo também apresentou um poder preditivo considerável ($R^2_{pred} = 99{,}89\%$) e um erro de R$ 118.038.866 (0,54%). Apesar de ambos os modelos apresentarem bons resultados, eles indicam tendências diferentes (crescimento no primeiro e queda no segundo). Isso demonstra que o modelo para um par de anos não apresenta consistência para o período de tempo subsequente. Deve-se levar em consideração que a base de dados compreende dois períodos socioeconomicamente distintos (pré e de pandemia de covid-19), o que pode ter uma grande influência nos resultados.

O segundo modelo construído é denominado Modelo Preditivo Regionalizado. Ele é um modelo híbrido com duas camadas: uma primeira camada de regionalização espacial e a segunda com uma regressão linear otimizada pela função *Step*. A variável resposta é o log da razão do valor arrecadado entre dois anos consecutivos e as variáveis preditoras são referentes ao ano anterior. O objetivo com esse modelo é identificar qual o impacto de um ano no valor arrecadado no ano seguinte. O modelo matemático é:

$$log\left(\frac{V.A._{t_1}}{V.A._{t_0}}\right) = \beta_0 + \beta_1 X_{1_t 0} + \beta_2 X_{2_t 0} + \cdots + \epsilon \tag{5.4}$$

Ao aplicar exponencial nos dois lados da equação tem-se:

$$V.A._{t_1} = V.A._{t_0} \times \exp(\beta_0 + \beta_1 X_{1_t 0} + \beta_2 X_{2_t 0} + \cdots + \epsilon) \tag{5.5}$$

$$V.A._{t_1} = V.A._{t_0} \times \exp(\beta_0) \times \exp(\beta_1 X_{1_t 0} + \beta_2 X_{2_t 0} + \cdots + \epsilon) \tag{5.6}$$

Onde: $V.A._{t_0}$ = valor arrecadado no ano t_0; $V.A._{t_1}$ = valor arrecadado no ano t_1; X_{t_0} = variável preditora no ano t_0; ϵ = erro do modelo.

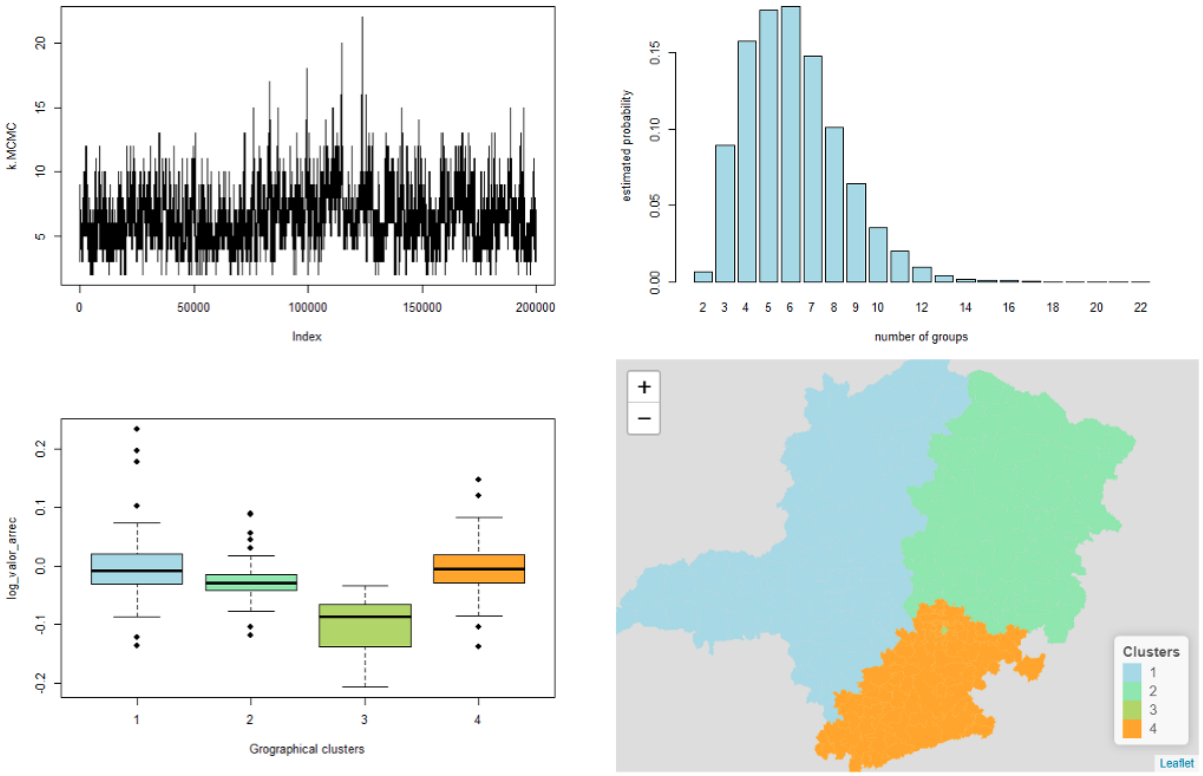


Figura 5.13 – Resultados da regionalização univariada utilizando o log da razão do valor arrecadado entre dois anos consecutivos.

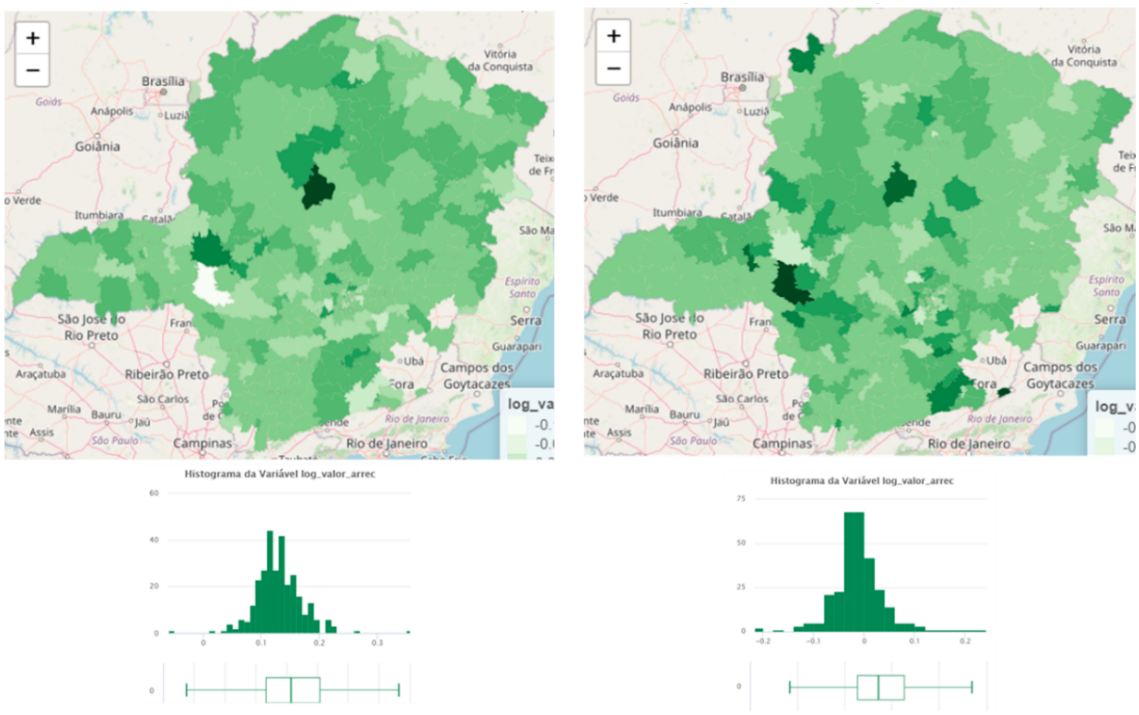
A regionalização utilizada partiu da análise da variável resposta, i.e., o log da razão do valor arrecadado entre dois anos consecutivos. Os resultados são apresentados na Figura 5.13.

Os resultados apresentam alta probabilidade para a existência de seis *clusters*, tal como apresentado na Figura 5.13. Apesar de seis *clusters* serem indicados, considerou-se essa uma divisão muito grande da região. Influenciados por estudos passados, optou-se por utilizar somente quatro *clusters*, que também apresentam uma grande probabilidade. Essas diferenças entre regiões e à cada ano podem ser vistas na Figura 5.14, que apresenta a distribuição da variável resposta em cada ano. É possível identificar as tais diferenças entre as regiões assim como mudanças no perfil da variável entre os anos, como por exemplo na região norte e no triângulo mineiro, conforme Figura 5.15.

Após diversos testes, optou por utilizar a variável Consumo Média Tensão para criar os *clusters*. Os resultados indicaram a existência de quatro *clusters*. O resultado se mostrou consistente ao analisar os dados de 2019 e 2020, com poucas mudanças de um ano para o outro, tal como visto na Figura 5.16.

Com a base de dados disponibilizada foram construídos dois Modelos Preditivos Regionalizados: um para 2019 (utilizando variáveis preditoras de 2018) e um para 2020 (utilizando variáveis preditoras de 2019).

O modelo de 2019 apresentou um $R^2_{pred} = 8{,}03\%$. Nesse modelo, as variáveis preditoras são o valor arrecadado em 2018 e diferentes percentis de umidade. As variáveis e os coeficiente



(a) Distribuição geográfica e histograma do log da razão do valor arrecadado para 2019.

(b) Distribuição geográfica e histograma do log da razão do valor arrecadado para 2020.

Figura 5.14 – Distribuição geográfica e histograma do log da razão do valor arrecadado entre dois anos consecutivos.

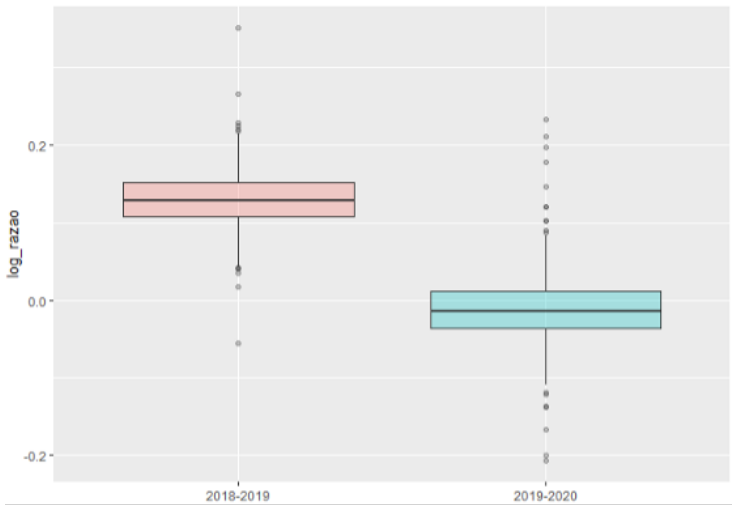
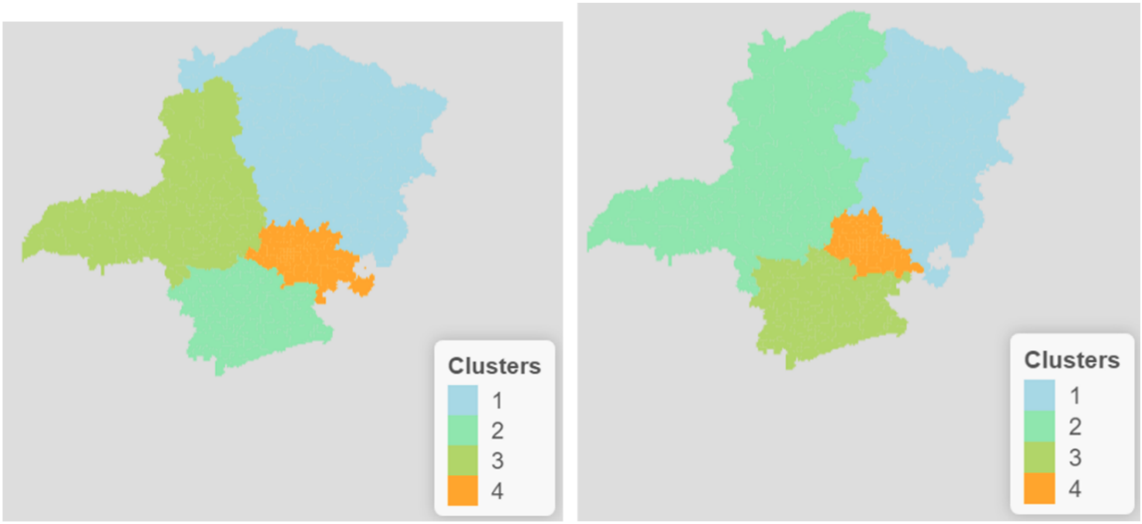


Figura 5.15 – Boxplot do log da razão do valor arrecadado entre 2018-2019 e 2019-2020.



(a) Regionalização de regressão entre o log da razão do valor arrecadado entre 2018 e 2019 e o Consumo Média Tensão para 2019.

(b) Regionalização de regressão entre o log da razão do valor arrecadado entre 2019 e 2020 e o Consumo Média Tensão para 2020.

Figura 5.16 – Regionalização de regressão entre o log da razão do valor arrecadado entre dois anos consecutivos e o Consumo Média Tensão para 2019 e 2020.

de cada *cluster* são apresentados na Tabela 5.2.

O modelo apresentou um erro relativo pequeno de R$ 52.936.770 (0,24%). Os resultados indicam que o *cluster* 1 foi o responsável pela taxa de atualização maior que 1 encontrada anteriormente, enquanto os demais *clusters* puxaram a taxa de atualização para baixo. Além disso, o *cluster* 1 foi o único que apresentou coeficiente menor que 1 para a variável climática umidade.

O modelo de 2020 apresentou um $R^2_{pred} = 15{,}61\%$. Nesse modelo as variáveis preditoras são o valor arrecadado em 2019, variáveis de consumo, quantidade de clientes, investimento e umidade. As variáveis e os coeficiente de cada *cluster* são apresentados na Tabela 5.3.

O modelo apresentou um erro relativo pequeno de R$ 47.897.855 (0,22%). Os resultados indicam que todos os *clusters* apresentaram uma tendência de queda no valor arrecadado, o que era previsto dado o resultado do modelo de taxa de atualização. Destaque para o *cluster 3* que apresentou o menor coeficiente no β_0. Outra curiosidade foi a mudança brusca nas variáveis significativas para cada *cluster*. Enquanto que em 2019 a umidade possuiu uma importância maior, em 2020 esta variável aparece somente no *cluster* 3. Os demais *clusters* apresentaram uma importância menor da variável climática e tiveram como destaque variáveis de consumo e perda GD. Uma explicação para essa diferença pode ser a mudança nas dinâmicas socioeconômicas após o início da pandemia.

Após a construção dos dois modelos foi realizado um experimento no sentido de tentar identificar qual dos dois apresenta mais consistência ao longo dos anos. O experimento foi realizado utilizando o modelo de 2019 para prever o valor arrecadado de 2020, e vice-versa, o modelo de 2020 para fazer a previsão do valor arrecadado de 2019.

O modelo de 2019 está associado à uma tendência de crescimento, enquanto que o de

Tabela 5.2 – Resultado do ajuste dos modelos de regressão múltipla para cada *cluster*.

Cluster 1 (cor azul)	
Variável	Coeficiente
Valor Arrecadado 2018 ($\exp(\beta_0)$)	1.4096824
Umidade (percentil 70%) ($\exp(\beta_1)$)	0.9974004
Cluster 2 (cor verde claro)	
Variável	Coeficiente
Valor Arrecadado 2018 ($\exp(\beta_0)$)	0.8373813
Umidade (percentil 80%) ($\exp(\beta_1)$)	1.0036335
Cluster 3 (cor verde musgo)	
Variável	Coeficiente
Valor Arrecadado 2018 ($\exp(\beta_0)$)	0.9509856
Umidade (percentil 70%) ($\exp(\beta_1)$)	1.0024111
Cluster 4 (cor laranja)	
Variável	Coeficiente
Valor Arrecadado 2018 ($\exp(\beta_0)$)	0.9895782
Umidade (percentil 50%) ($\exp(\beta_1)$)	1.0019806

2020 está associado à uma queda do valor arrecadado. Assim, ao utilizar o modelo de 2019 para prever o ano de 2020 espera-se que o modelo superestime o valor arrecadado, ou seja, que o modelo estime um valor arrecadado acima do valor real. Os resultados indicaram exatamente este comportamento, com um erro de R$ 3.323.758.372 (15,13%).

Já no caso contrário, ao utilizar o modelo de 2020 para realizar a previsão do valor arrecado para o ano de 2019, espera-se que o resultado seja menor que o valor real. Os resultados apresentaram um erro de R$ - 2.770.213.802 ($-12{,}40\%$). Tal como esperado, os dois modelos apresentam um erro considerável ao serem utilizados em anos diferente. Isso indica a dificuldade de construir um modelo preditivo para o valor arrecadado considerando os anos pré e de pandemia.

Tabela 5.3 – Resultado do ajuste dos modelos de regressão múltipla para cada *cluster*.

Cluster 1 (cor azul)	
Variável	Coeficiente
Valor Arrecadado 2019 ($\exp(\beta_0)$)	0.9918644
Consumo Faturado BT ($\exp(\beta_1)$)	0.6581830
Quantidade de Clientes AT ($\exp(\beta_2)$)	1.0133913
Cluster 2 (cor verde claro)	
Variável	Coeficiente
Valor Arrecadado 2019 ($\exp(\beta_0)$)	0.9891619
Consumo Faturado AT ($\exp(\beta_1)$)	1.1023774
Cluster 3 (cor verde musgo)	
Variável	Coeficiente
Valor Arrecadado 2019 ($\exp(\beta_0)$)	0.3632629
Investimento em Mercado ($\exp(\beta_1)$)	1.2358942
Umidade (percentil 80%) ($\exp(\beta_2)$)	1.0129001
Cluster 4 (cor laranja)	
Variável	Coeficiente
Valor Arrecadado 2019 ($\exp(\beta_0)$)	1.0144421
Perda GD ($\exp(\beta_1)$)	0.8359885

5.6 Considerações Finais

Este trabalho realiza uma síntese das metodologias e dos principais resultados decorrentes da execução do projeto de pesquisa e desenvolvimento: "Modelagem estatístico-computacional do modelo de negócio da CEMIG-D utilizando bases de dados e conhecimento técnico". Para tanto, foram pesquisadas e implementadas diversas metodologias tais como frameworks, canvas, equações estruturais, modelos lineares, não-lineares e híbridos e regionalização.

Os frameworks de modelo de negócio desenvolvidos apresentam resultados inovadores para a área de distribuição de energia elétrica. Os diagramas construídos permitem uma maior compreensão do funcionamento do setor e quais as variáveis consideradas mais relevantes.

O foco do trabalho nas principais variáveis que compõem o LAJIDA (DEC/Compensações Financeiras e Receita/Valor Arrecadado) resultou em dois modelos distintos e com alto poder preditivo. O modelo do indicador DEC, ao implementar diferentes modelos híbridos com regionalização, permitiu identificar os principais pontos que devem ser observados pela CEMIG-D para reduzir os seus gastos com OPEX.

Os modelos de valor arrecadado apresentaram uma dinâmica distinta do modelo do DEC. É possível identificar evidências do impacto que a pandemia e a quarentena trouxeram no consumo de energia elétrica. Os modelos ano a ano, apesar de apresentarem bons resultados, também deixam claro a dificuldade de se trabalhar com modelos preditivos em períodos de grande mudança socioeconômica.

Assim, conclui-se que o trabalho foi bem-sucedido na sua proposta ao analisar esse setor com metodologias inovadoras e resultados inéditos. Para o futuro, é de grande interesse continuar as análises com novas bases de dados para validar os modelos propostos e/ou atualizá-los, caso seja necessário.

Parte IV

Conclusão

6 Considerações Finais

Neste capítulo são apresentadas as interfaces desenvolvidas, o resumo dos resultados desta tese e possíveis direções para trabalhos futuros.

6.1 Interfaces desenvolvidas

Nesta seção são apresentadas as interfaces desenvolvidas ao longo da pesquisa de doutorado e do P&D-636. Essas interfaces contemplam grande parte da metodologia que foi desenvolvida e utilizada na pesquisa. Elas foram desenvolvidas em linguagem R através do pacote "shiny". Esse pacote permite a criação de interfaces web com pouco conhecimento de HTML, CSS e Javascript. As interfaces são adaptáveis para computadores, tablets e smartphones.

As interfaces foram desenvolvidas com três objetivos: 1) automatizar grande parte da rotina computacional; 2) facilitar o acesso às rotinas computacionais; 3) permitir o uso da metodologia desenvolvida na pesquisa por usuários leigos em linguagens de programação.

Os três objetivos foram alcançados com as interfaces aqui apresentadas. As interfaces permitem a realização de cálculos de interpolação, regionalização, modelos híbridos, simulação e otimização sem a necessidade de instalar ou abrir o software R. Foi realizado o upload das interfaces em um servidor, o que permite que elas sejam acessadas em qualquer dispositivo com acesso à internet. O usuário precisa selecionar poucos parâmetros para executar cada interface e obter um resultado simples e direto.

A Figura 6.1 apresenta o fluxograma proposto para o uso das interfaces. O seu objetivo é apresentar a sequência lógica para o uso adequado das interfaces. O início ocorre com a demanda de análise de uma base de dados. A base de dados pode estar segregada por municípios ou por conjuntos elétricos.

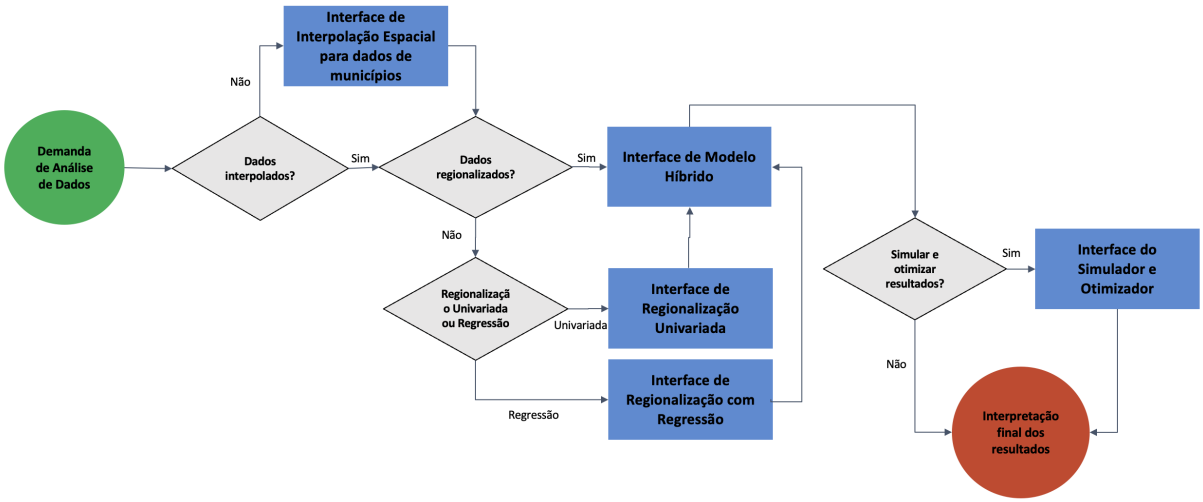


Figura 6.1 – Fluxograma de uso das interfaces.

A primeira interface é denominada “Interface de Interpolação Espacial para dados de municípios”. A demanda para esta interface surge quando a base de dados está segregada em municípios e se deseja converter para conjuntos elétricos. Os dados climáticos utilizados nos Capítulos 5 e 4 foram geradas utilizando esta interface.

A “Interface de Regionalização Univariada” e “Interface de Regionalização com Regressão” são utilizadas quando se deseja dividir os conjuntos elétricos em grupos. Esses grupos podem ser divididos com base em uma única variável (univariada) ou pela relação entre duas variáveis (regressão). Essa regionalização geralmente é utilizada como a primeira camada de um modelo híbrido. Os grupos apresentados nos resultados dos Capítulos 5 e 4 foram gerados utilizando estas interfaces.

A “Interface de Modelo Híbrido” possibilita a criação de modelos híbridos multicamadas em bases de dados divididas por conjunto elétrico. Essa interface adiciona mais duas camadas ao modelo (além da regionalização): uma primeira camada de regressão linear e uma segunda de modelo não-linear (*Random Forest*). Os modelos híbridos apresentados nos Capítulos 5 e 4 foram gerados utilizando esta interface.

A “Interface do Simulador e Otimizador” apresenta ferramentas para análise dos modelos gerados nas interfaces anteriores. As análises consistem de análise espacial, simulação dos resultados de todos os conjuntos elétricos, simulação dos resultados de um conjunto elétrico e otimização dos resultados de um conjunto elétrico.

O Apêndice B contém o manual de uso das interfaces. O manual apresenta: o método de instalação, detalhamento de todas as suas características e elementos gráficos, e exemplos de uso.

6.2 Resumo

Este trabalho de tese se propôs a realizar uma análise do setor de distribuição de energia elétrica (com eventual foco na CEMIG-D) desenvolvendo e aplicando ferramentas de modelagem estatístico-computacionais. Essa abordagem qualitativa e quantitativa permitiu alcançar resultados inovadores no setor.

A característica de monopólio natural inerente ao setor traz consigo diversas questões e dificuldades. A regulação executada pela ANEEL se faz totalmente necessária para mediar os interesses entre as empresas distribuidoras e os clientes consumidores. Mas também é necessário avaliar de maneira objetiva a efetividade da regulação.

O entendimento do setor e as suas inter-relações deve ser o ponto de partida para qualquer análise. A simples questão de quem é o principal cliente de uma empresa distribuidora pode por si só causar desdobramentos que levam a pesquisa para lados opostos. Enquanto que em um primeiro momento pode-se pensar somente no cliente consumidor, deve-se levar em consideração também que as empresas distribuidoras prestam um serviço de gestão dos ativos físicos elétricos para a União. A União é a real proprietária de tais ativos elétricos, e poderia ser considerada o principal cliente das empresas distribuidoras. Após diversas discussões, nossa abordagem se baseou somente na análise do cliente consumidor. Os diversos modelos apresentados (Ciclos virtuosos, Canvas, escolha e consequência) permitem a visualização do funcionamento de uma

empresa distribuidora de forma objetiva e bastante detalhada. A representação na Figura 2.10 se mostrou primordial para o desenvolvimento do restante da pesquisa. A Figura 2.10 permitiu identificar exatamente quais variáveis deveriam ser coletadas e como elas se relacionam entre si.

Os custos operacionais das empresas distribuidoras são objeto de constante controle e regulação pela ANEEL. Além dessa questão, a ANEEL também avalia de forma recorrente a sustentabilidade econômico-financeira das empresas distribuidoras. Essa avaliação ocorre à cada trimestre e separa as empresas em quatro grupos ordenados de acordo com o seu resultado em relação às demais empresas. Embora válido, esse método de avaliação pode ser considerado superficial pois não considera questões como o tipo de controle da empresa (público ou privado) e questões regionais. Nossas avaliações sugerem que ambas as variáveis possuem um impacto significativo no resultado de uma empresa. Vários modelos de regressão ordinal logísticos foram desenvolvidos para validar tais hipóteses. Existe uma clara distinção entre as empresas com controle público e privado, com as empresas públicas apresentando resultados consistentemente inferiores ao das empresas privadas. Um comportamento semelhante ocorre ao avaliar questões regionais. Empresas nas Regiões Sul e Sudeste apresentam resultados consideravelmente superiores quando comparados com empresas na região Norte. Além disso, por avaliar os resultados quase que exclusivamente por meio de comparações, o método utilizado pela ANEEL peca em avaliar o desenvolvimento geral do setor.

O índice de Duração Equivalente de interrupção por unidade consumidora (DEC) é utilizado pela ANEEL como o principal indicador para avaliar a qualidade do serviço de uma empresa distribuidora. O DEC pode ser facilmente calculado pela média de interrupções dos conjuntos elétricos de uma distribuidora. Mas determinar quais variáveis impactam e o quanto impactam não é trivial. O nosso modelo proposto abrange diversas variáveis de diferentes tipos (operacionais, climáticas, financeiras). Essas variáveis foram reduzidas (agrupadas) em variáveis latentes utilizando Modelos de Equação Estruturais. Em seguida aplicou-se a metodologia de modelos híbridos multicamadas com duas camadas: modelo de regressão linear e modelos não-lineares. Foram avaliadas combinações de modelo de regressão linear com modelos CART e *Random Forest*. O modelo final apresentou poder preditivo significativo ($R^2_{pred} \approx 78\%$) e indicou algumas variáveis importantes no resultado, como humidade, volume de chuva e temperatura.

Os resultados alcançados nesta tese se mostram muito relevantes tanto pelo ponto de vista metodológico quanto pela abordagem. Os modelos híbridos propostos e apresentados são válidos e significativos. A abordagem inovadora no setor apresentou resultados nunca antes vistos e servem como alicerce para o desenvolvimento de mais pesquisas nesse setor. As interfaces criadas durante o projeto possibilitam a realização de pesquisas avançadas sobre o setor sem a necessidade de conhecer linguagens de programação. As interfaces também apresentam flexibilidade para o uso de diferentes variáveis e diferentes empresas distribuidoras. Isso as torna uma ferramenta com grande potencial para análise e estudo do setor elétrico.

6.3 Trabalhos futuros

Nesta seção são sugeridos direcionamentos para trabalhos futuros:

- Aprofundar o modelo híbrido multicamadas para a receita;
- Aplicar modelos DEA, malmquist e tornqvist para análise da sustentabilidade econômico-financeira das distribuidoras;
- Aplicar modelos econométricos para análise da sustentabilidade econômico-financeira das distribuidoras;
- Agregar dados mais recentes à análise de sustentabilidade econômico-financeira das distribuidoras;
- Validar o modelo híbrido multicamadas para o DEC em anos mais recentes;
- Coletar base de dados com mais variáveis relacionadas ao DEC e à receita;
- Atualizar o modelo de negócio com novas tendências de mercado (e.g. geração distribuída).

Referências

Abdelkafi, N.; Täuscher, K. Business models for sustainability from a system dynamics perspective. *Organization & Environment*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 74–96, 2016. Citado na página 34.

ABRADEE. *Tarifas de energia*. 2018. Acessado em: 11-06-2019. Disponível em: <https://www.abradee.org.br/setor-de-distribuicao/tarifas-de-energia/>. Citado 2 vezes nas páginas 18 e 19.

Agresti, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018. Citado na página 53.

Aguiar, F. S. et al. Classification and regression tree (cart) model to predict pulmonary tuberculosis in hospitalized patients. *BMC pulmonary medicine*, Springer, v. 12, n. 1, p. 1–8, 2012. Citado na página 77.

Andrade, G. N. d. et al. Contribuição para o desenvolvimento de uma metodologia de avaliação da eficiência no setor de transmissão de energia elétrica. Programa de Pós-graduação em Engenharia de Produção, 2008. Citado na página 68.

Andrade, G. N. de et al. Evaluating electricity distributors efficiency using self-organizing map and data envelopment analysis. *IEEE Latin America Transactions*, IEEE, v. 12, n. 8, p. 1464–1472, 2014. Citado na página 68.

ANEEL. *Fontes de Energia no Brasil*. 2016. Acessado em: 20-03-2019. Disponível em: <https://www.aneel.gov.br/documents/656877/15142444/Fontes+de+Energia+no+Brasil/2eb48f5c-cc7f-4f63-867e-b2a4f3603418?version=1.0v>. Citado na página 28.

ANEEL. *Nota técnica n111/2016: Instituição de Indicadores Públicos de Sustentabilidade Econômico-Financeira*. 2016. Acessado em: 20-03-2019. Disponível em: <https://www2.aneel.gov.br/aplicacoes/consulta_publica/documentos/Nota_Tecnica_2016_111.pdf>. Citado 2 vezes nas páginas 47 e 54.

ANEEL. *Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional - PRODIST*. 2016. Acessado em: 13-09-2019. Disponível em: <https://www.aneel.gov.br/prodist>. Citado 2 vezes nas páginas 71 e 72.

ANEEL. *Procedimentos de Regulação Tarifária - PRORET*. 2016. Acessado em: 13-09-2019. Disponível em: <https://www.aneel.gov.br/procedimentos-de-regulacao-tarifaria-proret>. Citado 2 vezes nas páginas 18 e 100.

ANEEL. *Quinto Termo Aditivo aos Contratos de Concessão de Serviço Público de Distribuição de Energia Elétrica N 002/1997-DNAEE, N 003/1997-DNAEE, N 004/1997-DNAEE e N 005/1997-DNAEE*. 2016. Acessado em: 20-03-2019. Disponível em: <https://www.aneel.gov.br/documents/10184//15063035//Quinto+Termo+Aditivo.pdf>. Citado na página 72.

ANEEL. *Sustentabilidade Econômico-Financeira*. 2016. Acessado em: 05-07-2019. Disponível em: <https://www.aneel.gov.br/informacoes-tecnicas/-/asset_publisher/CegkWaVJWF5E/content/sustentabilidade-economico-financeira/656815?inheritRedirect=false>. Citado na página 54.

ANEEL. *Sistema de Informações Geográficas do Setor Elétrico - SIGEL*. 2017. Disponível em: <https://sigel.aneel.gov.br/portal/home/index.html>. Citado na página 54.

ANEEL. *Agência Nacional de Energia Elétrica*. 2019. Acessado em: 20-03-2019. Disponível em: <https://www.aneel.gov.br/>. Citado na página 17.

Arend, R. J. The business model: Present and future—beyond a skeumorph. *Strategic Organization*, Sage Publications Sage UK: London, England, v. 11, n. 4, p. 390–402, 2013. Citado na página 32.

Aspara, J. et al. Corporate business model transformation and inter-organizational cognition: The case of nokia. *Long Range Planning*, Elsevier, v. 46, n. 6, p. 459–474, 2013. Citado na página 32.

Ateba, B. B.; Prinsloo, J. J. Strategic management for electricity supply sustainability in south africa. *Utilities Policy*, Elsevier, v. 56, p. 92–103, 2019. Citado na página 48.

Ateba, B. B.; Prinsloo, J. J.; Gawlik, R. The significance of electricity supply sustainability to industrial growth in south africa. *Energy Reports*, Elsevier, v. 5, p. 1324–1338, 2019. Citado na página 48.

Begić, F.; Afgan, N. H. Sustainability assessment tool for the decision making in selection of energy system—bosnian case. *Energy*, Elsevier, v. 32, n. 10, p. 1979–1985, 2007. Citado na página 47.

Berends, H. et al. Learning while (re) configuring: Business model innovation processes in established firms. *Strategic Organization*, Sage Publications Sage UK: London, England, v. 14, n. 3, p. 181–219, 2016. Citado na página 33.

Bilder, C. R.; Loughin, T. M. *Analysis of categorical data with R*. [S.l.]: CRC Press, 2014. Citado na página 52.

Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 99, n. suppl 3, p. 7280–7287, 2002. Citado na página 35.

Brasil. Lei n 9.427, de 26 de dezembro de 1996. *Institui a Agência Nacional de Energia Elétrica - ANEEL, disciplina o regime das concessões de serviços públicos de energia elétrica e dá outras providências.*, 1996. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L9427compilada.htm>. Citado na página 17.

Breiman, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 74 e 108.

Breiman, L. et al. Classification and regression trees. wadsworth int. *Group*, v. 37, n. 15, p. 237–251, 1984. Citado 3 vezes nas páginas 77, 89 e 107.

Brown, M. A.; Sovacool, B. K. Developing an'energy sustainability index'to evaluate energy policy. *Interdisciplinary Science Reviews*, Taylor & Francis, v. 32, n. 4, p. 335–349, 2007. Citado na página 48.

Burger, S. P.; Luke, M. Business models for distributed energy resources: A review and empirical analysis. *Energy Policy*, Elsevier, v. 109, p. 230–248, 2017. Citado na página 30.

Cao, Z. et al. Multi-factor analysis and modeling of net energy of lactation (nel) prediction in primiparous dairy cows. *Measurement*, Elsevier, v. 162, p. 107881, 2020. Citado na página 76.

Casadesus-Masanell, R.; Ricart, J. E. From strategy to business models and onto tactics. *Long range planning*, Elsevier, v. 43, n. 2-3, p. 195–215, 2010. Citado na página 31.

Casadesus-Masanell, R.; Ricart, J. E. How to design a winning business model. *Harvard business review*, v. 89, n. 1/2, p. 100–107, 2011. Citado na página 97.

Castro, N. de et al. Indicadores de sustentabilidade econômico-financeira das empresas de distribuição de energia elétrica. 2017. Citado 3 vezes nas páginas 35, 40 e 96.

CEMIG. *CEMIG*. 2012. Acessado em: 04-12-2019. Disponível em: <https://www.cemig.com.br/quem-somos/>. Citado 2 vezes nas páginas 21 e 29.

CEMIG. *História da eletricidade no Brasil*. 2012. Acessado em: 14-11-2019. Disponível em: <http://www.cemig.com.br/pt-br/a_cemig/Nossa_Historia/Paginas/historia_da_eletricidade_no_brasil.aspx>. Citado na página 21.

Cf, O. Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*, 2015. Citado 2 vezes nas páginas 47 e 50.

Chen, T. et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, p. 1–4, 2015. Citado 2 vezes nas páginas 77 e 109.

Choubin, B. et al. River suspended sediment modelling using the cart model: A comparative study of machine learning techniques. *Science of the Total Environment*, Elsevier, v. 615, p. 272–281, 2018. Citado na página 76.

Cosenz, F.; Noto, G. A dynamic business modelling approach to design and experiment new business venture strategies. *Long Range Planning*, Elsevier, v. 51, n. 1, p. 127–140, 2018. Citado na página 34.

Costa, J. et al. Adaptive learning for dynamic environments: A comparative approach. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 65, p. 336–345, 2017. Citado na página 76.

Costa, M. et al. Failure detection in robotic arms using statistical modeling, machine learning and hybrid gradient boosting. *Measurement*, Elsevier, v. 146, p. 425–436, 2019. Citado 6 vezes nas páginas 25, 74, 75, 77, 78 e 99.

Costa, M. A. et al. Bayesian detection of clusters in efficiency score maps: An application to brazilian energy regulation. *Applied Mathematical Modelling*, Elsevier, v. 68, p. 66–81, 2019. Citado 2 vezes nas páginas 109 e 110.

Costa, M. de A. *Tópicos em ciência dos dados: Introdução aos modelos paramétricos e seus aplicações utilizando o R*. [S.l.]: Bonecker, 2019. Citado na página 78.

Demil, B.; Lecocq, X. Business model evolution: in search of dynamic consistency. *Long range planning*, Elsevier, v. 43, n. 2-3, p. 227–246, 2010. Citado na página 97.

Doege, R.; Lakoski, J. C. Análise comparativa de rentabilidade e lucratividade dos negócios geração, transmissão e distribuição de energia elétrica. In: *Anais do Congresso Brasileiro de Custos-ABC*. [S.l.: s.n.], 2012. Citado na página 28.

Dong, X. et al. Multiscale feature extraction from the perspective of graph for hob fault diagnosis using spectral graph wavelet transform combined with improved random forest. *Measurement*, Elsevier, v. 176, p. 109178, 2021. Citado na página 77.

Doukas, H. et al. Assessing energy sustainability of rural communities using principal component analysis. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 16, n. 4, p. 1949–1957, 2012. Citado na página 52.

Energia, C. *Consulta pública discute revisão dos indicadores de qualidade*. 2019. Acessado em: 15-12-2019. Disponível em: <https://www.canalenergia.com.br/noticias/53120751/consulta-publica-discute-revisao-dos-indicadores-de-qualidade>. Citado na página 18.

Farquharson, D.; Jaramillo, P.; Samaras, C. Sustainability implications of electricity outages in sub-saharan africa. *Nature Sustainability*, Nature Publishing Group, v. 1, n. 10, p. 589–597, 2018. Citado na página 50.

FIRJAN, F. das Indústrias do Estado do Rio de J. *Diretor-geral da Aneel defende desoneração tarifária em Conselho da Firjan*. 2019. Acessado em: 23-10-2019. Disponível em: <https://www.firjan.com.br/noticias-1/diretor-geral-da-aneel-defende-desoneracao-tarifaria-em-conselho-da-firjan.htm>. Citado na página 18.

Fisher, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, JSTOR, v. 85, n. 1, p. 87–94, 1922. Citado na página 77.

Friedman, J. Stochastic gradient boosting. *Computational statistics & data analysis*, Elsevier, v. 38, n. 4, p. 367–378, 2002. Citado na página 77.

Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1. Citado 3 vezes nas páginas 77, 99 e 109.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado 3 vezes nas páginas 25, 78 e 99.

Fuks, M.; Salazar, E. Applying models for ordinal logistic regression to the analysis of household electricity consumption classes in rio de janeiro, brazil. *Energy Economics*, Elsevier, v. 30, n. 4, p. 1672–1692, 2008. Citado na página 52.

Ganhadeiro, T. G. L. et al. Evaluation of energy distribution using network data envelopment analysis and kohonen self organizing maps. *Energies*, Multidisciplinary Digital Publishing Institute, v. 11, n. 10, p. 2677, 2018. Citado na página 68.

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*, Elsevier, v. 64, p. 373–383, 2017. Citado na página 68.

Grace, J. B. *Structural equation modeling and natural systems*. [S.l.]: Cambridge University Press, 2006. Citado na página 73.

Harrel, F. *Regression Modeling strategies: general aspects of fitting regression models*. [S.l.]: New York, NY, Springer, 2001. Citado na página 52.

Hoerl, A. E.; Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis Group, v. 12, n. 1, p. 55–67, 1970. Citado 2 vezes nas páginas 76 e 107.

Hoptroff, R. G. The principles and practice of time series forecasting and business modelling using neural nets. *Neural Computing & Applications*, Springer, v. 1, n. 1, p. 59–66, 1993. Citado na página 37.

Horkoff, J. et al. Strategic business modeling: representation and reasoning. *Software & Systems Modeling*, Springer, v. 13, n. 3, p. 1015–1041, 2014. Citado na página 35.

Iddrisu, I.; Bhattacharyya, S. C. Sustainable energy development index: A multi-dimensional indicator for measuring sustainable energy development. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 50, p. 513–530, 2015. Citado 2 vezes nas páginas 47 e 51.

Jayawardena, S.; Epps, J.; Ambikairajah, E. Ordinal logistic regression with partial proportional odds for depression prediction. *IEEE Transactions on Affective Computing*, IEEE, 2020. Citado na página 52.

Jia, L.; Li, E.; Yu, J. Designing neurofuzzy system based on icart algorithm and its application for modeling jet fuel endpoint of hydrocracking process. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 16, p. 11–19, 2003. Citado na página 76.

Jiang, X.; Li, S. A dual path optimization ridge estimation method for condition monitoring of planetary gearbox under varying-speed operation. *Measurement*, Elsevier, v. 94, p. 630–644, 2016. Citado na página 76.

Johnson, M. W.; Christensen, C. M.; Kagermann, H. Business model. *Harvard Business Review*, v. 86, n. 12, p. 51–59, 2008. Citado na página 31.

Kalaiselvi, B.; Thangamani, M. An efficient pearson correlation based improved random forest classification for protein structure prediction techniques. *Measurement*, Elsevier, v. 162, p. 107885, 2020. Citado na página 77.

Kang, H. J.; Shin, J.-G.; Lee, J. K. A business model-based design of a damage control support system for naval ships. *Systems Engineering*, Wiley Online Library, v. 15, n. 1, p. 14–27, 2012. Citado na página 34.

Karger, C. R.; Hennings, W. Sustainability evaluation of decentralized electricity generation. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 13, n. 3, p. 583–593, 2009. Citado 2 vezes nas páginas 47 e 50.

Kauffman, R. J.; Wang, B. Tuning into the digital channel: evaluating business model characteristics for internet firm survival. *Information Technology and Management*, Springer, v. 9, n. 3, p. 215–232, 2008. Citado na página 33.

Kaygusuz, K. Energy for sustainable development: A case of developing countries. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 16, n. 2, p. 1116–1126, 2012. Citado na página 47.

Kemmler, A.; Spreng, D. Energy indicators for tracking sustainability in developing countries. *Energy policy*, Elsevier, v. 35, n. 4, p. 2466–2480, 2007. Citado 2 vezes nas páginas 47 e 50.

Keramat-Jahromi, M. et al. Real-time moisture ratio study of drying date fruit chips based on on-line image attributes using knn and random forest regression methods. *Measurement*, Elsevier, v. 172, p. 108899, 2021. Citado na página 77.

Knorr-Held, L.; Raßer, G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, Wiley Online Library, v. 56, n. 1, p. 13–21, 2000. Citado na página 110.

Kraus, M.; Feuerriegel, S.; Oztekin, A. Deep learning in business analytics and operations research: Models, applications and managerial implications. *arXiv preprint arXiv:1806.10897*, 2018. Citado na página 35.

Li, H. et al. Vehicle classification with single multi-functional magnetic sensor and optimal mns-based cart. *Measurement*, Elsevier, v. 55, p. 142–152, 2014. Citado na página 76.

Liu, Y.; Wei, J. Business modeling for entrepreneurial firms: four cases in china. *Chinese management studies*, Emerald Group Publishing Limited, 2013. Citado na página 35.

Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, IEEE, v. 7, p. 154096–154113, 2019. Citado na página 74.

Magretta, J. *Why business models matter*. [S.l.]: Harvard Business School Boston, MA, 2002. Citado 2 vezes nas páginas 29 e 30.

May, J. R.; Brennan, D. J. Sustainability assessment of australian electricity generation. *Process Safety and Environmental Protection*, Elsevier, v. 84, n. 2, p. 131–142, 2006. Citado na página 50.

Meakin, S. *The Rio Earth summit: summary of the United Nations conference on environment and development*. [S.l.]: Library of Parliament, Research Branch, 1992. v. 317. Citado na página 47.

Millan, J.; Lora, E.; Micco, A. Sustainability of the electricity sector reforms in latin america. *Research Department, Inter-American Development Bank*, Citeseer, 2001. Citado na página 50.

Moeinaddini, M. et al. Proposing a new score to measure personal happiness by identifying the contributing factors. *Measurement*, Elsevier, v. 151, p. 107115, 2020. Citado na página 72.

Morris, M.; Schindehutte, M.; Allen, J. The entrepreneur's business model: toward a unified perspective. *Journal of business research*, Elsevier, v. 58, n. 6, p. 726–735, 2005. Citado na página 36.

Murthy, S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, Springer, v. 2, n. 4, p. 345–389, 1998. Citado na página 76.

Nandagopal, V. et al. Feasible analysis of gene expression–a computational based classification for breast cancer. *Measurement*, Elsevier, v. 140, p. 120–125, 2019. Citado 2 vezes nas páginas 75 e 76.

Nelder, J.; Baker, R. *Generalized linear models. Encyclopedia of statistical sciences*. [S.l.]: Wiley, New York, 1972. Citado 3 vezes nas páginas 75, 77 e 99.

Neves, A. R.; Leal, V. Energy sustainability indicators for local energy planning: Review of current practices and derivation of a new framework. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 14, n. 9, p. 2723–2735, 2010. Citado 2 vezes nas páginas 47 e 50.

Ngouna, R. H. et al. A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 95, p. 103822, 2020. Citado na página 77.

Osterwalder, A.; Pigneur, Y. *Business model generation: a handbook for visionaries, game changers, and challengers*. [S.l.]: John Wiley & Sons, 2010. Citado 3 vezes nas páginas 32, 36 e 98.

Osterwalder, A.; Pigneur, Y. Designing business models and similar strategic objects: the contribution of is. *Journal of the Association for information systems*, v. 14, n. 5, p. 3, 2012. Citado na página 36.

Ovans, A. What is a business model. *Harvard business review*, v. 23, n. January, 2015. Citado na página 31.

Oyedepo, S. O. Energy and sustainable development in nigeria: the way forward. *Energy, Sustainability and Society*, Springer, v. 2, n. 1, p. 1–17, 2012. Citado na página 47.

Özbuğday, F. C.; Öğünlü, B.; Alma, H. The sustainability of turkish electricity distributors and last-resort electricity suppliers: What did transition from vertically integrated public monopoly to regulated competition with privatized and unbundled firms bring about? *Utilities Policy*, Elsevier, v. 39, p. 50–67, 2016. Citado na página 50.

Posner, R. A. *Natural monopoly and its regulation*. [S.l.]: Cato Institute, 1999. Citado 2 vezes nas páginas 17 e 71.

Powers, D.; Xie, Y. *Statistical methods for categorical data analysis*. [S.l.]: Emerald Group Publishing, 2008. Citado na página 52.

Prete, C. L. et al. Sustainability and reliability assessment of microgrids in a regional electricity market. *Energy*, Elsevier, v. 41, n. 1, p. 192–202, 2012. Citado 2 vezes nas páginas 47 e 50.

Pugesek, B. H.; Tomer, A.; Eye, A. V. *Structural equation modeling: applications in ecological and evolutionary biology*. [S.l.]: Cambridge University Press, 2003. Citado 3 vezes nas páginas 72, 73 e 98.

Queiroz, L. C.; Costa, M. A.; Lopes, A. L. M. Avaliação das eficiências operacionais das gerências cemig-d: Uma abordagem utilizando data envelopment analysis e regressão truncada em segundo estágio. In: In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 49., 2017, Blumenau .... [S.l.], 2017. Citado na página 72.

Reineking, B. et al. Constrain to perform: regularization of habitat models. *Ecological Modelling*, Elsevier, v. 193, n. 3-4, p. 675–690, 2006. Citado na página 76.

Rösch, C. et al. Indicator system for the sustainability assessment of the german energy system and its transition. *Energy, Sustainability and Society*, Springer, v. 7, n. 1, p. 1–13, 2017. Citado 2 vezes nas páginas 47 e 51.

Rösch, C. et al. Sustainability assessment of the german energy transition. *Energy, Sustainability and Society*, BioMed Central, v. 8, n. 1, p. 1–23, 2018. Citado na página 47.

Saccoccio, M. et al. Optimal regularization in distribution of relaxation times applied to electrochemical impedance spectroscopy: ridge and lasso regression methods-a theoretical and experimental study. *Electrochimica Acta*, Elsevier, v. 147, p. 470–482, 2014. Citado na página 75.

Sanders, M. P. et al. Energy policy by beauty contests: the legitimacy of interactive sustainability policies at regional levels of the regulatory state. *Energy, sustainability and society*, Springer, v. 4, n. 1, p. 1–13, 2014. Citado na página 47.

Sarangi, G. K. et al. Indian electricity sector, energy security and sustainability: An empirical assessment. *Energy Policy*, Elsevier, v. 135, p. 110964, 2019. Citado 2 vezes nas páginas 47 e 51.

Schumacker, R.; Lomax, R. A guide to structural equations modeling. *Hillsdale, NJ: Erl-baum*, 1996. Citado na página 72.

Sharma, T.; Balachandra, P. Benchmarking sustainability of indian electricity system: An indicator approach. *Applied Energy*, Elsevier, v. 142, p. 206–220, 2015. Citado 2 vezes nas páginas 47 e 51.

Silva, A. V. D. et al. Benchmarking modeling for cost incentive regulation of brazilian electricity companies. Universidade Federal de Minas Gerais, 2019. Citado na página 68.

Silva, A. V. da et al. Performance benchmarking models for electricity transmission regulation: Caveats concerning the brazilian case. *Utilities Policy*, Elsevier, v. 60, p. 100960, 2019. Citado na página 68.

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*, Elsevier, v. 65, p. 111–126, 2019. Citado na página 68.

Sioshansi, F.; Pfaffenberger, W. *Electricity market reform: an international perspective*. [S.l.]: Elsevier, 2006. Citado na página 71.

Snedecor, G. W.; Cochran, W. G. Statistical methods, eight edition. *Iowa state University press, Ames, Iowa*, 1989. Citado na página 77.

Streimikiene, D.; Siksnelyte, I. Sustainability assessment of electricity market models in selected developed world countries. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 57, p. 72–82, 2016. Citado 2 vezes nas páginas 47 e 51.

Suganthi, L. Sustainability indices for energy utilization using a multi-criteria decision model. *Energy, Sustainability and Society*, BioMed Central, v. 10, n. 1, p. 1–31, 2020. Citado na página 48.

Swan, L. G.; Ugursal, V. I. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, Elsevier, v. 13, n. 8, p. 1819–1835, 2009. Citado na página 35.

Tavassoli, M.; Ketabi, S.; Ghandehari, M. Developing a network dea model for sustainability analysis of iran's electricity distribution network. *International Journal of Electrical Power & Energy Systems*, Elsevier, v. 122, p. 106187, 2020. Citado na página 52.

Team, R. C. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Citado na página 55.

Teece, D. J. Business models, business strategy and innovation. *Long range planning*, Elsevier, v. 43, n. 2-3, p. 172–194, 2010. Citado na página 36.

Teixeira, L.; Lopes, H. E. G. Aplicação do modelo canvas para o modelo de negócios do banco do brasil e da caixa econômica federal. *Revista Gestão & Tecnologia*, v. 16, n. 2, p. 73–99, 2016. Citado 2 vezes nas páginas 34 e 36.

Thrän, D. et al. Governance of sustainability in the german biogas sector—adaptive management of the renewable energy act between agriculture and the energy sector. *Energy, Sustainability and Society*, Springer, v. 10, n. 1, p. 1–18, 2020. Citado na página 48.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citado na página 75.

Uimonen, S. et al. A machine learning approach to modelling escalator demand response. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 90, p. 103521, 2020. Citado na página 77.

Velu, C. Business model innovation and third-party alliance on the survival of new firms. *Technovation*, Elsevier, v. 35, p. 1–11, 2015. Citado na página 33.

Vera, I.; Langlois, L. Energy indicators for sustainable development. *Energy*, Elsevier, v. 32, n. 6, p. 875–882, 2007. Citado na página 47.

Vithayasrichareon, P.; MacGill, I. F.; Nakawiro, T. Assessing the sustainability challenges for electricity industries in asean newly industrialising countries. *Renewable and Sustainable Energy Reviews*, Elsevier, v. 16, n. 4, p. 2217–2233, 2012. Citado 2 vezes nas páginas 47 e 51.

Voronov, S.; Jung, D.; Frisk, E. A forest-based algorithm for selecting informative variables using variable depth distribution. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 97, p. 104073, 2021. Citado na página 77.

Wang, T.; Lin, Q. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *Journal of Machine Learning Research*, v. 22, n. 137, p. 1–38, 2021. Citado na página 74.

Wirtz, B. et al. Business models: Origin, development and future research perspectives. *Long range planning*, Elsevier, v. 49, n. 1, p. 36–54, 2016. Citado 2 vezes nas páginas 31 e 97.

Xu, K. et al. Application of ordinal logistic regression analysis to identify the determinants of illness severity of covid-19 in china. *Epidemiology & Infection*, Cambridge University Press, v. 148, 2020. Citado na página 52.

Zhang, C.; Ma, Y. *Ensemble machine learning: methods and applications*. [S.l.]: Springer, 2012. Citado na página 77.

Zhang, S. et al. A temporal lasso regression model for the emergency forecasting of the suspended sediment concentrations in coastal oceans: Accuracy and interpretability. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 100, p. 104206, 2021. Citado na página 76.

Zhou, Q. et al. Estimation of the instantaneous rotational frequency of gear transmission with large speed variations using short-time angular resampling and ridge-enhancing techniques. *Measurement*, Elsevier, v. 172, p. 108844, 2021. Citado na página 76.

Zott, C.; Amit, R.; Massa, L. The business model: recent developments and future research. *Journal of management*, SAGE Publications Sage CA: Los Angeles, CA, v. 37, n. 4, p. 1019–1042, 2011. Citado na página 36.

Apêndices

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

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

Marcelo Azevedo Costa[a,*], Leandro Brioschi Mineti[b], Álvaro Lédo Ferreira[a]

[a]*Department of Industrial Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil*
macosta@ufmg.br, alvaroledoferreira@gmail.com
[b]*Falconi Consultants for Results, São Paulo, SP 04543-011, Brazil*
leandromineti@falconi.com

Abstract

The quality of electrical energy distribution services is undoubtedly negatively correlated with power outages. Nevertheless, power outage is affected by both managerial and non-managerial factors. For instance, environmental factors are known drivers of power outage, worldwide. To tackle the complex behavior of power outage, this work proposes a Bayesian spatial regression model to estimate spatial clusters and the regression coefficients within each cluster. The number of spatial clusters, their locations and sizes are unknown and are estimated using prior probability distributions. The regression coefficients are estimated based on managerial and environmental factors. The case study and the main motivation is the prediction of the power outage indicator in the largest Brazilian distribution service operator, located in the southeast region. The estimated model will drive future management decisions to reduce compensation paid to consumers and fines charged by the regulator, consequently increasing investments in network expansion and quality of the services.

Keywords:
hierarchical Bayesian modeling, reversible-jump mcmc, spatial clustering.

*Corresponding author
Email address: macosta@ufmg.br (Marcelo Azevedo Costa)

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, 2018). 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 and 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 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 and the Brazilian database is also presented in section 2. The Brazilian data set analysis using the proposed methodology are presented in section 3. Discussion is presented in section 4 and conclusion is presented in section 5. Additional information regarding the proposed algorithm, simulation study and univariate regression results are found in the appendix.

2. Material and methods

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 et al. (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 and 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 and 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 and 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 et al. (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 and 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 and 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 et al. (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 and 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 and 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 et al. (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 and 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.

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, 2016a). 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, 2016b).

The DEC indicator is calculated using Equation 1,

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

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.

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:

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

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 2 can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3}$$

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, $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I})$ (Seber and 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 and Lee, 2012).

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:

$$P(\boldsymbol{\beta}, \sigma^2) \propto P(\boldsymbol{\beta}|\sigma^2)P(\sigma^2). \tag{4}$$

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

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_{k+1}(\boldsymbol{\mu}_0; \sigma^2 \boldsymbol{\Lambda}_\mathbf{0}^{-1}).$$

In a weakly informative *prior* specification for $\boldsymbol{\beta}$, we set the mean and covariance matrix as $\boldsymbol{\mu}_\mathbf{0} = \mathbf{0}$, $\boldsymbol{\Lambda}_\mathbf{0} = \lambda_0 \mathbf{I}$. 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 et al. (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,

$$P(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \propto P(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \times P(\boldsymbol{\beta}, \sigma^2). \tag{5}$$

The application of Equation 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

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

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:

$$L(\mathbf{y} \mid \mathbf{X}, \mathbf{B}_k, \sigma^2) = \prod_{j=1}^{k}\prod_{i \in C_j}\frac{1}{\sigma}\phi\left(\frac{y_i - \mu_i}{\sigma}\right) \tag{6}$$

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.

2.5.1. Prior *distribution for the number of clusters*

As suggested by Knorr-Held and 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.

$$P(k) \propto (1 - c)^k. \tag{7}$$

2.5.2. Prior distribution for the coefficients and variance

As defined in section 2.4, a weakly informative Normal prior distribution is assumed for the vectors $\boldsymbol{\beta}_j$:

$$\boldsymbol{\beta}_j | \sigma^2 \sim \mathcal{N}_{p+1}(\mathbf{0}; \sigma^2(\lambda_0 \mathbf{I})^{-1}). \tag{8}$$

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.

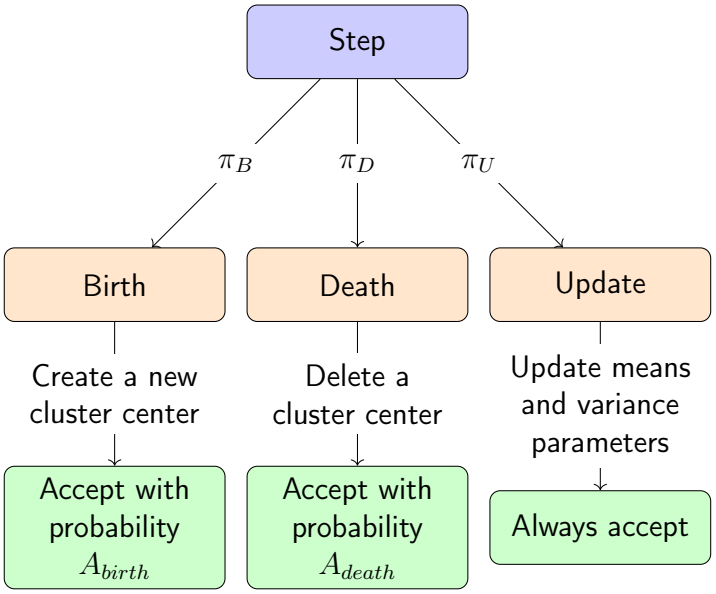


Figure 1: RJMCMC algorithm flowchart.

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. (2019). However, the number of possible steps in the Reversible Jump Markov Chain Monte Carlo was reduced from five to three. Costa et al. (2019) claims that the *shift* and *switch* steps can be removed without affecting the performance of the algorithm. Thus, only *birth*, *death* and *update* steps are implemented. Figure 1 presents a diagram of the proposed algorithm.

Similar to 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 and Costa, 2018). Further details about RJMCMC sampler are presented in the appendix.

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 et al., 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 et al., 2001) and the predictive coefficient of determination ($R^2_{prediction}$) (Montgomery et al., 2012) is proposed to estimate the optimal number of clusters.

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 (reduction in dimensionality) was applied. Initially, a statistical factor analysis (Johnson et al., 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 1.

Table 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. Results

3.1. 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 2 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 $R^2 = 0.5702$.

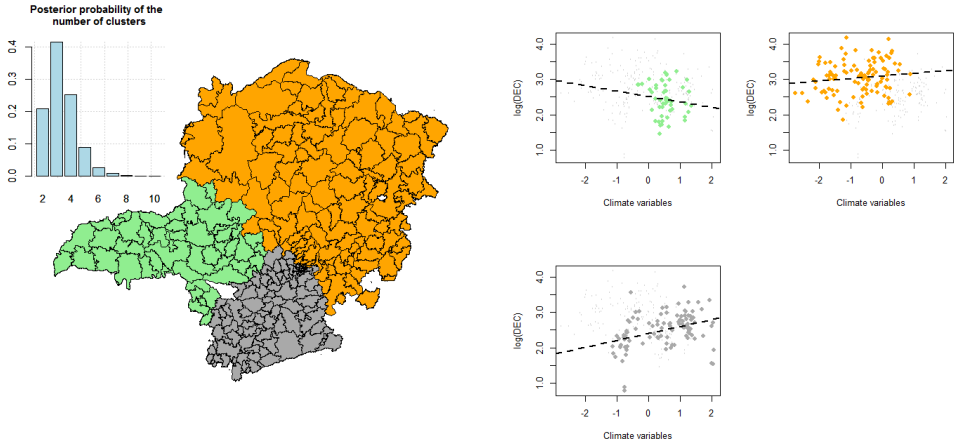
Table 2: Multiple linear regression results using the estimated latent variables and the logarithm of the power outage indicator.

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

As opposed to using the proposed Bayesian spatial regression model including all seven variables, univariate models were primarily used to investigate, for each variable, 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 2 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 2); i.e., the larger the climate variable, the larger the power outage. The estimated coefficient for the cluster located in the west is 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.

Figure 2: Results using the univariate Bayesian spatial regression model and the climate variable as the predictor.

Similar results were found for the remaining variables. Detailed analyses are found in Appendix C. 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. Table 3 summarizes the detected clusters using the proposed univariate Bayesian spatial regression model. In general, three to four spatial clusters were detected. The smaller clusters detected using demand for electrical services I, demand for electrical services II and electrical assets I variables are located closer to the state capital and surrounding areas. 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

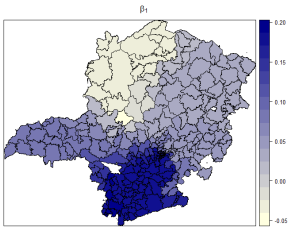
coefficient for the climate variable, as previously presented in Table 2. By dividing the data into spatial clusters, the estimated coefficients within some clusters are positive, as technically expected. These results indicate a concept known as the Simpson's paradox, or reversal paradox (Simpson, 1951), 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.

Table 3: Mode of the estimated number of clusters and number of areas in each cluster sorted in increasing order.

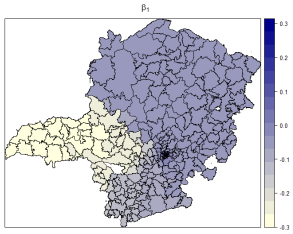
Evaluated variables	Mode of the number of clusters	Estimated number of areas in each cluster			
		cluster 1	cluster 2	cluster 3	cluster 4
Geographical assets	3	70	86	110	-
Electrical assets I	4	24	53	75	114
Electrical assets II	4	19	44	53	140
Climate variable	3	52	105	109	-
Demand for electrical services I	4	25	55	74	112
Demand for electrical services II	4	27	32	55	152
Operational and capital costs	4	36	39	51	140

Figure 3 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.

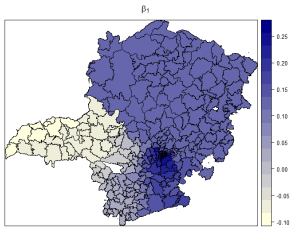
Figure 5 illustrates the Bayesian univariate spatial regression results using a simulated scenario with no spatial clusters, i.e., only one cluster. In this case, the proposed model correctly identified a single partition on the map and that the spatial estimates of the regression coefficient (β_1) are homogeneous.



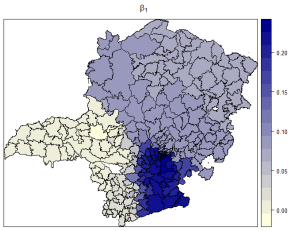
(a) Climate variable



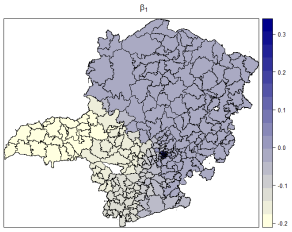
(b) Demand for electrical services I



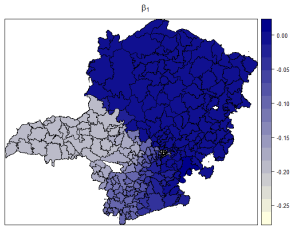
(c) Demand for electrical services II



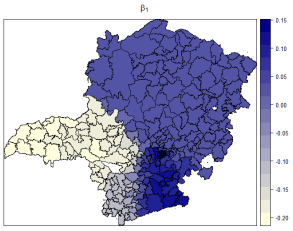
(d) Geographical assets



(e) Electrical assets I

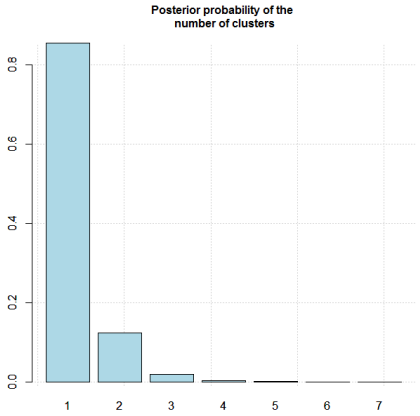


(f) Electrical assets II

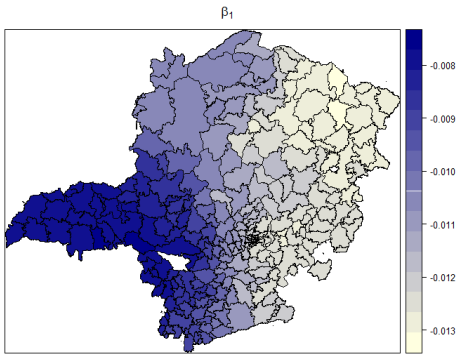


(g) Operational and capital costs

Figure 3: 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.

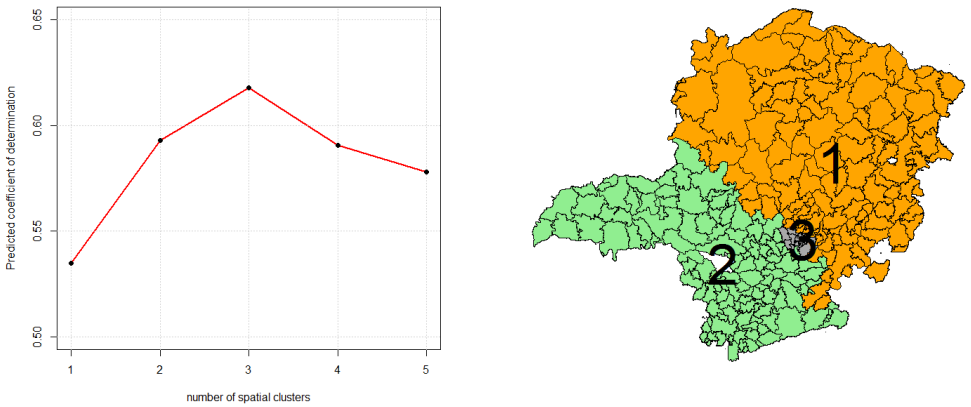
Figure 4: Results using the univariate Bayesian spatial regression model and a simulated scenario with no clusters.

Table 4 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 whereas 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 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.

Table 4: 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%

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 et al., 2012) was calculated and the cluster configuration achieving the maximum value of R^2_{pred} was selected. Figure 5(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 5(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.



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

(b) Electrical areas divided into 3 spatial clusters.

Figure 5: Final selection of the number of clusters based on the predictive coefficient of determination for different number of clusters.

Table 5 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 6 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 value for electrical assets II.

Table 7 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 ex-

Table 5: Estimated univariate and multiple liner regression coefficients for the spatial cluster 1.

Predictor variable	Expected sign	Cluster 1			
		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					

Table 6: Estimated univariate and multiple liner regression coefficients for the spatial cluster 2.

Predictor variable	Expected sign	Cluster 2			
		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					

pected. 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.

Results shown in Tables 5, 6 and 7 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.

4. Discussion

As previously mentioned, Brazil has continental dimensions and some of the Brazilian DSOs has concession areas larger than many european coun-

Table 7: Estimated univariate and multiple liner regression coefficients for the spatial cluster 3.

Predictor variable	Expected sign	Cluster 3			
		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					

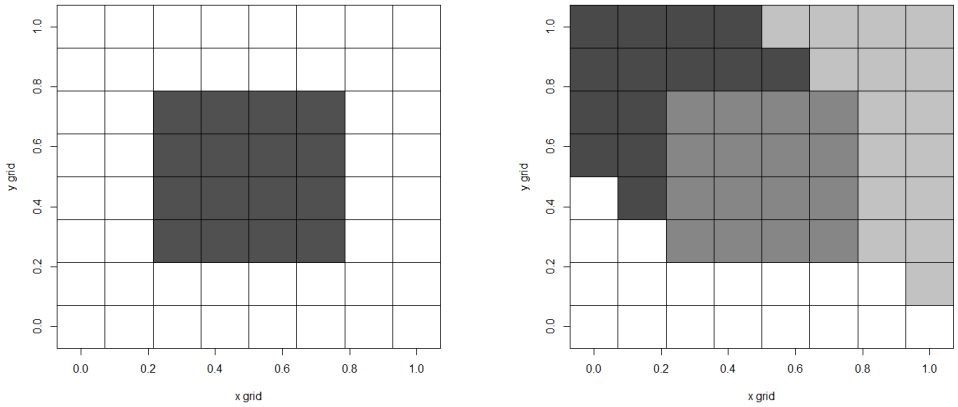
tries. Consequently, the electrical distribution service faces many challenges related to weather, vegetation and socioeconomic factors. In the case study, the geographical heterogeneity was technically known for engineers and management staff. Nonetheless, providing a proper treatment was a difficult 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 6 illustrates one limitation of the proposed clustering-based Bayesian spatial model. The original spatial partition algorithm, as proposed by Knorr-Held and Raßer (2000), may overestimate the number of clusters if their locations do not fit the spatial partition algorithm. For instance, Figure 6(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 6(b). Conse-

quently, 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.

Figure 6: Overestimation of the number of clusters using the spatial Bayesian approach. In order to detect the central cluster, two additional clusters are created.

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.

References

Andersen, T. B., Dalgaard, C.-J., 2013. Power outages and economic growth in africa. Energy Economics 38, 19–23.

ANEEL, 1996. Agência Nacional de Energia Elétrica. `aneel.gov.br`, [Online; accessed 08-September-2020].

ANEEL, 2016a. Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST. https://www.aneel.gov.br/prodist, [Online; accessed 13-September-2019].

ANEEL, 2016b. Qualidade do Serviço. https://www.aneel.gov.br/qualidade-do-servico2, [Online; accessed 20-March-2019].

ANEEL, 2018. Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST Módulo 8 – Qualidade da Energia Elétrica. [Online; accessed 08-September-2020].
URL http://www.aneel.gov.br/procedimentos-de-regulacao-tarifaria-proret

ANEEL, 2019. Painel de Desempenho das Distribuidoras de Energia Elétrica. www.aneel.gov.br/painel-de-desempenho, [Online; accessed 08-September-2020].

Baarsma, B. E., Hop, J. P., 2009. Pricing power outages in the netherlands. Energy 34 (9), 1378–1386.

Beenstock, M., Goldin, E., Haitovsky, Y., 1997. The cost of power outages in the business and public sectors in israel: revealed preference vs. subjective valuation. The Energy Journal 18 (2).

Biswas, S., Goehring, L., 2019. Load dependence of power outage statistics. EPL (Europhysics Letters) 126 (4), 44002.

Carlsson, F., Martinsson, P., 2007. Willingness to pay among swedish households to avoid power outages: a random parameter tobit model approach. The Energy Journal 28 (1).

Carlsson, F., Martinsson, P., Akay, A., 2011. The effect of power outages and cheap talk on willingness to pay to reduce outages. Energy Economics 33 (5), 790–798.

Castillo, A., 2014. Risk analysis and management in power outage and restoration: A literature survey. Electric Power Systems Research 107, 9–15.

Cole, M. A., Elliott, R. J., Occhiali, G., Strobl, E., 2018. Power outages and firm performance in sub-saharan africa. Journal of Development Economics 134, 150–159.

Costa, M. A., Mineti, L. B., Mayrink, V. D., Lopes, A. L. M., 2019. Bayesian detection of clusters in efficiency score maps: An application to brazilian energy regulation. Applied Mathematical Modelling 68, 66–81.

Cressie, N., 2015. Statistics for spatial data. John Wiley & Sons.

Feng, W., Lim, C. Y., Maiti, T., Zhang, Z., 2016. Spatial regression and estimation of disease risks: A clustering-based approach. Statistical Analysis and Data Mining: The ASA Data Science Journal 9 (6), 417–434.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. Vol. 1. Springer series in statistics New York.

Fujita, G., Shirai, G., 1997. Estimation of power outage size based on the dominating differential equation. Electrical engineering in Japan 118 (3), 39–49.

Gelman, A., et al., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian analysis 1 (3), 515–534.

Gil, G. D. R., Costa, M. A., Lopes, A. L. M., Mayrink, V. D., 2017. Spatial statistical methods applied to the 2015 brazilian energy distribution benchmarking model: Accounting for unobserved determinants of inefficiencies. Energy Economics 64, 373–383.

Green, P. J., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82 (4), 711–732.

Guha, S., Moss, A., Naor, J., Schieber, B., 1999. Efficient recovery from power outage. In: Proceedings of the thirty-first annual ACM symposium on Theory of computing. pp. 574–582.

Johnson, R. A., Wichern, D. W., et al., 2002. Applied multivariate statistical analysis. Vol. 5. Prentice hall Upper Saddle River, NJ.

Knorr-Held, L., Raßer, G., 2000. Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56 (1), 13–21.

Mineti, L., Costa, M., Jun. 2018. leandromineti/gbdcd: an R package implementing the Bayesian Detection of Clusters and Discontinuities.
URL `https://doi.org/10.5281/zenodo.1291501`

Moeltner, K., Layton, D. F., 2002. A censored random coefficients model for pooled survey data with application to the estimation of power outage costs. Review of Economics and Statistics 84 (3), 552–561.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2012. Introduction to linear regression analysis. Vol. 821. John Wiley & Sons.

Morrissey, K., Plater, A., Dean, M., 2018. The cost of electric power outages in the residential sector: A willingness to pay approach. Applied energy 212, 141–150.

Mukherjee, S., Nateghi, R., Hastak, M., 2018. A multi-hazard approach to assess severe weather-induced major power outage risks in the US. Reliability Engineering & System Safety 175, 283–305.

Ng, A. Y., Jordan, M. I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems. pp. 849–856.

Reilly, A., Guikema, S., 2015. Bayesian multiscale modeling of spatial infrastructure performance predictions with an application to electric power outage forecasting. Journal of infrastructure systems 21 (2), 04014036.

Seber, G. A., Lee, A. J., 2012. Linear regression analysis. Vol. 329. John Wiley & Sons.

Simpson, E. H., 1951. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological) 13 (2), 238–241.

Taimoor, N., Khosa, I., Jawad, M., Akhtar, J., Ghous, I., Qureshi, M. B., Ansari, A. R., Nawaz, R., 2020. Power outage estimation: The study of revenue-led top affected states of US. IEEE Access.

Zachariadis, T., Poullikkas, A., 2012. The costs of power outages: A case study from cyprus. Energy Policy 51, 630–641.

Appendix A. Details of the RJMCMC sampler algorithm

The proposed RJMCMC sampler comprises three main steps (*Birth*, *Death* or *Update*) described below.

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:

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

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:

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

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 $k+1$.

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:

$$A_{death} = \frac{L(\mathbf{y} \mid \mathbf{B}_{k-1}, G_{k-1}, \sigma^2)}{L(\mathbf{y} \mid \mathbf{B}_k, G_k, \sigma^2)} \cdot \frac{1}{(1-c)} \cdot \frac{\varphi(\boldsymbol{\beta}_r)}{\varphi_0(\boldsymbol{\beta}_r)}, \tag{A.3}$$

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 $k-1$.

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 A.1. 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:

$$\begin{aligned} a_n &= a_0 + \tfrac{n}{2} \\ b_n &= b_0 + \tfrac{\sum_{i=1}^{n}(y_i - \mu_i)^2}{2} \end{aligned} \tag{A.4}$$

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*.

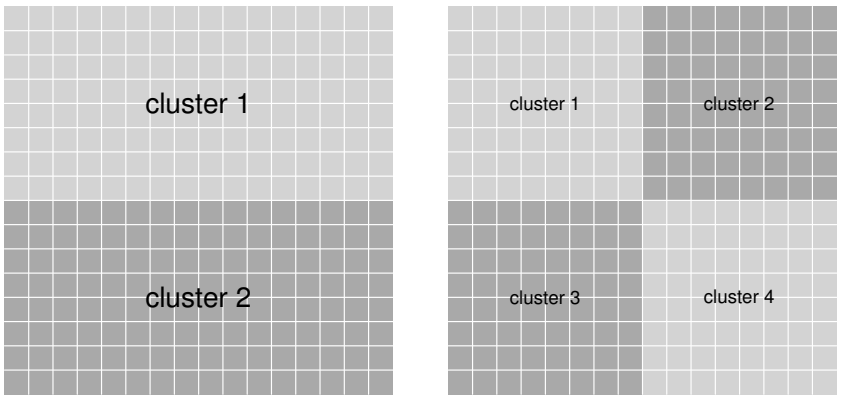
The complete algorithm is presented next.

Algorithm 1 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; π_B, π_D and π_{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. (A.1))
  for s=1 **to** Steps **do**
    Randomly choose one of the three moves with probabilities π_B, π_D, π_{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. (A.1).
      $\mathbf{B}^*_{k+1} \leftarrow$ 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. (A.2))
      $\alpha \leftarrow \min(1, A_{Birth})$
      $u \leftarrow$ 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. (A.3))
      $\alpha \leftarrow \min(1, A_{Death})$
      $u \leftarrow$ 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. (A.1)).
      $\sigma^2 \leftarrow$ Update the variance parameter using an inverse-Gamma random number, $\sigma^2 \sim \text{IG}(a_n, b_n)$ (Eq. (A.4))
    end if
  end for

Appendix B. 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 B.1. Thus the sample size is $n = 256$. Three different scenarios with one, two and four clusters were simulated. Figures B.1(a) and B.1(b) show simulated scenarios with two and four clusters, respectively.



(a) Simulated scenario with 2 clusters.

(b) Simulated scenario with 4 clusters.

Figure B.1: 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

$$\hat{\beta}^{[A]} - \hat{\beta}^{[B]} \sim Normal\left(\beta^{[A]} - \beta^{[B]}; 2k^2\sigma^2\right) \tag{B.1}$$

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 B.2.

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

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 B.1(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.

Results

Table B.1 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.

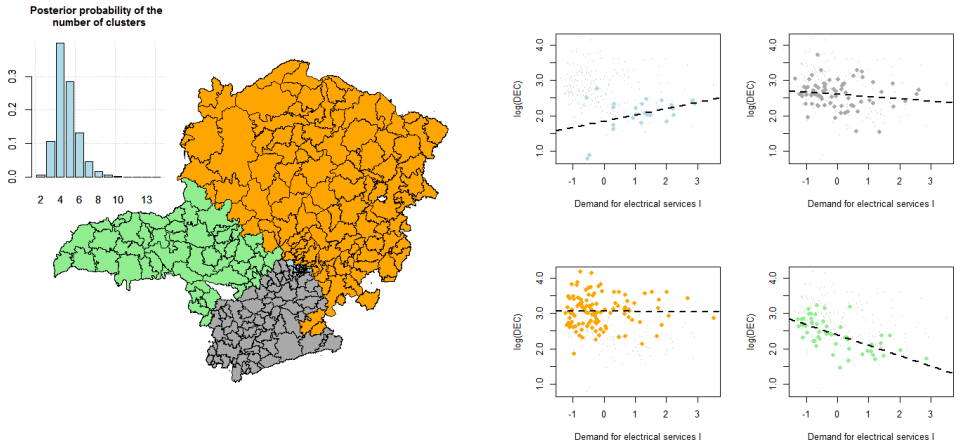
Appendix C. Univariate Bayesian regression results

Using the demand for electrical services I, Figure C.2 shows that four clusters were estimated. The largest cluster comprises the north region. The

Table B.1: 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 clusters	Prior c parameter	Z_α	Average R^2_{simul}	$\bar{k}\|\mathbf{Y},\mathbf{X}$	Average HPD size	HPD proportion
	0.01	NA	21.9%	1.26	6.48	98.0%
1	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%
	0.01	1.96	1.9%	4.5	16.2	97.5%
2	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%
	0.01	1.96	3.4%	3.9	16.3	94.0%
4	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%

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 2).

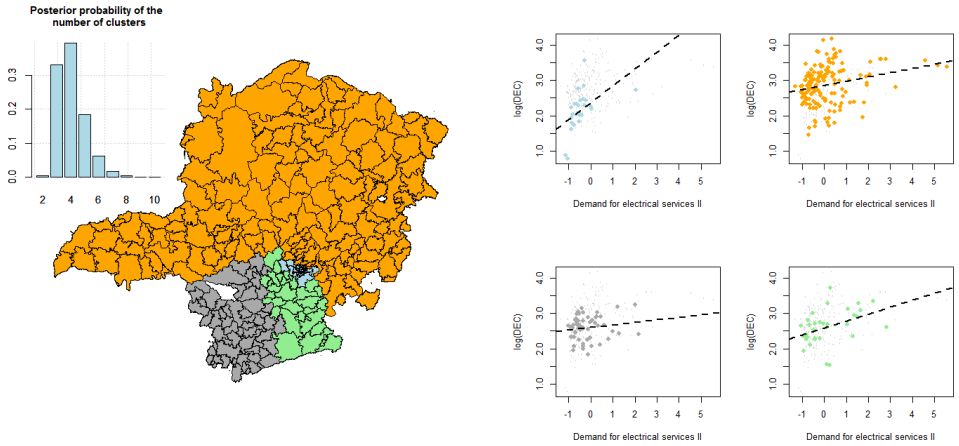


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figure C.2: 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 C.3 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.

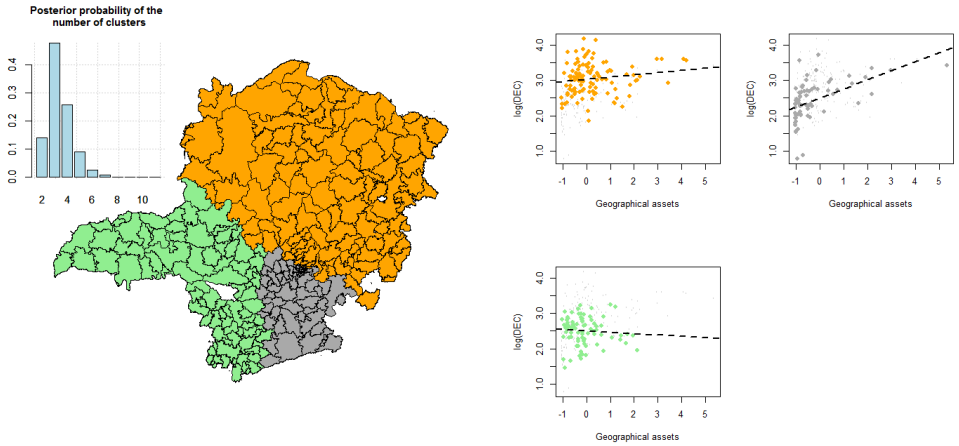


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figure C.3: Results using the univariate Bayesian spatial regression model and the demand for electrical services II as the predictor.

Using the geographical assets, Figure C.4 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 2); 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.

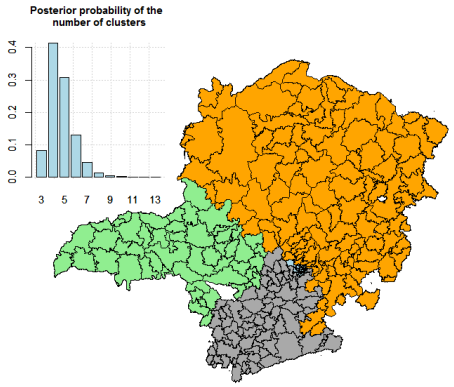


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

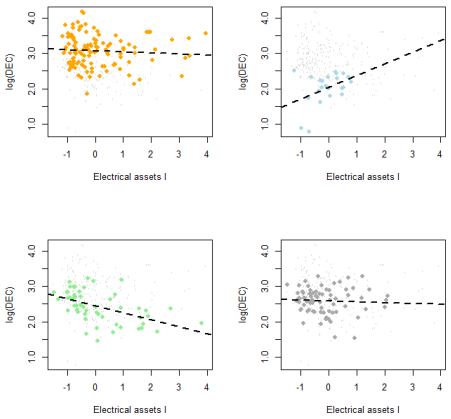
(b) Univariate regression models estimated for each cluster.

Figure C.4: Results using the univariate Bayesian spatial regression model and the geographical assets as the predictor.

Using the electrical assets I, Figure C.5 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 2).



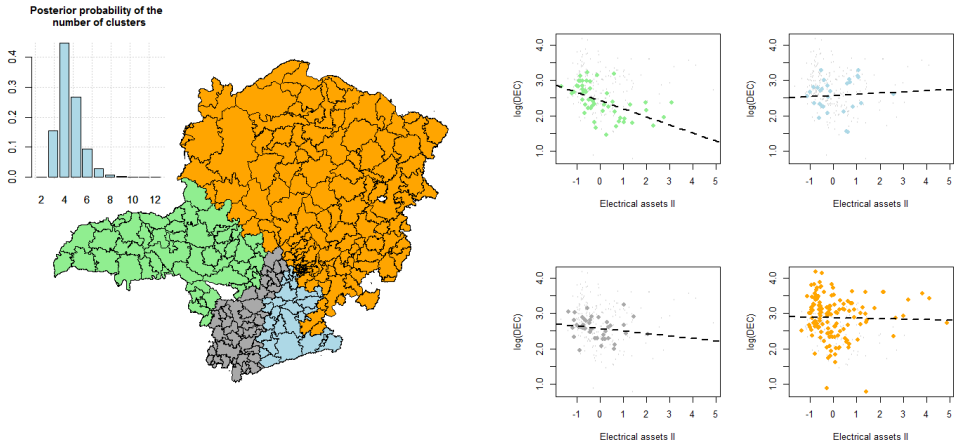
(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.



(b) Univariate regression models estimated for each cluster.

Figure C.5: Results using the univariate Bayesian spatial regression model and the electrical assets I as the predictor.

Using the demand for electrical services II, Figure C.6 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.

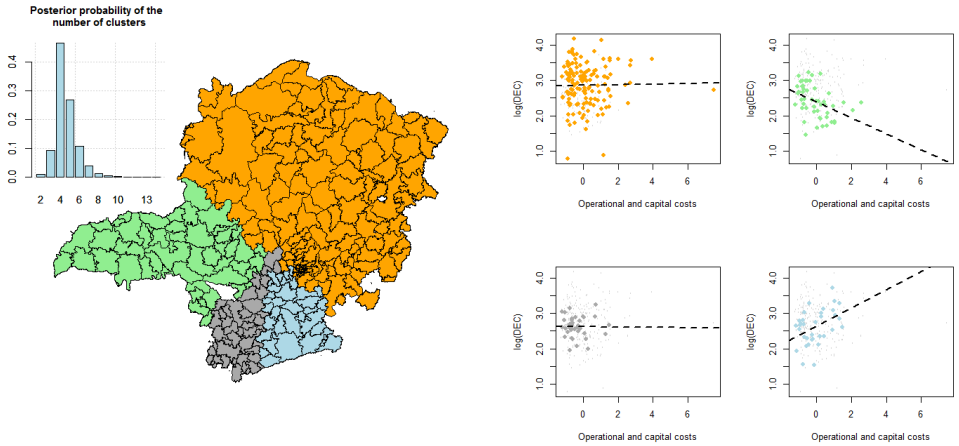


(a) The *a posteriori* distribution of the number of clusters and the most likely partition using the mode as the point estimate.

(b) Univariate regression models estimated for each cluster.

Figure C.6: Results using the univariate Bayesian spatial regression model and the electrical assets II as the predictor.

Using the operational and capital costs, Figure C.7 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.

Figure C.7: Results using the univariate Bayesian spatial regression model and the operational and capital costs as the predictor.

APÊNDICE B – PeD636 - Manual de uso do aplicativo

PeD636 Manual de Uso do Aplicativo

Versão 2.00

Luiz Henrique Cardoso Oliveira
Tomás Cadar de Castro
Álvaro Lédo Ferreira
Marcelo Azevedo Costa

Abril/2021

Sumário

INSTALAÇÃO DO PACOTE PeD363

No escopo do projeto P&D 0636, a ferramenta computacional proposta será disponibilizada como um pacote para o ambiente R. A instalação do pacote é possível a partir de um arquivo compactado (.tar.gz) chamado "**ped636_2.00.tar.gz**". De posse deste arquivo, os detalhes para a instalação do pacote "**ped636**" são mostrados a seguir.

1. Antes de iniciar a instalação do pacote "**ped636**" o usuário deve baixar e instalar o programa Rtools (necessário para a execução dos códigos em C++ pelo pacote). O programa pode ser baixado no link: https://cran.r-project.org/bin/windows/Rtools/. Após baixar e instalar o programa o usuário deve executar o seguinte código dentro do RStudio:

```
> writeLines('PATH="${RTOOLS40_HOME}\\usr\\bin;${PATH}"', con
= "~/.Renviron")
```

2. Para confirmar que a instalação do Rtools foi bem-sucedida o usuário deve reiniciar o seu RStudio e executar o código a seguir:

```
> Sys.which("make")
```

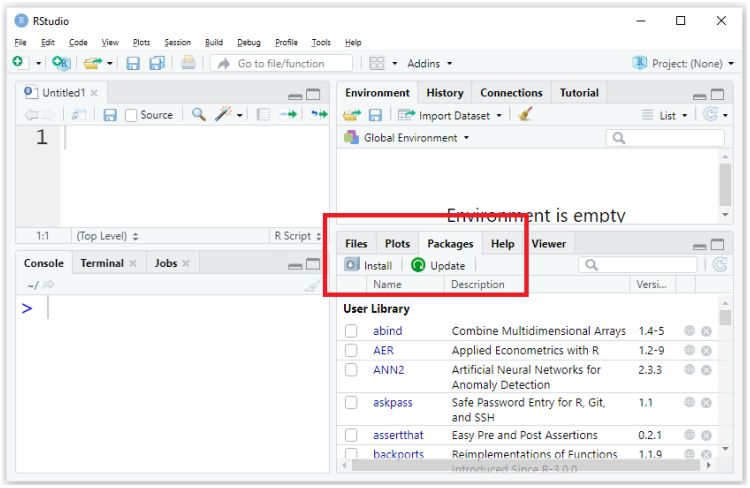
3. Caso o console retorne a mensagem abaixo, a instalação foi bem-sucedida. Caso contrário, volte ao link da instalação do Rtools para correção de erros.

```
> "C:\\rtools40\\usr\\bin\\make.exe"
```

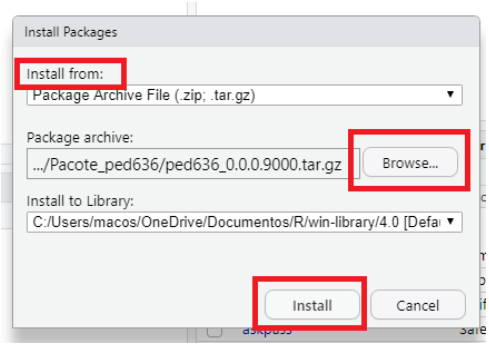
4. Após instalar o programa Rtools, o usuário deve instalar manualmente do CRAN todos os pacotes dos quais o pacote "**ped636**" é dependente. Para instalar os pacotes necessários deve executar o código a seguir:

```
> install.packages(c("dashboardthemes", "classInt", "Rcpp",
"RcppArmadillo", "dplyr", "DT", "exploreR", "forecast",
"geoR", "ggplot2", "gstat", "highcharter", "htmlwidgets",
"inline", "kableExtra", "leaflet", "lubridate", "maptools",
"MASS", "modeest", "mvtnorm", "openxlsx", "packHV",
"plotly", "plyr", "purrr", "randomForest", "RColorBrewer",
"readr", "readxl", "rgdal", "scales", "shiny", "shinyalert",
"shinyBS", "shinycssloaders", "shinydashboard", "shinyjs",
"shinyWidgets", "sp", "spData", "spdep", "StanHeaders",
"stringr"))
```

5. Para instalar o pacote "**ped636**" utilizando a interface gráfica do R Studio, acesse a opção "**Packages**" e selecione a opção "**Install**", como indicado na figura a seguir.



6. Em seguida, na opção "**Install from:**" selecione a opção "**Package Archive File (.zip; .tar.gz)**" e utilizando o botão "**Browse**", selecione o arquivo "**ped636_2.00.tar.gz**" no seu computador. Clique em "**Install**" para instalar o pacote.



7. Após a instalação do pacote, digite no console a opção:

```
> require(gstat);require(spdep);require(ped636)
```

MÓDULO DE INTERPOLAÇÃO ESPACIAL DOS DADOS DE MUNICÍPIO

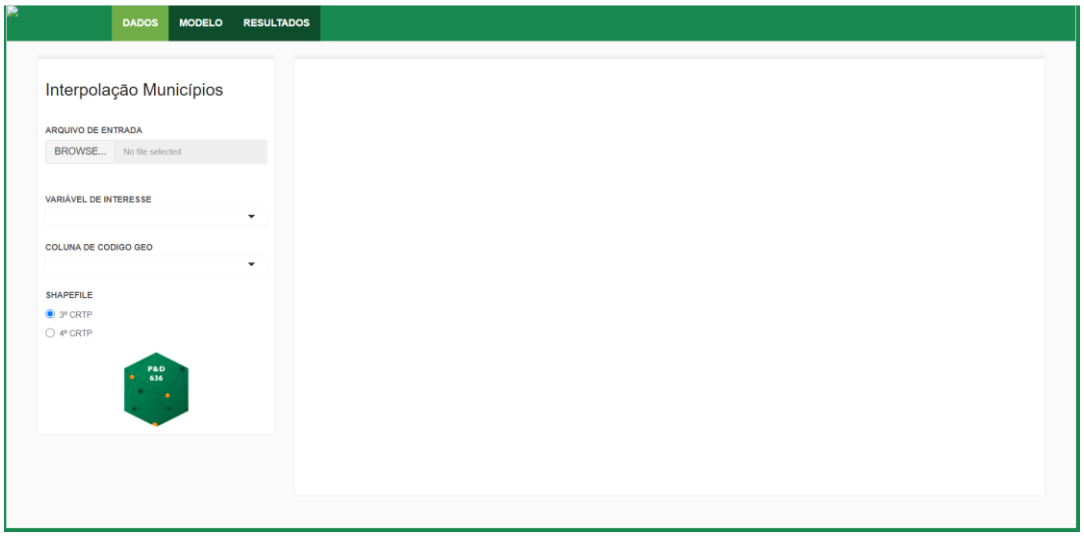
OBJETIVO DO APLICATIVO: A interface computacional de interpolação de dados municipais foi desenvolvida com o objetivo de agregar ou projetar as variáveis que estão segmentadas pelos municípios de MG para a configuração dos conjuntos elétricos da CEMIG-D. Por se tratar de uma ferramenta computacional para o tratamento de dados, sugere-se que a mesma seja utilizada inicialmente para a adição de novas variáveis ao banco de dados.

INSTRUÇÕES DE USO:

1. Para acessar o módulo de interpolação de dados municipais, será necessário chamar a função "**interp_CEMIG**", como descrito abaixo:

   ```
   > interp_CEMIG()
   ```

2. Após a execução da função, o aplicativo Shiny com o módulo de interpolação espacial dos dados municipais será aberto automaticamente na aba "**DADOS**", como apresentado na imagem abaixo:



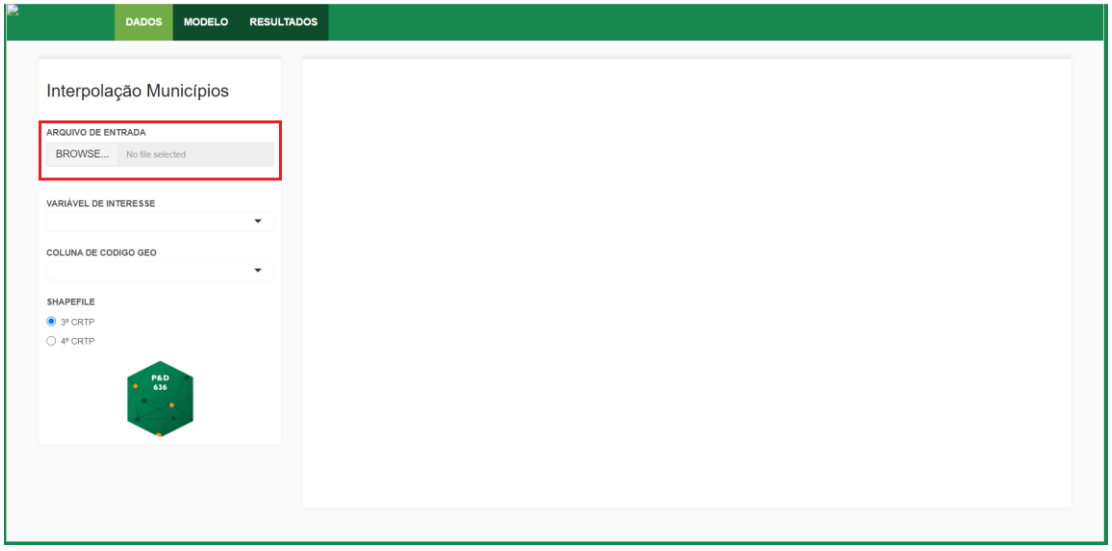
3. O aplicativo já possui uma base interna para a realização da operação de interpolação, com os seguintes indicadores socioeconômicos configurados:
   - Média do produto Interno Bruto (PIB) dos municípios de Minas Gerais nos anos de 2014 a 2017;
   - Taxa de homicídios dos municípios de Minas Gerais em 2017;
   - Capacidade de pagamento (CAPAG) dos municípios de Minas Gerais no ano de 2020;
   - Índice de desenvolvimento da educação básica (IDEB) coletadas do estado de Minas Gerais no ano de 2019.

Caso as interpolações a serem realizadas utilizarem as variáveis descritas, siga para o tópico 8 deste manual. Caso deseje inserir novos dados municipais a serem interpolados a partir de uma base de dados externa, siga para o próximo passo.

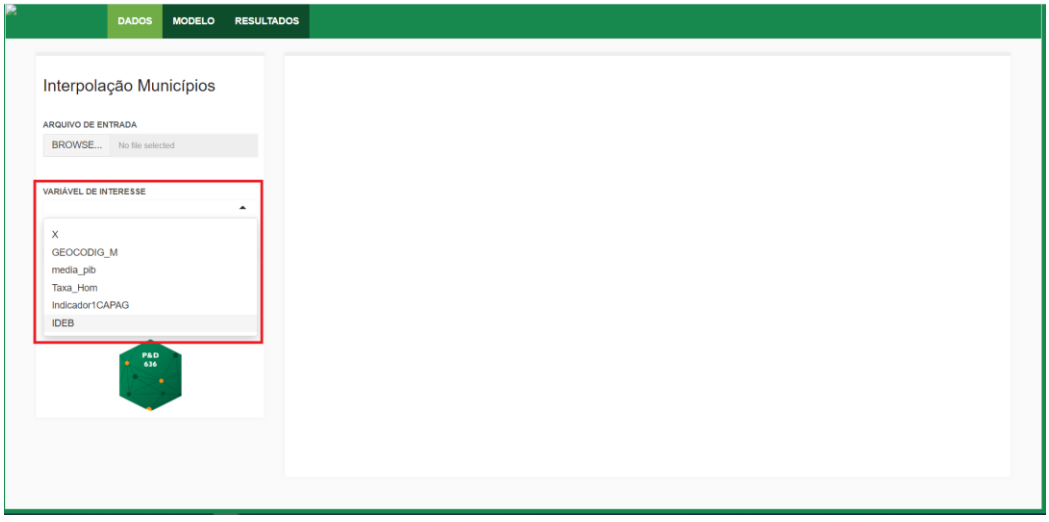
4. Para configurar uma base de dados externa é importante que a mesma siga a seguinte estrutura:
    - O arquivo deve estar em formato .xlsx (Excel);
    - É necessário que exista uma coluna contendo o geocódigo do IBGE referente ao município ao qual o dado pertence;

Vale ressaltar também que é fundamental que o formato da variável de interesse seja numérico, para que não haja problemas no processamento da interface durante a interpolação.

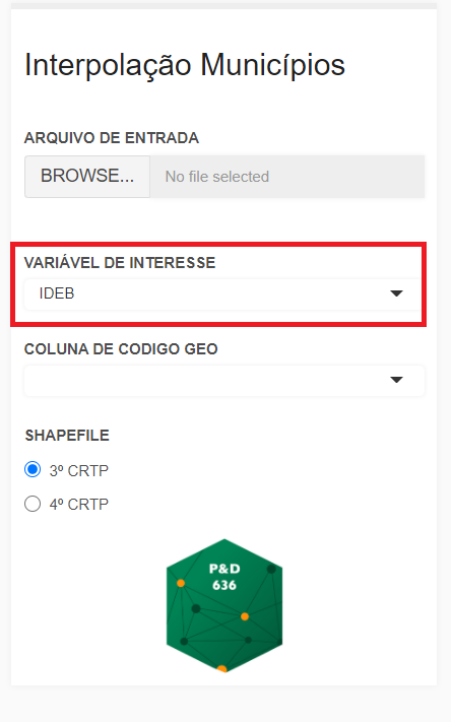
5. Com a base de dados em formato excel estruturada corretamente, basta clicar em "**BROWSE**" no campo "**ARQUIVO DE ENTRADA**" e selecionar o arquivo a ser utilizado, como mostra a figura a seguir. Feito isso, a base de dados será carregada para a interface.



6. No campo "**VARIÁVEL DE INTERESSE**" dever ser selecionada a variável a ser interpolada.

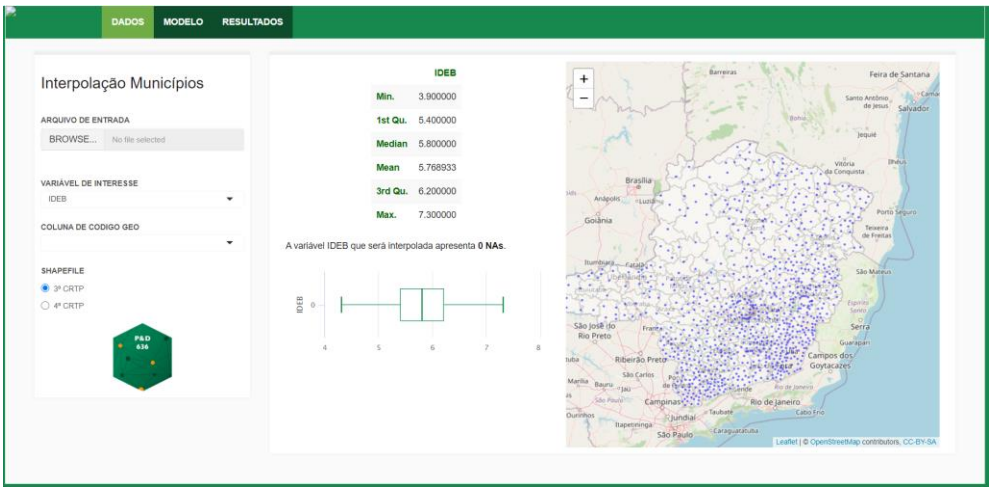


7. Uma vez carregada a base de dados, todas as colunas presentes serão disponibilizadas. Caso não seja selecionada uma variável numérica, a interface retornará uma mensagem de erro. No exemplo a seguir, será selecionada a variável referente ao IDEB:

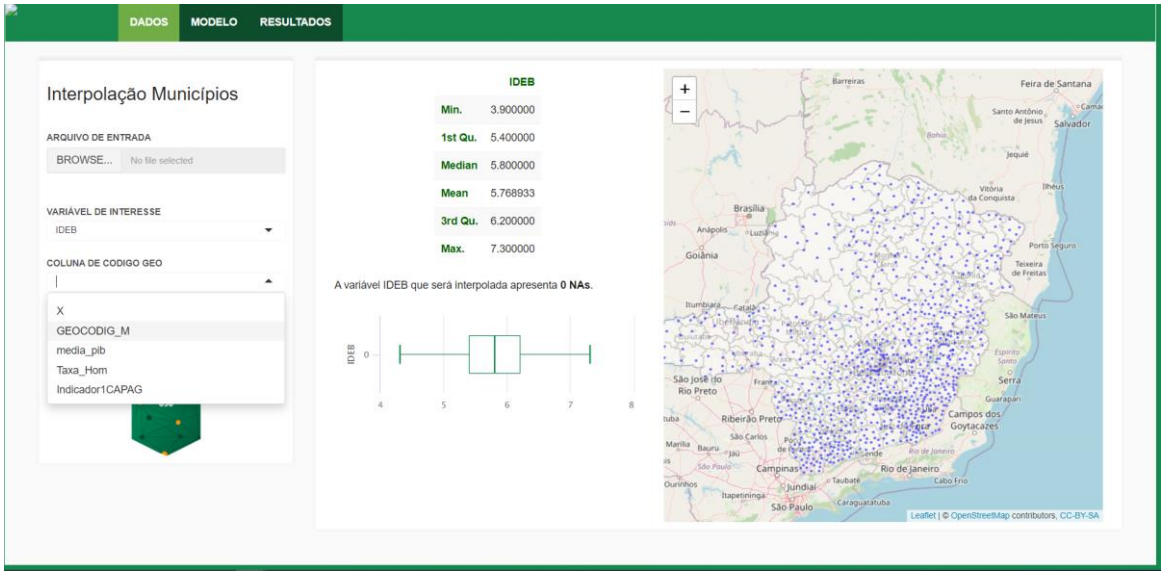


8. Após selecionada a variável de interesse, serão apresentadas automaticamente na tela as análises descritivas, contendo informação do valor mínimo, do primeiro quartil (valor que abrange 25% dos dados), mediana (valor central, que corresponde a 50% dos dados), média, terceiro quartil (valor que abrange 75% dos dados) e o valor máximo. Todas essas informações são resumidas num gráfico do tipo Boxplot.

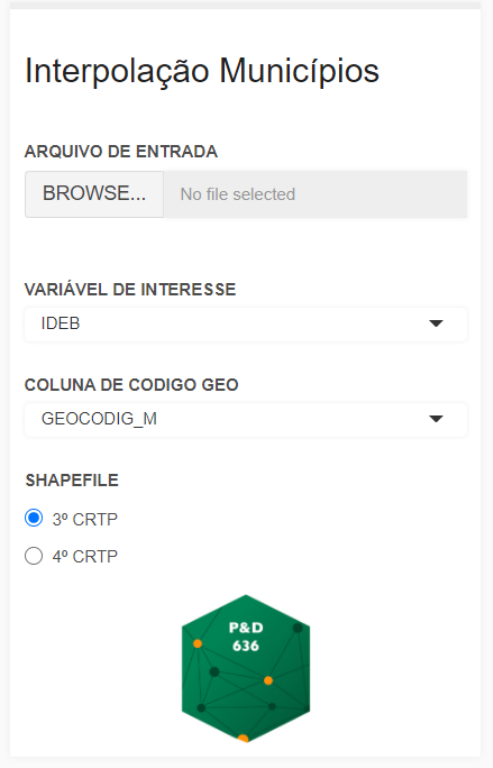
Além disso, também é informada a quantidade de valores nulos encontrados na variável de interesse e um mapa inicial é apresentado. Os pontos em azul do mapa são os centróides (pontos centrais) dos 853 municípios de Minas Gerais.



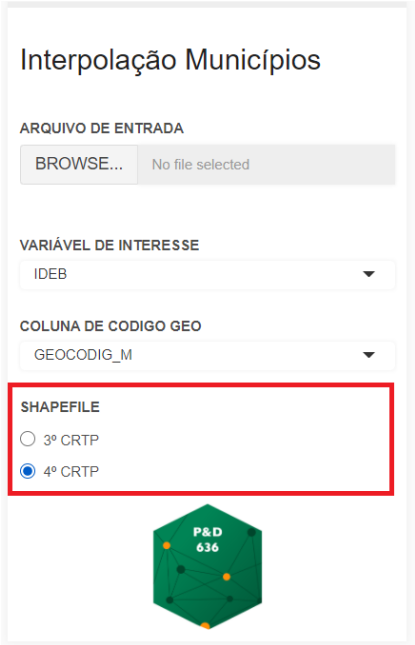
9. Em seguida, na opção "**COLUNA DE CÓDIGO GEO**", deve ser selecionada qual coluna da base de dados representa o geocódigo municipal do IBGE, como mostrado a seguir.



10. Caso não seja selecionada uma variável não reconhecida como geocódigo (na estrutura definida pelo IBGE) a interface retornará uma mensagem de erro. Segue exemplo de uma variável selecionada, neste caso, denominada GEOCODIG_M.



11. Por fim, o último dado a ser configurado diz respeito a qual shapefile da CEMIG será utilizado na interpolação. O shapefile do 3º CRTP (Ciclo de revisão tarifária periódica) diz respeito à configuração utilizada pela CEMIG até o ano de 2018, em que existiam 271 conjuntos elétricos e o shapefile do 4º CRTP, refere-se à configuração definida a partir de 2019 com 295 conjuntos elétricos.

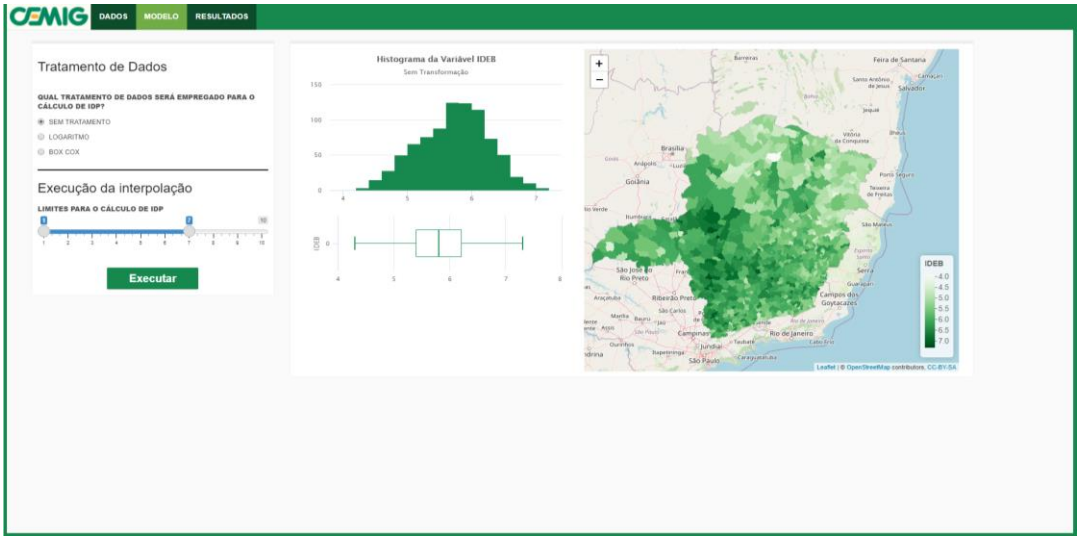


12. Com todos as opções selecionadas, o usuário pode prosseguir para a aba “**MODELO**”.

MODELO

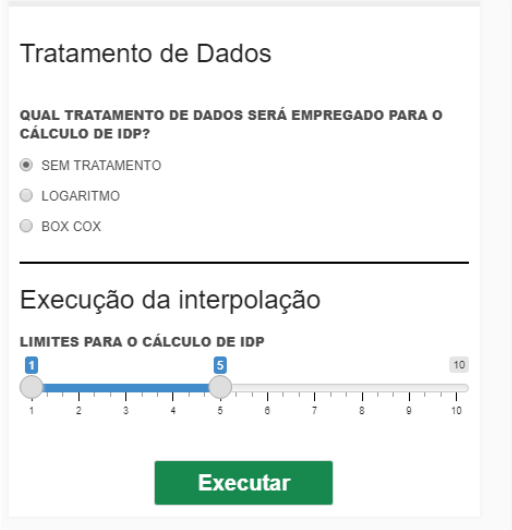
Na aba “**MODELO**”, o usuário configura o procedimento de Interpolação Espacial. De forma sucinta, o parâmetro IDP define a taxa de decaimento dos pesos quando na interpolação espacial. Maiores detalhes sobre o parâmetro IDP são descritos no apêndice.

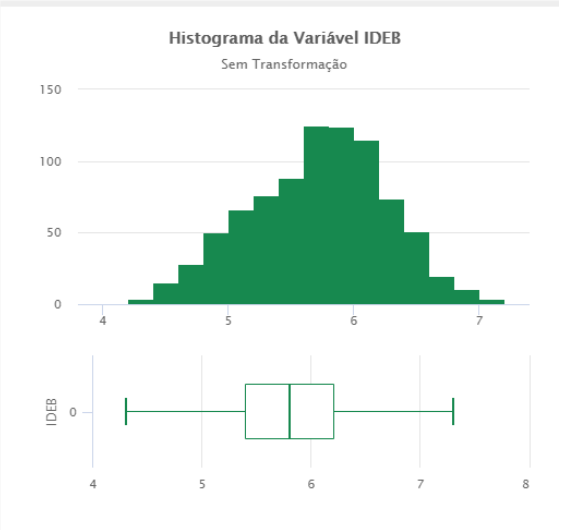
Na aba “**MODELO**”, são apresentados gráficos de análise exploratória da variável escolhida previamente (aba "**DADOS**").

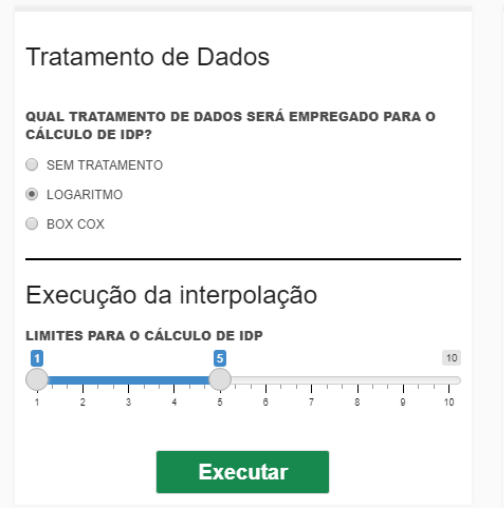


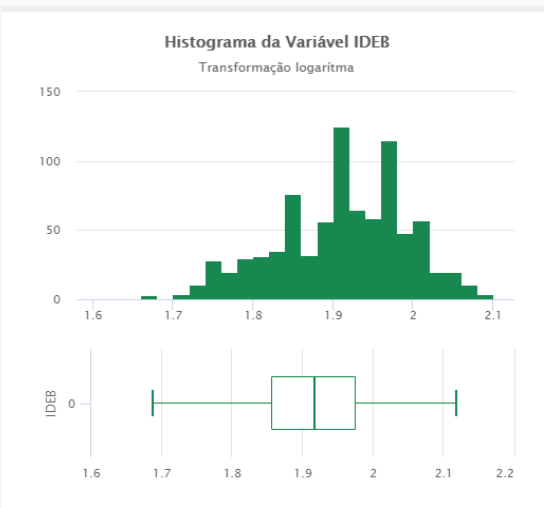
O histograma e o boxplot dizem respeito à distribuição das observações da variável de interesse definida na aba "**DADOS**", enquanto o mapa de Minas Gerais, segregado por municípios, apresenta visualmente como essas observações estão distribuídas ao longo do estado. Ressalta-se que a coloração do mapa é intuitiva, de forma que os maiores valores apresentam a coloração mais escura. Destaca-se que o mapa é interativo, permitindo que o usuário tenha acesso a informações específicas de cada município ao passar o mouse ou clicar sobre eles. Além disso, os gráficos são atualizados à medida que os parâmetros dessa aba são definidos. Para utilizar a aba "**MODELO**", siga os passos a seguir:

13. O usuário deve definir qual é o tratamento de dados que será usado para a definição do valor do IDP usado na interpolação. É interessante que os dados apresentem uma distribuição semelhante a uma distribuição simétrica (como uma distribuição normal, por exemplo) para se obter melhores resultados. Portanto, caso os dados brutos (sem tratamento) apresentem uma distribuição fortemente assimétrica, dois tratamentos podem ser escolhidos: (a) a transformação logarítmica ou (b) a transformação de Box-Cox. O histograma e o boxplot serão modificados automaticamente para apresentar a distribuição dos dados seguindo a transformação desejada, garantindo uma visualização fácil e um auxílio necessário na identificação de qual deve ser a opção mais adequada.



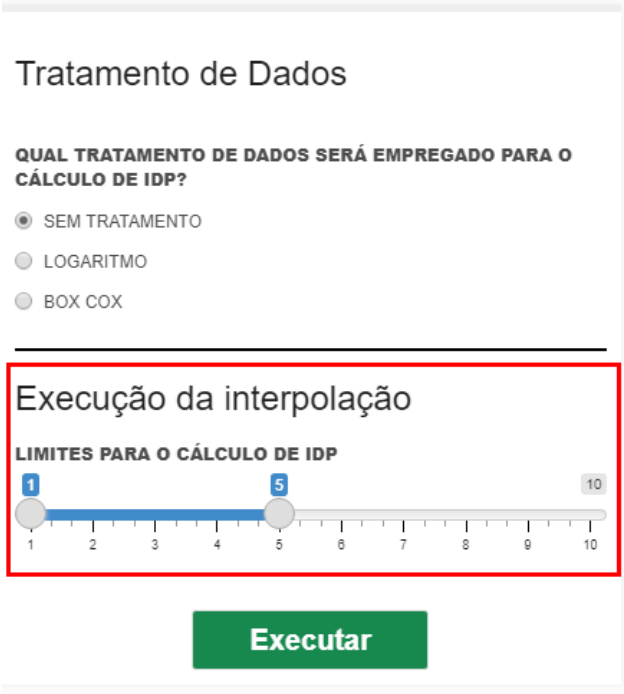




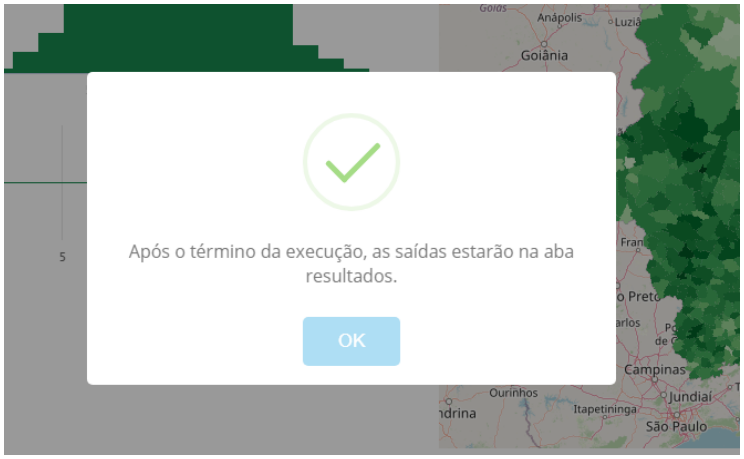


OBS: É importante ressaltar que a essa escolha (entre dados brutos, transformação log e transformação de Box-Cox) impactará somente o cálculo do IDP, e não será usada para o resultado final da interpolação. Isso significa que cada conjunto elétrico receberá valores para os dados brutos da variável de interesse, e não o log(variável de interesse), por exemplo, mesmo que use a transformação log para o cálculo de IDP.

14. Na sequência, o usuário deve definir o intervalo para o cálculo do valor ótimo do IDP, que minimiza o erro da interpolação espacial. Como o processo de busca do valor que minimiza o erro é computacionalmente intenso, demandando um certo tempo para ser executado, um grid de tamanho fixo de potenciais valores para o parâmetro IDP será usado na tentativa de se encontrar o menor erro possível. Caso o usuário deseje obter resultados mais precisos para o IDP, os limites de busca podem ser reduzidos, mantendo a solução do IDP obtida em uma primeira tentativa em uma posição central nos novos limites.



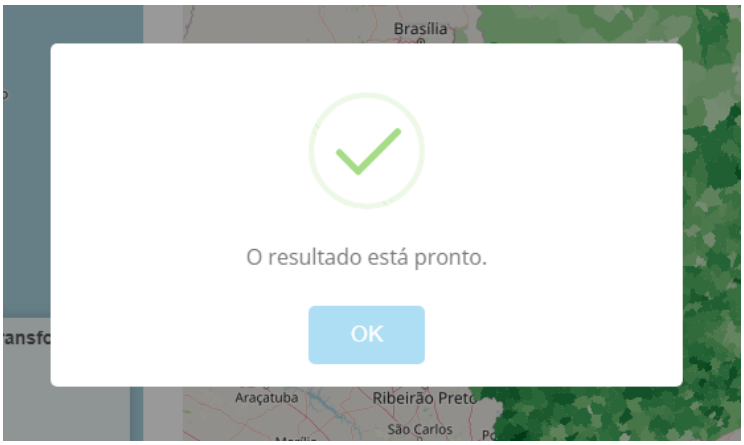
15. Com o tratamento de dados do cálculo do IDP e o intervalo de busca definidos, o usuário poderá então iniciar o processo de execução da Interpolação Espacial. Para tanto, basta clicar no botão de Executar. Uma mensagem de início da execução será apresentada ao se clicar no botão, e o usuário poderá fechar a mensagem e aguardar pelos resultados na aba "**RESULTADOS**".



RESULTADOS

Na aba "**RESULTADOS**", símbolos de execução estarão presentes enquanto os resultados são gerados. Ressalta-se que o procedimento é computacionalmente intenso, e que alguns minutos serão necessários para a execução de todos os cálculos.

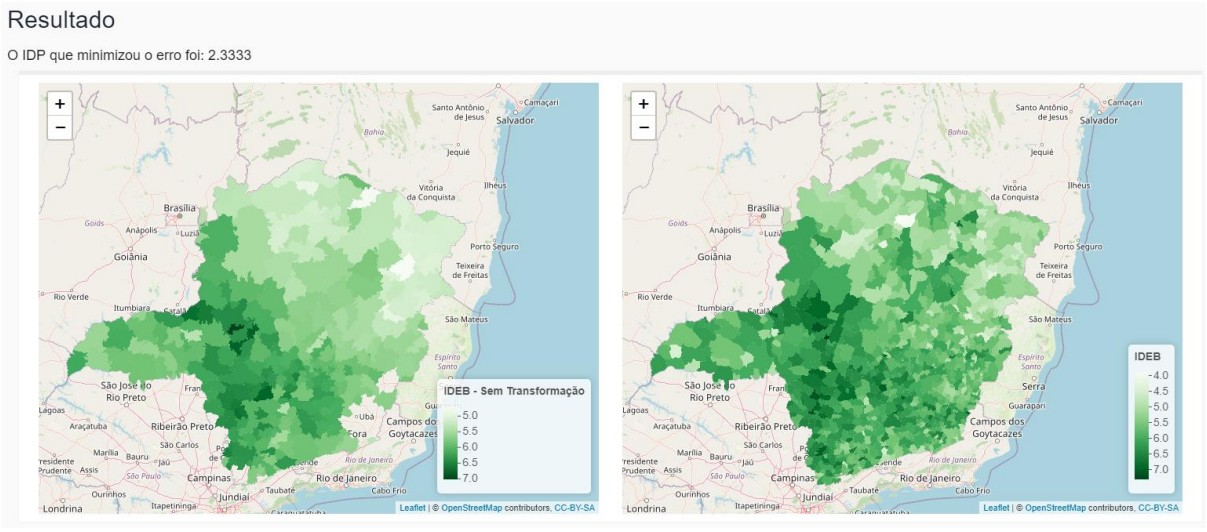
16. Quando o resultado estiver pronto, uma mensagem irá aparecer para o usuário, independentemente de qual aba ele estiver, e os mapas apresentando os resultados serão exibidos.



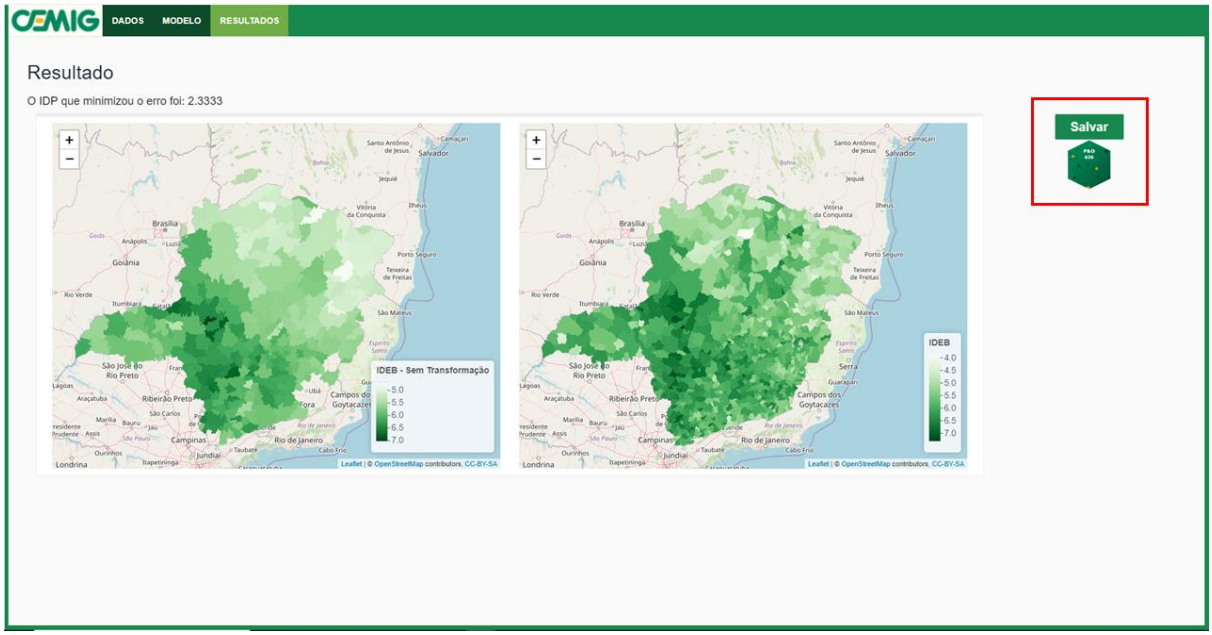
Após a execução da interpolação espacial, a aba “**RESULTADOS**” apresentará dois gráficos.

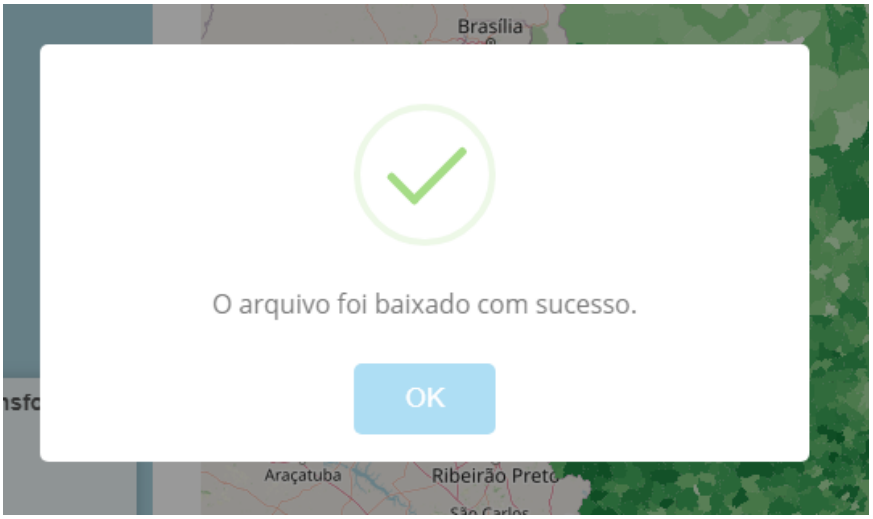
O gráfico da esquerda representa a distribuição da variável de interesse em relação aos conjuntos elétricos da CEMIG-D segundo o CRTP escolhido. Este gráfico é o resultado da interpolação, ou seja, os valores referentes aos municípios foram alocadas aos conjuntos elétricos. Acima do gráfico, é possível identificar o valor do IDP que minimizou o erro.

O gráfico da direita representa a distribuição da variável de interesse em relação aos municípios de Minas Gerais. Ele é o mesmo gráfico apresentado na aba "**MODELO**", e se encontra também em "**RESULTADOS**" para uma análise comparativa entre o resultado da interpolação (por conjuntos elétricos) e o gráfico por municípios. Gráficos semelhantes indicam que a interpolação foi bem executada.



17. Finalmente, para o usuário exportar uma base de dados contendo o resultado da interpolação espacial, basta clicar no botão de "salvar". Ao clicar nesse botão, a interface irá automaticamente gerar uma planilha no excel contendo os valores obtidos da variável resposta para cada conjunto elétrico. Essa planilha terá o nome "**Saida_Interp.xlsx**". Uma mensagem de sucesso no download será apresentada.





18. A base de dados exportada apresenta as informações sobre os conjuntos elétricos (código, descrição e nome) presentes nos arquivos georreferenciados referentes ao 3º ou 4º CRTP, juntamente com o dado interpolado na última coluna. Sendo assim, encerra-se o uso da interface, e o usuário pode aplicar os dados interpolados em suas futuras análises.

	A	B	C	AW
1	Codigo	DSC_C_N	conjunt	IDEB_estimado
2	15303	CENTRALINA	CENTRALINA	5,887007896
3	15312	ITABIRITO	ITABIRITO	6,133096604
4	15130	BH SION	BH SION	6,10805005
5	15357	CARMO DO F	CARMO DO F	6,586651823
6	15343	ENGENHEIRC	ENGENHEIRC	5,596344116
7	15160	COUTO MAG	COUTO MAG	5,663811122
8	15350	CORDISBURG	CORDISBURG	5,788495144
9	15138	BRASOPOLIS	BRASOPOLIS	5,705279057
10	15120	BH ATALAIA	BH ATALAIA	6,097125722
11	15293	JANAUBA 1	JANAUBA 1	5,40815463
12	15108	BAMBUI	BAMBUI	6,265137253
13	15151	CI CONTAGE	CI CONTAGE	6,060324908
14	15321	IPATINGA 2	IPATINGA 2	5,919804012
15	15157	CLAUDIO 1	CLAUDIO 1	6,381208753
16	15104	ARAGUARI 2	ARAGUARI 2	6,273903354
17	15145	CARLOS CHA	CARLOS CHA	5,226479338
18	15201	SACRAMENT	SACRAMENT	6,167819811
19	15260	NOVA LIMA	NOVA LIMA	6,197706256
20	15333	GOVERNADC	GOVERNADC	5,312404368
21	15099	ALPINOPOLI	ALPINOPOLI	6,558474946

MÓDULO DE REGIONALIZAÇÃO UNIVARIADA

OBJETIVO DO APLICATIVO: A interface computacional de regionalização univariada foi desenvolvida com o intuito de possibilitar a identificação de *clusters espaciais* (agrupamentos) de conjuntos elétricos que apresentam resultados semelhantes para uma determinada variável numérica de interesse. A interface possibilita, portanto, que as regiões geográficas que apresentam semelhanças com relação ao valor médio da variável de interesse sejam devidamente identificadas.

INSTRUÇÕES DE USO:

1. Para acessar o módulo de regionalização univariada, é necessário chamar a função "**shiny_regio_uni**"como descrito a seguir:

   ```
   > shiny_regio_uni()
   ```

2. Após a execução da função, o aplicativo Shiny com o módulo de regionalização univariada será aberto automaticamente na aba "**DADOS**", como apresentado na imagem abaixo:

DADOS

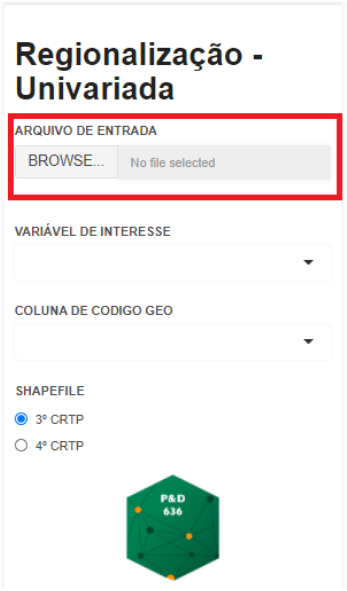
3. Para iniciar o uso da interface, será preciso importar a base de dados que contém a variável de interessa a ser usada na análise. Essa base deve respeitar os seguintes requisitos:

   - O arquivo deve estar em formato .xlsx (Excel);
   - É necessário que exista uma coluna contendo o código dos conjuntos elétricos da CEMIG-D;

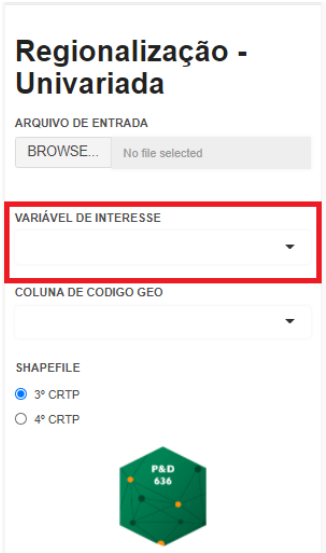
- O usuário precisa ter conhecimento de qual configuração de conjuntos elétricos a base está relacionada (3º ou 4º CRTP).

Vale ressaltar também que o formato da variável de interesse em questão deve ser numérico, para evitar possíveis problemas de processamento durante a execução da análise.

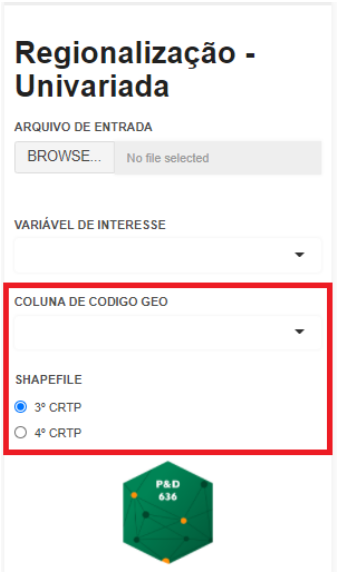
4. Com a base de dados em formato Excel estruturada corretamente, basta clicar em "**BROWSE**" no campo "**ARQUIVO DE ENTRADA**" e selecionar o arquivo a ser utilizado, como mostra a figura a seguir. Feito isso, a base de dados será carregada para a interface.



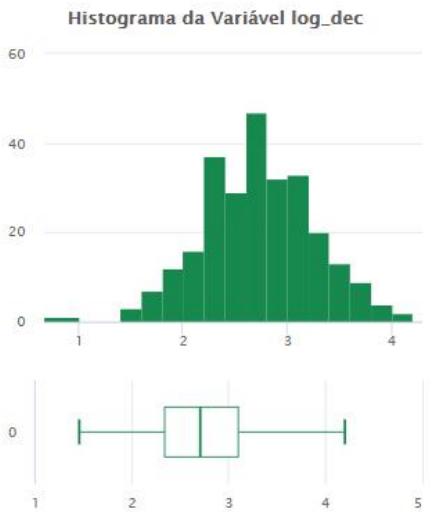
5. Após a definição da base e a conclusão de seu upload, o usuário deverá então definir qual será a variável de interesse a ser regionalizada. Essa atribuição será feita no campo "**VARIÁVEL DE INTERESSE**". Ressalta-se que as opções apresentadas no campo serão exatamente as variáveis contidas na base de dados importada.

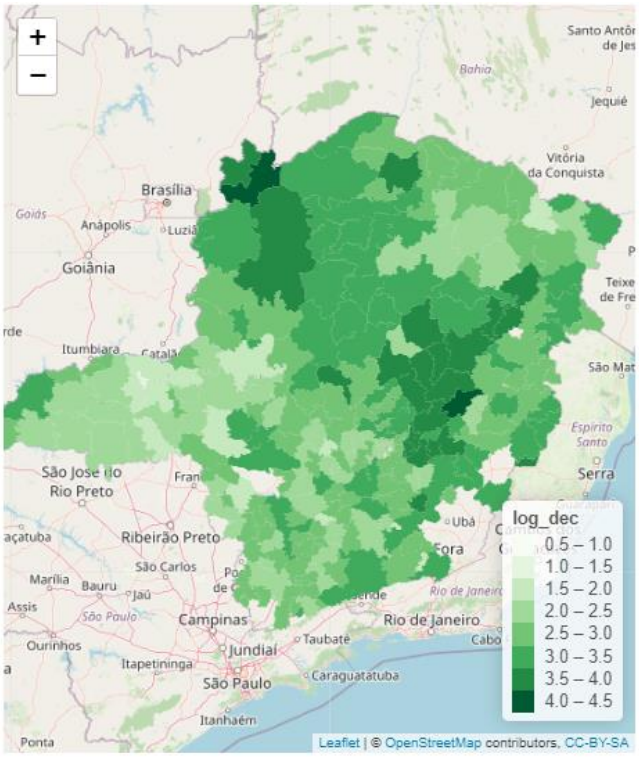


6. O usuário deverá também identificar, no campo "**COLUNA DE CODIGO GEO"** qual é a coluna da base de dados referente aos códigos dos conjuntos elétricos, para que os dados sejam devidamente atribuídos aos conjuntos. Além disso, é preciso apontar no campo "**SHAPEFILE"** se os conjuntos elétricos em questão seguem a estrutura referente ao 3º ou 4º CRTP.



Uma vez com todos os campos da aba "**DADOS**" devidamente preenchidos, a interface irá apresentar três gráficos relacionados à variável de interesse especificada:





a. O primeiro deles será um histograma da variável de interesse, apresentando a distribuição dos dados em diversas classes e a frequência absoluta com que o valor da classe ocorre, contendo todas as observações presentes na base de dados;

b. O segundo recurso gráfico apresentado é um gráfico Boxplot, que busca analisar a distribuição dos dados a partir de outra perspectiva. O gráfico resume a informação do valor mínimo, do primeiro quartil (valor que abrange 25% dos dados), mediana (valor central, que corresponde a 50% dos dados), terceiro quartil (valor que abrange 75% dos dados) e o valor máximo de todas as observações.

c. O terceiro será um mapa apresentando a configuração dos conjuntos elétricos da CEMIG-D de acordo com o CRTP escolhido. A escala de coloração dos conjuntos elétricos representa o valor da variável de interesse definida. Ressalta-se que esse mapa é dinâmico, e mais informações podem ser obtidas caso o usuário clique sobre um determinado conjunto (nome do conjunto, valor da variável para o conjunto correspondente).

Com os campos devidamente preenchidos, o usuário poderá então iniciar a manipulação dos parâmetros da metodologia de regionalização na aba "**RESULTADOS**".

Observação: Para o desenvolvimento desse manual, foi utilizado de maneira arbitrária a variável "log_dec" como variável de interesse, somente para visualização dos passos a serem executados.

RESULTADOS

Na aba "**RESULTADOS**", o usuário deverá ajustar parâmetros específicos para o processamento da regionalização univariada.

7. O primeiro campo é o campo **"PARÂMETROS DE AJUSTE (C, STEPS, BURN-IN)"**, responsável pela atribuição de parâmetros mais técnicos necessários para a análise. Nele, o usuário precisará definir os valores para:

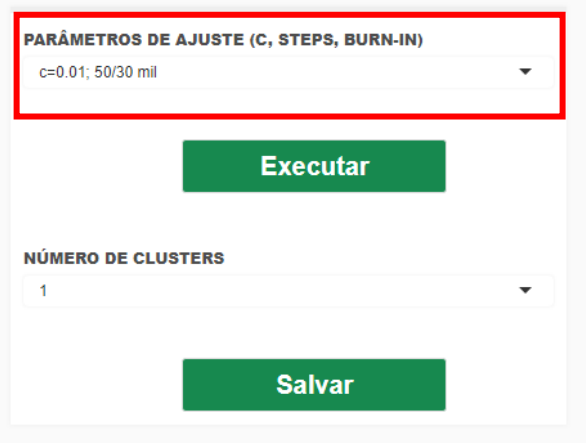
    a. **c**: parâmetro para ajuste a priori do tamanho dos clusters (caso c = 0.01, os clusters serão maiores; caso c = 0.35, os clusters serão mais compactos).

    b. **Steps**: define o comprimento da cadeia de Markov implementada.

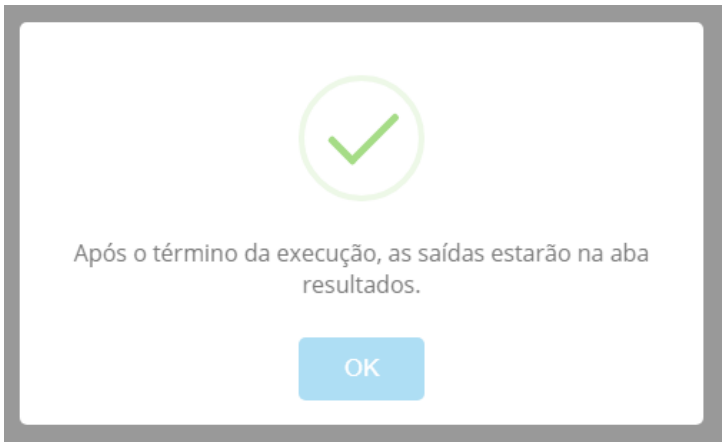
    c. **Burn-in**: número de valores no início da cadeia que serão descartados.

Ressalta-se que o parâmetro **Steps** impactará diretamente no tempo necessário para o processamento e execução da análise. Quanto maior o comprimento da cadeia de Markov, maior o tempo de processamento.

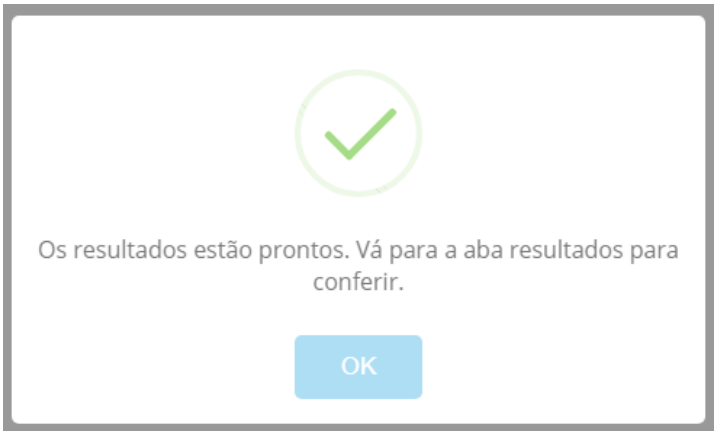
Para aumentar a comodidade e a usabilidade da interface e garantir resultados pertinentes de acordo com o método usado, algumas configurações apropriadas desses três parâmetros foram previamente definidas. O usuário pode escolher qual dessas configurações usar.

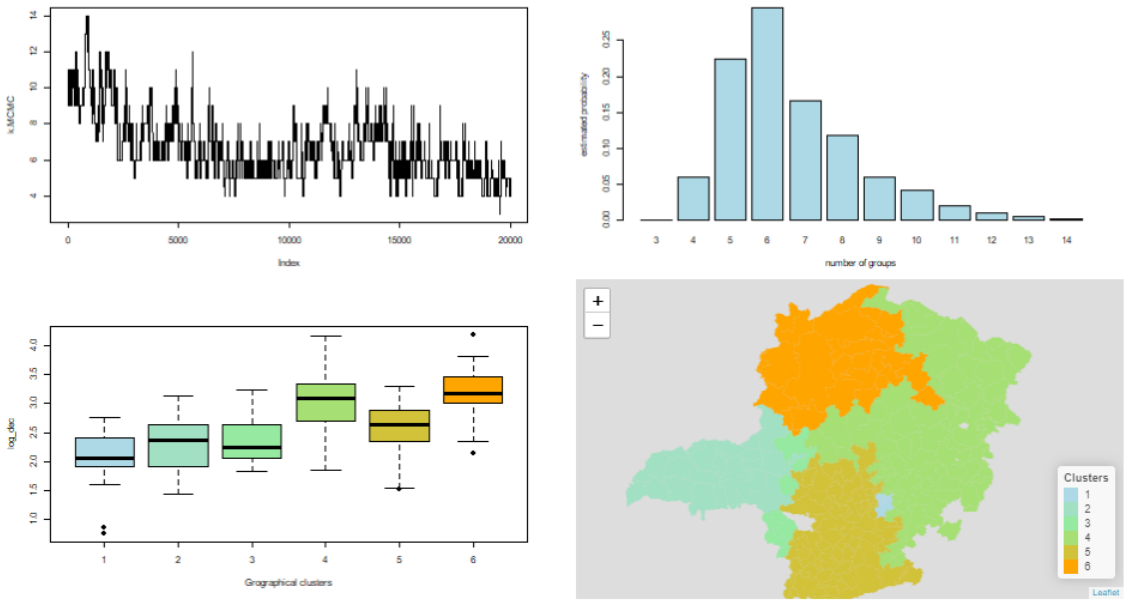


8. Após a definição dos parâmetros no campo "**PARÂMETROS DE AJUSTE**", o usuário deverá então ativar o botão "**EXECUTAR**" para que o processamento da regionalização univariada comece. Uma mensagem de início da execução será apresentada, e o ícones de carregamento aparecerão na própria aba de **RESULTADOS**.



Uma vez terminado o processamento, uma mensagem de finalização também será apresentada, indicando que os resultados foram compilados. Na aba **RESULTADOS**, quatro gráficos serão então construídos.





Ressalta-se que os dois gráficos superiores deverão ser analisados primeiramente. Isso se deve ao fato de que eles apresentam resultados referentes ao processamento da regionalização univariada, enquanto os dois gráficos inferiores apresentam informações sobre a clusterização escolhida. Seguem suas descrições:

9. O gráfico **superior esquerdo** indica os resultados do Método de Cadeia de Markov Monte Carlo (MCMC – *Markov Chain Monte Carlo*).

10. O gráfico **superior direito** apresenta a probabilidade estimada de que um determinado número de clusters seja o mais apropriado para aquela regionalização. O valor de "number of groups" que apresentar a maior probabilidade estimada é o valor que indicará o número de clusters mais adequado à variável de interesse ao longo da área de concessão da CEMIG-D. A título de exemplo, é possível perceber que 6 clusters é a quantidade mais provável de existir para o caso da variável de interesse "log_dec".

A definição de qual será o número de clusters irá alterar os dois gráficos inferiores, e ela deve ser feita tendo conhecimento de qual é o valor mais adequado. É por esse motivo que os dois gráficos superiores são analisados inicialmente.

Com as informações dos gráficos superiores devidamente analisadas, o usuário poderá então estudar os gráficos inferiores, que dizem respeito à configuração dos conjuntos elétricos nos clusters identificados.

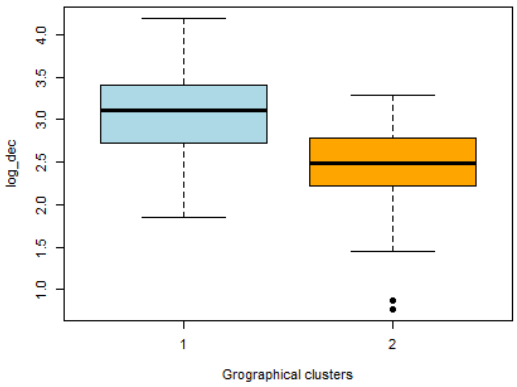
11. O gráfico **inferior esquerdo** é um conjunto de boxplots representando as observações em cada cluster. A coloração dos boxplots determina qual é o cluster geográfico considerado em cada um dos gráficos (se os marcadores forem verdes, o gráfico diz respeito ao cluster com coloração verde no mapa). Esses gráficos permitem uma comparação de como se comporta a distribuição dos dados em cada um dos clusters.

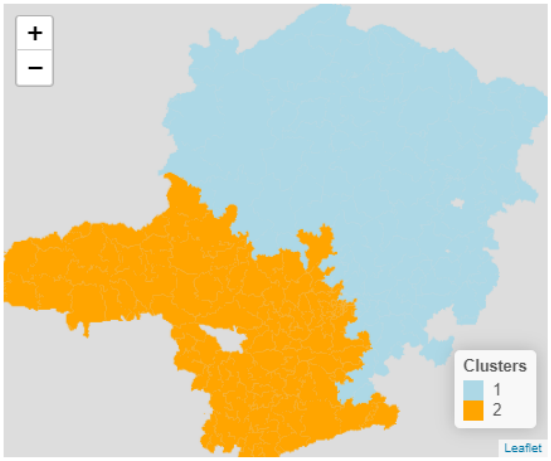
12. O gráfico **inferior direito** apresenta a repartição territorial dos clusters, onde cada cor indica um agrupamento específico de conjuntos elétricos que apresentaram um comportamento específico em seu valor médio. O mapa é interativo, de forma que o

usuário pode clicar em um determinado conjunto elétrico para identificar em qual cluster ele ficou situado.

13. O usuário poderá notar que, no término da execução da regionalização, o número de clusters será automaticamente alterado para o valor com maior probabilidade estimada, e os gráficos inferiores serão montados de acordo com esse número de clusters. Entretanto, para fomentar as análises e a visualização de outras possíveis formatações de agrupamentos, o usuário poderá livremente alterar o número de clusters usando o campo "**NÚMERO DE CLUSTERS**". Aconselha-se que outros números de clusters que apresentaram um valor razoável de probabilidade estimada também sejam visualizados, para melhor compreensão do comportamento da regionalização.
    Ao alterar o valor em "**NÚMERO DE CLUSTERS**", a nova informação será processada, e os dois gráficos inferiores serão devidamente alterados após um período de processamento.

A título de exemplo, mudando o número de clusters da variável de interesse "log_dec" para 2, os resultados dos dois gráficos inferiores serão alterados para:

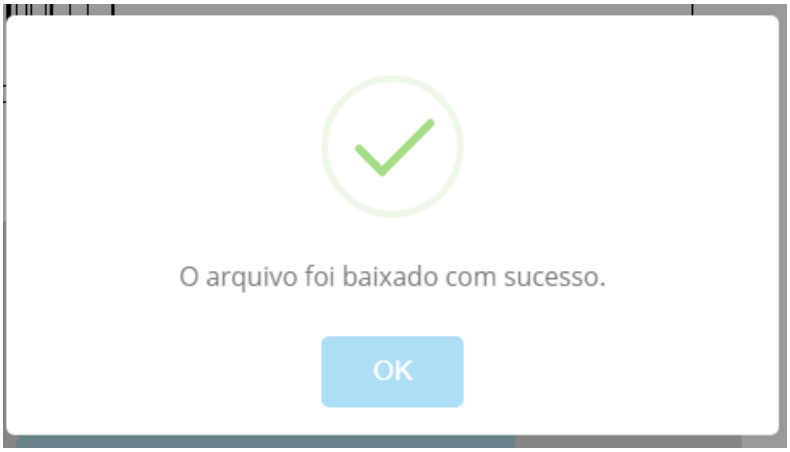




14. Finalmente, com todos os cenários de interesse do usuário devidamente analisados, o usuário poderá então salvar a base de dados contendo a informação do índice do cluster ao qual cada conjunto elétrico se encontra, utilizando o botão "**SALVAR**". A base de saída apresentará as mesmas informações da base original, importada no início do procedimento, com a adição de uma nova coluna denominada "**grupos**" que indicará o

índice do cluster alocado para cada um dos conjuntos elétricos. Vale lembrar que o usuário, antes de salvar, poderá alterar o valor de "**NÚMERO DE CLUSTERS**" para aquele que lhe foi mais interessante.

Antes de salvar, o usuário poderá visualizar uma tabela contendo código dos conjuntos, a variável de interesse e os grupos definidos na aba "**TABELA**", e ao clicar no botão de salvar, uma mensagem de confirmação de download será apresentada.



Segue uma imagem da tabela final exportada, cujo nome padrão é "**geoUniOutput.xlsx**":

	A	B	AW
1	codigo	dsc_conj_n	grupos
2	15092	ARACAGI	1
3	15093	ABADIA DOS	3
4	15094	ARCOS 1	3
5	15095	AREADO 2	3
6	15096	ABAETE 2	3
7	15097	ALFENAS 1	3
8	15098	AGUAS FORM	1
9	15099	ALPINOPOLI	3
10	15100	ALMENARA	1
11	15101	ANDRADAS 2	3
12	15102	AIMORES	1
13	15103	ARAPORA	3
14	15104	ARAGUARI 2	3
15	15106	AVATINGUAI	3
16	15107	ARAXA 1	3
17	15108	BAMBUI	3
18	15109	BARBACENA	2

MÓDULO DE REGIONALIZAÇÃO DE REGRESSÕES LINEARES SIMPLES

OBJETIVO DO APLICATIVO: A interface computacional de regionalização de regressões lineares simples foi desenvolvida com o intuito de possibilitar a identificação de *clusters* (agrupamentos) de conjuntos elétricos que apresentam resultados homogêneos quando no ajuste de modelos de regressão linear simples, ou seja, define-se uma determinada variável resposta Y e uma determinada variável preditora **x**. O modelo estatístico implementado nesta análise é descrito a seguir:

$$Y_i = \beta_{0j} + \beta_{1j} x_i + \epsilon_i$$

onde $i = 1, \dots, n$ e $j = 1, \dots, K$. O objetivo é encontrar K partições geográficas contíguas dos conjuntos elétricos onde, em cada partição, os parâmetros de intercepto (β_0) e coeficiente de regressão (β_1) são distintos. O método proposto permite estimar o número de partições K, suas localizações e seus respectivos parâmetros de regressão. Maiores detalhes da metodologia proposta podem ser encontrados em Costa e al. (2021).

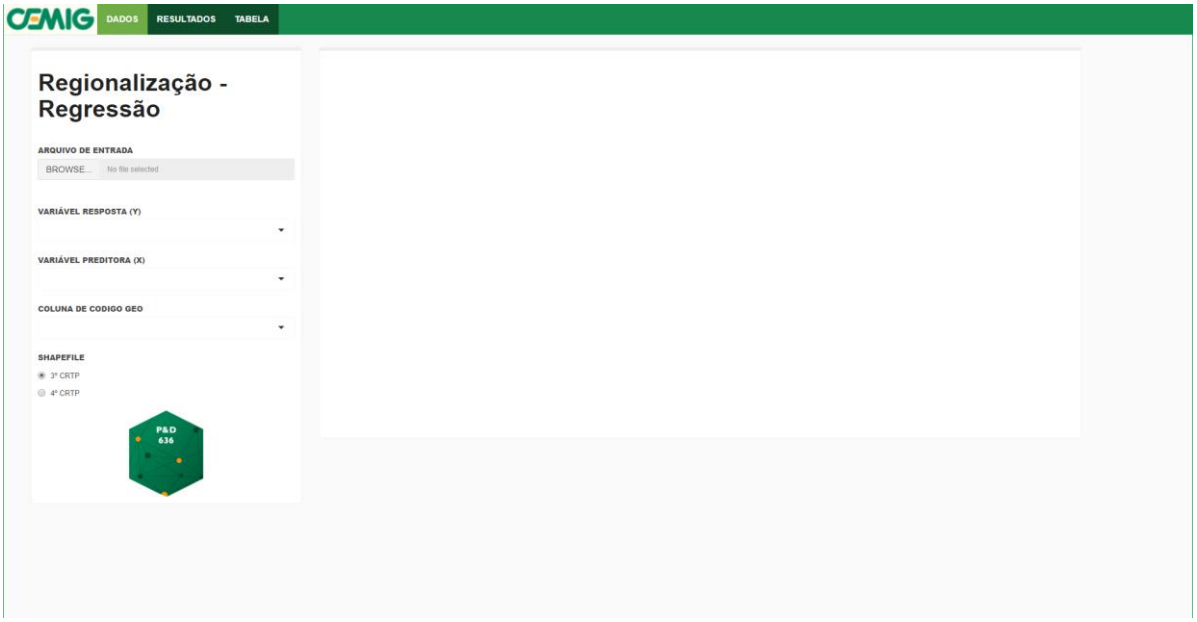
A interface possibilita, portanto, que os agrupamentos geográficos e os respectivos parâmetros de regressão sejam estimados.

INSTRUÇÕES DE USO:

1. Para acessar o módulo de regionalização de regressão, será necessário chamar a função “shiny_regio_reg()como descrito abaixo:

   ```
   > shiny_regio_reg()
   ```

2. Após a execução da função, o aplicativo Shiny com o módulo de regionalização de regressão será aberto automaticamente na aba “**DADOS**”, como apresentado na imagem abaixo:



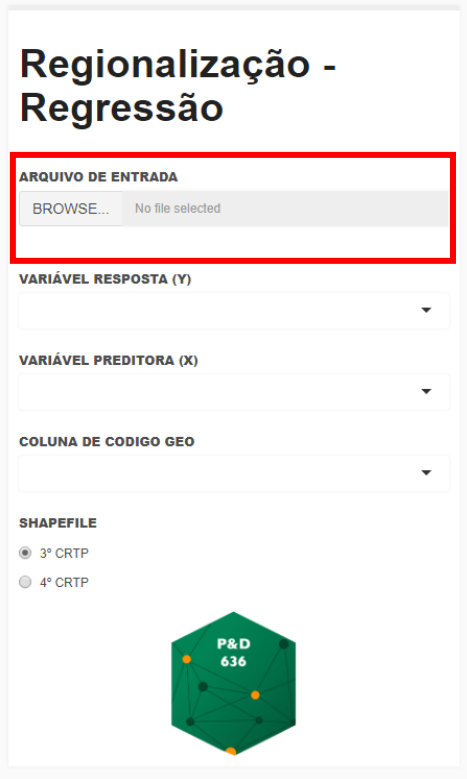
DADOS

a. Para iniciar o uso da interface, será preciso importar a base de dados que contém a variável resposta Y e a variável preditora X. A base deve estar em formato Excel (.xlsx) e deve conter uma coluna com o código dos conjuntos elétricos da CEMIG-D;

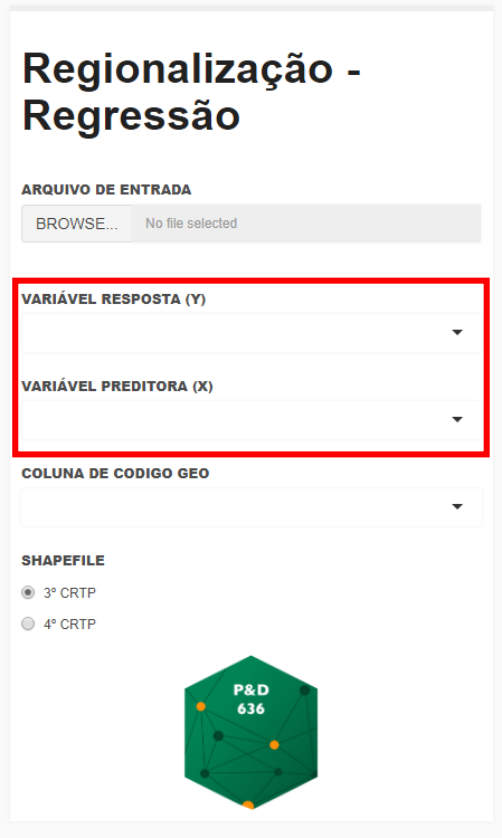
b. O usuário precisa ter conhecimento de qual configuração de conjuntos elétricos a base está relacionada (3º ou 4º CRTP).

Vale ressaltar também que o formato das variáveis X e Y em questão deve ser numérico, para evitar possíveis problemas de processamento durante a execução da análise.

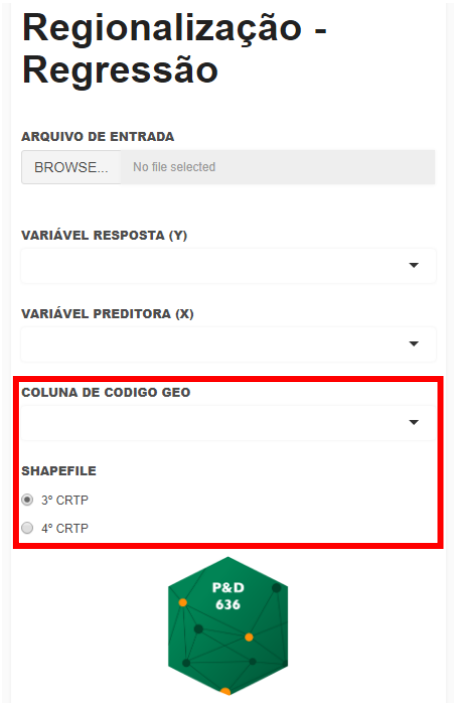
3. Com a base de dados em formato Excel estruturada corretamente, basta clicar em "**BROWSE**" no campo "**ARQUIVO DE ENTRADA**" e selecionar o arquivo a ser utilizado, como mostra a figura a seguir. Feito isso, a base de dados será carregada para a interface.



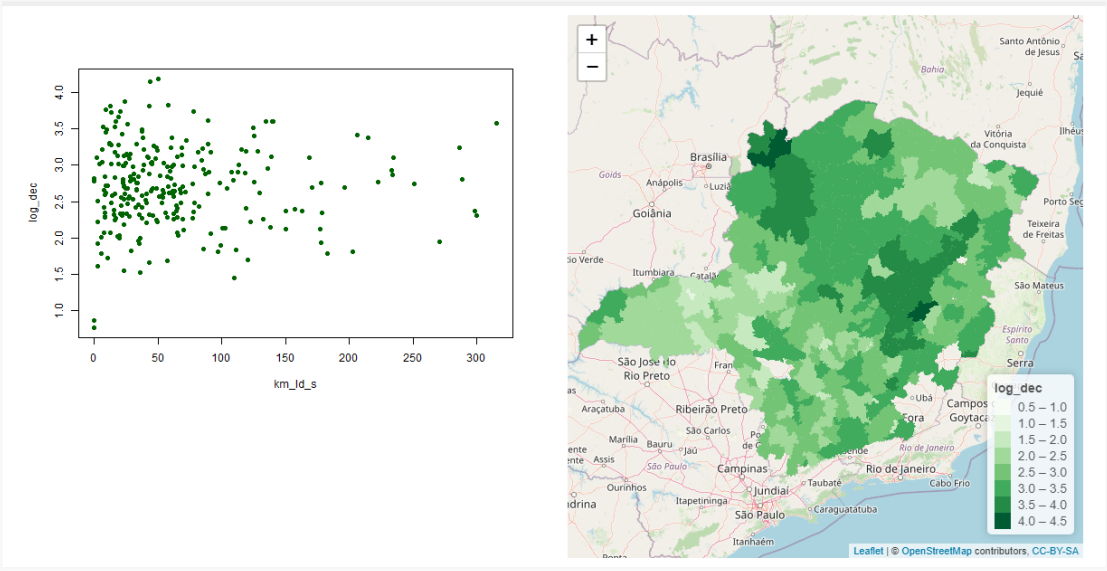
4. Após a definição da base e a conclusão de seu *upload*, o usuário deverá definir a variável resposta (Y) e a variável preditora (X). Essas atribuições serão feitas nos campos "**VARIAVEL RESPOSTA (Y)**" e "**VARIAVEL PREDITORA (X)**". Ressalta-se que as opções apresentadas em ambos os campos serão exatamente as variáveis contidas na base de dados importada.



5. O usuário deverá também identificar, no campo "**COLUNA DE CODIGO GEO"** qual é a coluna da base de dados referente aos códigos dos conjuntos elétricos, para que os dados sejam devidamente atribuídos a aos conjuntos. Além disso, é preciso apontar no campo "**SHAPEFILE"** se os conjuntos elétricos em questão seguem a estrutura referente ao 3º ou 4º CRTP.



Uma vez com todos os campos da aba "**DADOS**" devidamente preenchidos, a interface irá apresentar dois gráficos relacionados às variáveis especificadas:



a. O primeiro será um gráfico de dispersão, apresentando a variável Y em função da variável X, contendo todas as observações presentes na base de dados;

b. O segundo será um mapa apresentando a configuração dos conjuntos elétricos da CEMIG-D de acordo com o CRTP escolhido. A escala de coloração dos conjuntos elétricos representa o valor da variável resposta definida (Y). Ressalta-se que esse mapa é dinâmico, e mais informações podem ser obtidas caso o usuário clique sobre um determinado conjunto (nome do conjunto, valor de Y, valor de X).

Com os campos devidamente preenchidos, o usuário poderá então iniciar a manipulação da aba "**RESULTADOS**".

Observação: Para o desenvolvimento desse manual, foi utilizado de maneira arbitrária a variável "**log_dec**" como variável resposta (Y) e "**km_ld_s**" como variável preditora (x), somente para visualização dos passos a serem executados.

RESULTADOS

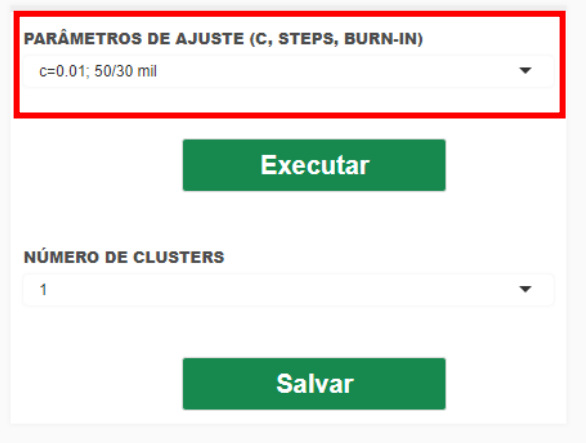
Na aba "**RESULTADOS**", o usuário deverá ajustar parâmetros específicos para o processamento da regionalização da regressão linear simples.

6. O primeiro campo, **"PARÂMETROS DE AJUSTE (C, STEPS, BURN-IN)"**, é responsável pela atribuição de parâmetros técnicos necessários para a execução do algoritmo de regionalização. Neste campo, o usuário precisará definir os valores para:
   a. **c**: parâmetro para ajuste a priori do tamanho dos clusters (caso c = 0.01, o algoritmo favorece a ocorrência de um número maior de clusters; caso c = 0.35, o algoritmo favorece a ocorrência de um número menor de clusters).
   b. **Steps**: define o comprimento da cadeia de Markov implementada.

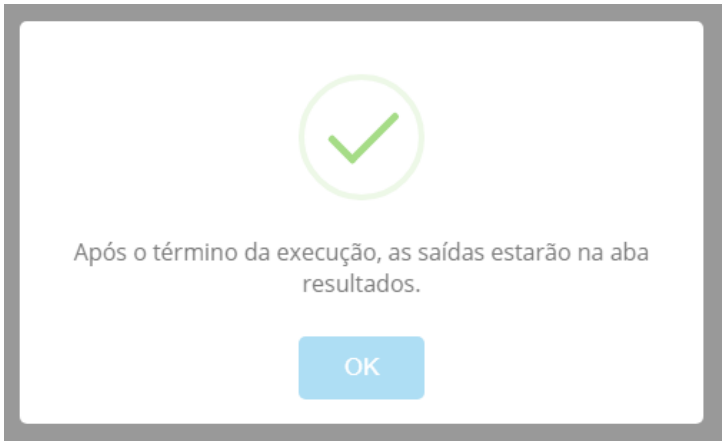
c. **Burn-in**: número de valores no início da cadeia que serão descartados.

Ressalta-se que o parâmetro **Steps** impactará diretamente no tempo necessário para o processamento e execução da análise. Quanto maior é a cadeia de Markov, maior é o tempo de processamento.

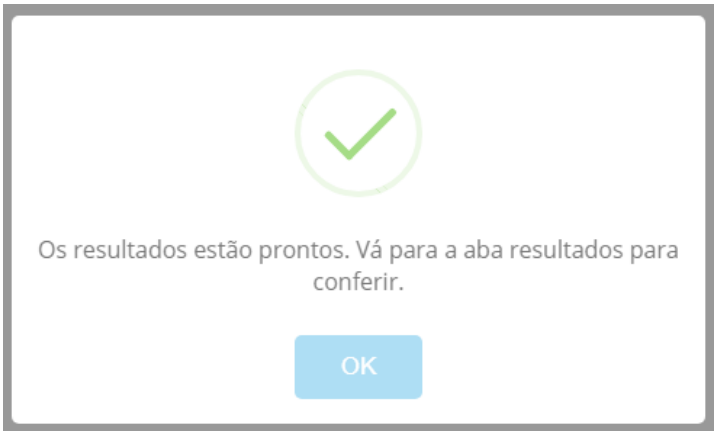
Para simplificar a seleção dos parâmetros **c**, **Steps** e **Burn-in**, configurações pré-definidas estão disponíveis na interface no formato "**c=X.XX; Steps/Burn-in**" e o usuário poderá escolher qual dessas configurações usar. É aconselhado o uso inicial da configuração "c=0.01; 50/30 mil" para que o usuário possa avaliar o tempo computacional necessário para execuções mais longas.

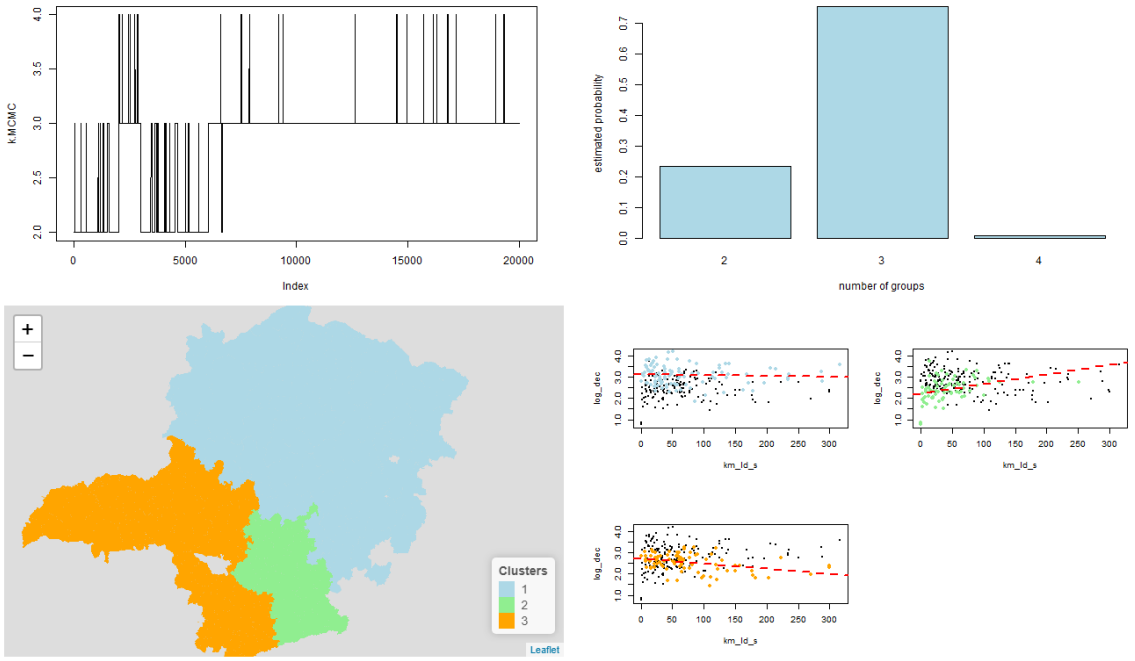


7. Após a definição dos parâmetros no campo "**PARÂMETROS DE AJUSTE**", o usuário deverá então ativar o botão "**EXECUTAR**" para que o processamento da metodologia de regionalização seja iniciado. Uma mensagem de início da execução será apresentada, e o ícones de carregamento aparecerão na própria aba de **RESULTADOS**.



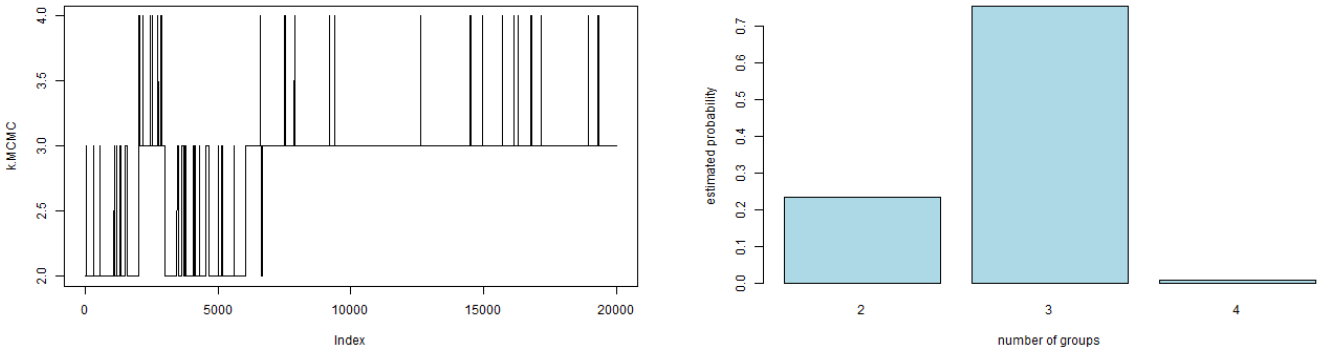
Uma vez terminado o processamento, uma mensagem de finalização também será apresentada, indicando que os resultados foram compilados. Na aba **RESULTADOS**, quatro gráficos serão construídos.





Ressalta-se que os dois gráficos superiores deverão ser analisados primeiramente. Isso se deve ao fato de que eles apresentam resultados referentes ao processamento da regionalização de regressões, enquanto os dois gráficos inferiores apresentam informações sobre a clusterização estimada. Seguem suas descrições:

8. O gráfico **superior esquerdo** indica os resultados do Método de Monte Carlo na Cadeia de Markov. Sucintamente, apresenta o número de clusters nos passos posteriores ao **Burn-in**.
9. O gráfico **superior direito** apresenta a probabilidade estimada do número de clusters. O valor de "number of groups" que apresentar a maior probabilidade estimada é o valor que indicará o número de clusters mais adequado à relação entre X e Y ao longo do território de concessão da CEMIG-D. A título de exemplo, é possível perceber que 3 clusters é a quantidade mais provável de existir para a regressão de "log_dec" em função de "km_ld_s".



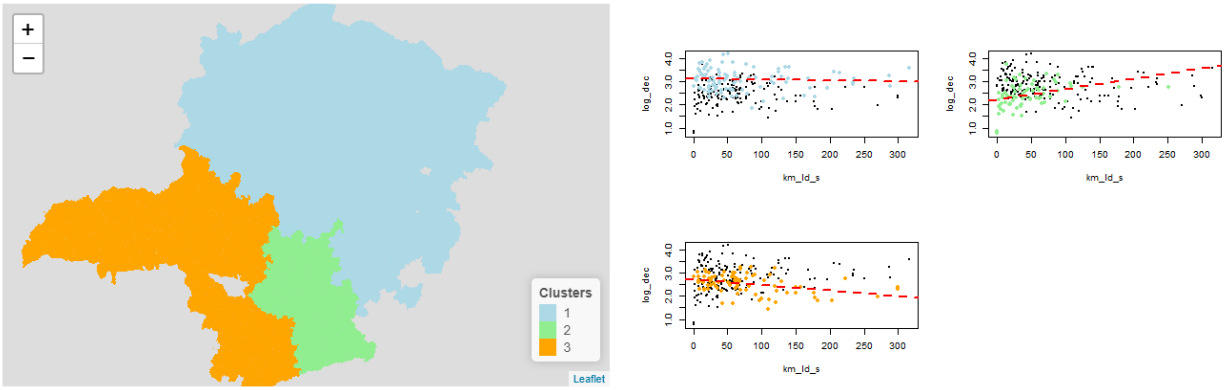
A definição de qual será o número de clusters irá alterar os dois gráficos seguintes (inferiores),e deve ser feita tendo conhecimento de qual é o valor mais adequado. Por esse motivo, os dois gráficos superiores são analisados inicialmente.

Os gráficos inferiores dizem respeito à regionalização dos conjuntos e o resultado dos modelos de regressão linear simples em cada conjunto.

10. O gráfico **inferior esquerdo** apresenta a localização dos clusters. Cada cor indica um agrupamento específico de conjuntos elétricos. Cada agrupamento possui um modelo específico de regressão. O mapa é interativo, de forma que o usuário pode clicar em um determinado conjunto elétrico para identificar em qual cluster ele está localizado.

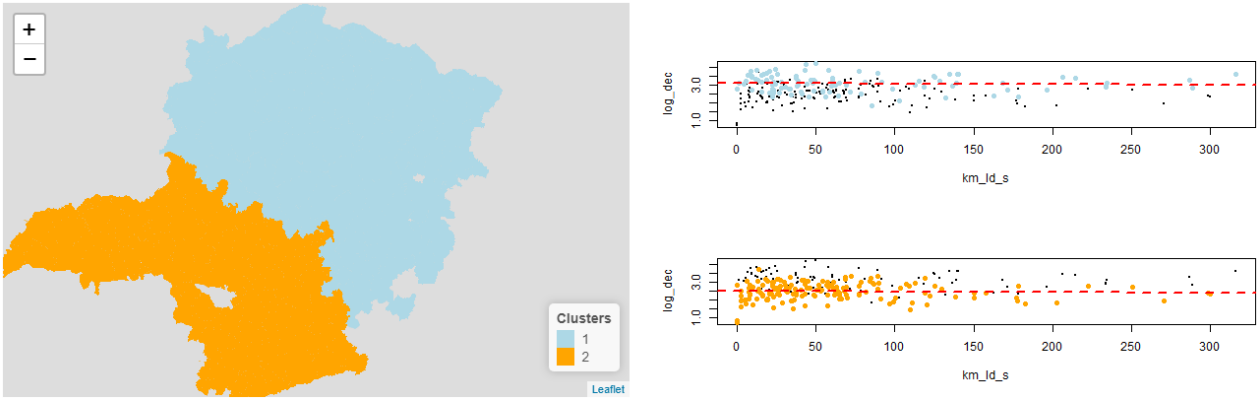
11. O gráfico **inferior direito** é um conjunto de gráficos que apresentam o resultado da regressão linear para cada cluster. A coloração das observações determina qual é o cluster considerado em cada um dos gráficos (se os marcadores forem verdes, o gráfico diz respeito ao cluster com coloração verde no mapa), e a reta vermelha representa a regressão linear levando em consideração somente as observações dos conjuntos elétricos contidos naquele cluster especificamente. Esses gráficos permitem a visualização do modelo de regressão em cada cluster (o comportamento da variável resposta Y em função da variável preditora X).

    Em todos os gráficos, as observações que não são de um determinado cluster estão plotadas como pequenos pontos pretos, facilitando a visualização da distribuição das observações específicas de cada clusters em relação às demais.

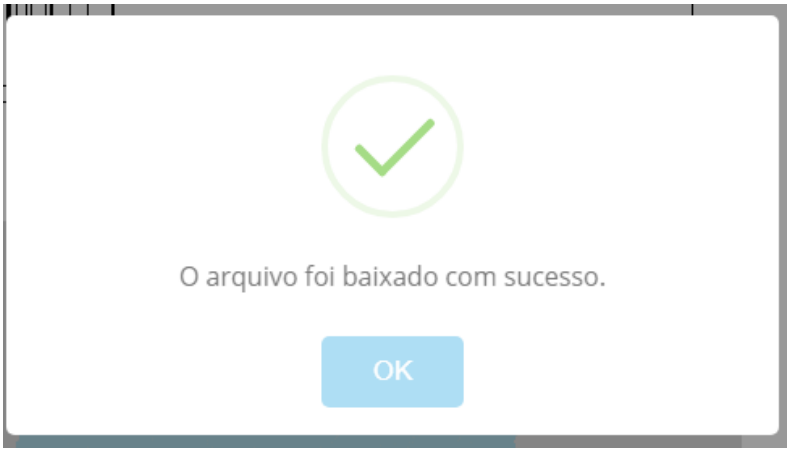


12. Ao término da execução da metodologia de regionalização, o número de clusters será automaticamente alterado para o valor com maior probabilidade estimada. Os gráficos inferiores serão configurados de acordo com esse número de clusters. Entretanto, o usuário poderá alterar o número de clusters usando o campo "**NÚMERO DE CLUSTERS**". Aconselha-se que outros números de clusters que apresentaram um valor razoável de probabilidade estimada também sejam visualizados, para melhor compreensão do comportamento da metodologia de regionalização. Ao alterar o valor em "**NÚMERO DE CLUSTERS**", a nova informação será processada, e os dois gráficos inferiores serão devidamente atualizados.

A título de exemplo, mudando o número de clusters da análise "log_dec" em função de "km_ld_s" para 2, os resultados dos dois gráficos inferiores serão alterados para:



13. Finalmente, uma vez finalizadas as análises, o usuário poderá então salvar a base de dados contendo a informação do cluster no qual cada conjunto elétrico se encontra, utilizando o botão "**SALVAR**". A base de dados de saída apresentará as mesmas informações da base original, importada no início do procedimento, com a adição de uma coluna denominada "grupos" que indicará o cluster alocado a cada um dos conjuntos. Vale lembrar que o usuário, antes de salvar, deverá alterar o valor de "**NÚMERO DE CLUSTERS**" para aquele que lhe for mais interessante.
    Antes de salvar, o usuário poderá visualizar uma tabela contendo código dos conjuntos, as variáveis X e Y e os grupos definidos na aba "**TABELA**", e ao clicar no botão de salvar, uma mensagem de confirmação de download será apresentada.



Segue uma imagem da tabela final exportada, cujo nome padrão é "**geoRegressOutput.xlsx**":

	A	B	AW
1	codigo	dsc_conj_n	grupos
2	15092	ARACAGI	1
3	15093	ABADIA DOS	3
4	15094	ARCOS 1	3
5	15095	AREADO 2	3
6	15096	ABAETE 2	3
7	15097	ALFENAS 1	3
8	15098	AGUAS FORM	1
9	15099	ALPINOPOLI	3
10	15100	ALMENARA	1
11	15101	ANDRADAS 2	3
12	15102	AIMORES	1
13	15103	ARAPORA	3
14	15104	ARAGUARI 2	3
15	15106	AVATINGUAI	3
16	15107	ARAXA 1	3
17	15108	BAMBUI	3
18	15109	BARBACENA	2

MÓDULO DO MODELO HÍBRIDO MULTICAMADAS

OBJETIVO DO APLICATIVO: A interface computacional do modelo híbrido multicamadas foi desenvolvida com o objetivo de construir modelos estatísticos multi-camadas com uma camada linear e uma segunda camada não-linear. É possível ainda criar modelos multi-camadas segmentados por clusters regionalizados dos conjuntos elétricos da CEMIG-D. Por se tratar de uma ferramenta computacional para o ajuste de modelos multi-camadas, sugere-se que a mesma seja utilizada após a construção e adição de todas as variáveis necessárias ao banco de dados. Para o ajuste de modelos multi-camadas regionalizados, os índices dos clusters devem ser previamente estimados utilizando a metodologia de regionalização univariada ou a metodologia de regressão linear simples regionalizada.

INSTRUÇÕES DE USO:

1. Para acessar o módulo do modelo híbrido multicamadas, será necessário chamar a função “shiny_modelo_hib”, como descrito abaixo:

   ```
   > shiny_modelo_hib()
   ```

2. Após a execução da função, o aplicativo Shiny com o módulo do modelo híbrido multicamadas será aberto automaticamente na aba “**DADOS**”, como apresentado nas imagens abaixo:

DADOS

3. Para configurar uma base de dados externa que possa ser utilizada pela interface é importante que a mesma siga a seguinte estrutura:
   - O arquivo deve estar em formato .xlsx (Excel);
   - É necessário que exista uma coluna contendo a informação geográfica de clusters para cada conjunto elétrico já realizada em outra interface (caso a base

de dados não possua divisão em grupos crie uma coluna contendo somente o número 1);

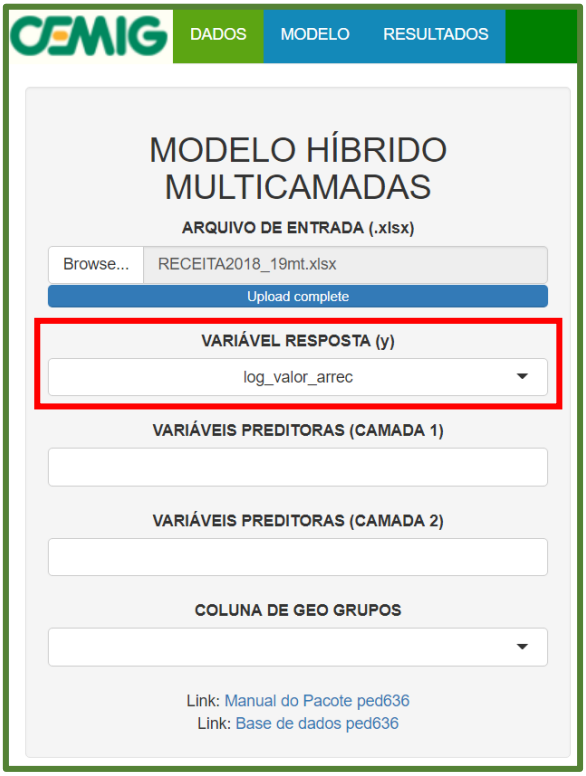
- O usuário precisa ter conhecimento de qual configuração de conjuntos elétricos a base está relacionada (3º ou 4º CRTP).

Vale ressaltar também que é fundamental que o formato das variáveis resposta e preditoras sejam numéricas para que não haja problemas na execução do modelo.

4. Com a base de dados em formato Excel estruturada corretamente, basta clicar em "**BROWSE**" no campo "**ARQUIVO DE ENTRADA**" e selecionar o arquivo a ser utilizado, como mostra a figura a seguir. Feito isso, a base de dados será carregada para a interface.

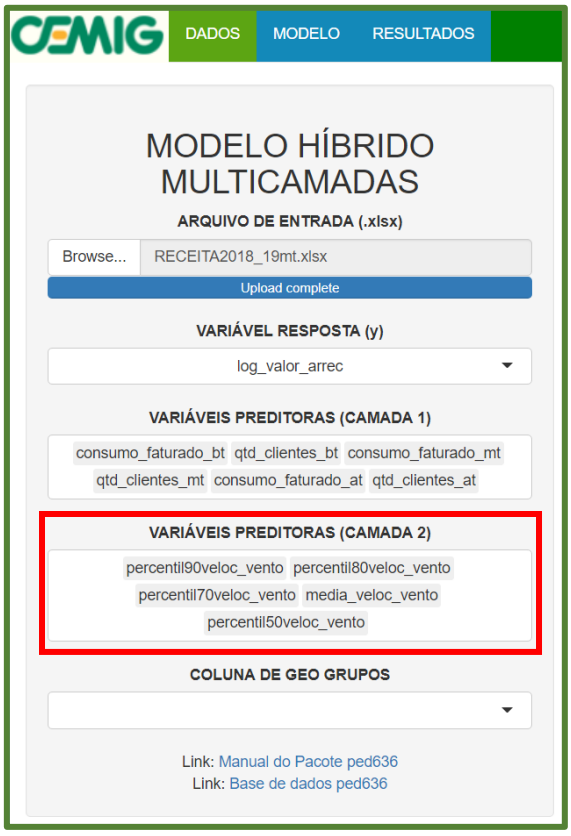
5. Quando a base de dados é carregada as informações são apresentadas em um formato de tabela, como mostrado a seguir.

6. No campo "**VARIÁVEL RESPOSTA (y)**" dever ser selecionada a variável resposta do modelo. Como valor padrão, será selecionada a primeira varável da base de dados.

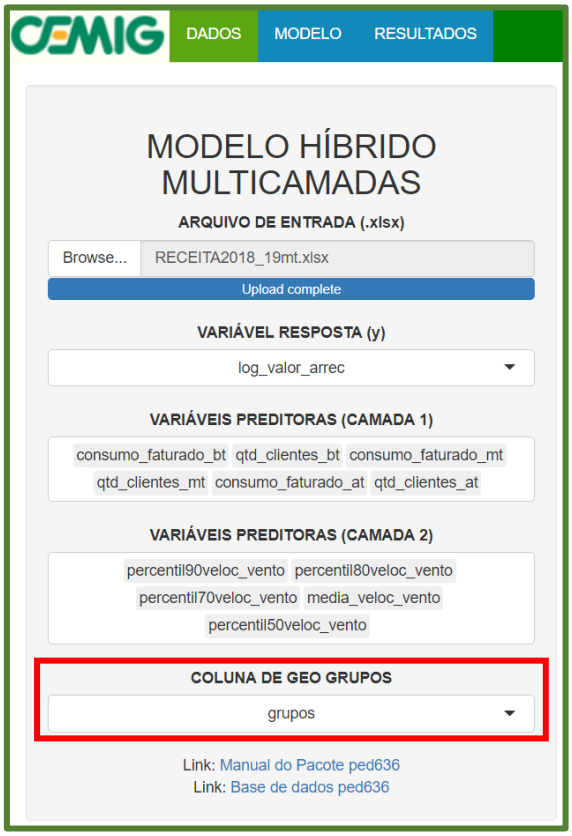


7. Em seguida devem ser selecionadas as variáveis da primeira camada (linear) do modelo multi-camadas no campo "**VARIÁVEIS PREDITORAS (CAMADA 1)**". O usuário pode escolher uma ou mais variáveis. Atentar para não selecionar a variável resposta nessa etapa.

8. Em seguida o usuário deve selecionar as variáveis da segunda camada (não-linear) do modelo no campo "**VARIÁVEIS PREDITORAS (CAMADA 2)**". O usuário pode escolher uma ou mais variáveis. O usuário pode selecionar variáveis que já foram escolhidas na primeira camada. Atentar para não selecionar a variável resposta nessa etapa.



9. Por fim, o usuário deve selecionar a variável que contém a informação geográfica (clusters) dos conjuntos elétricos no campo "**COLUNA DE GEO GRUPOS**". Caso não exista tal variável, deve ser criada uma variável na base de dados contendo somente o número 1 para cada amostra.



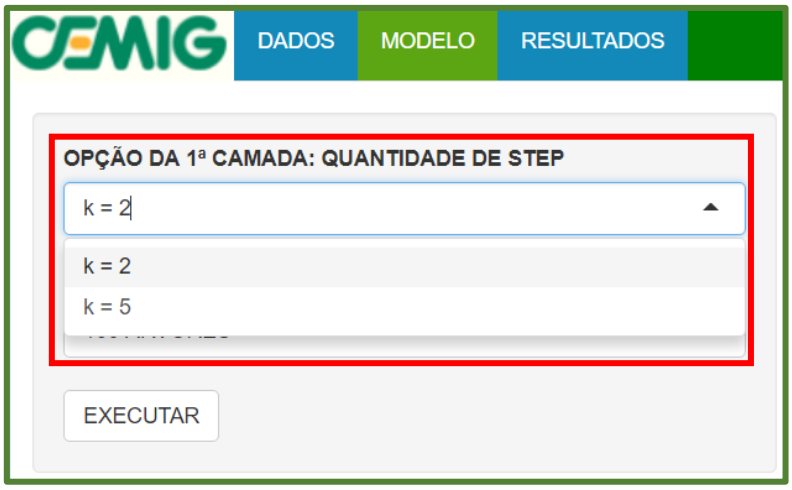
10. Com todos as opções selecionadas, o usuário pode prosseguir para a aba "**MODELO**".

MODELO

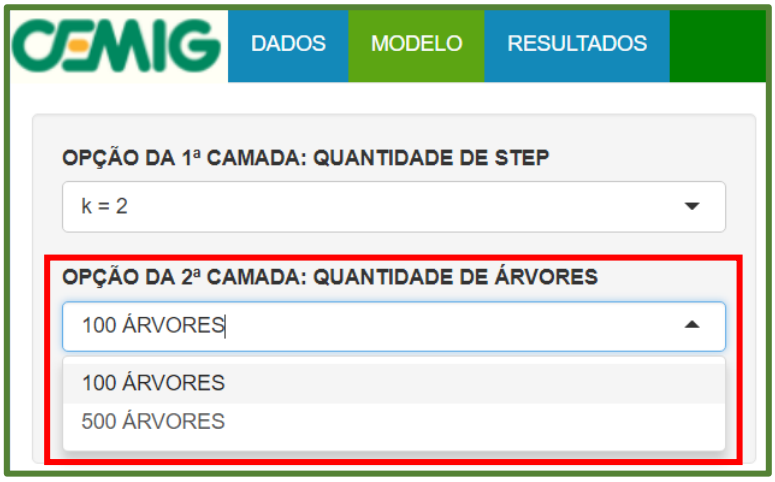
Na aba "**MODELO**", o usuário configura os modelos de primeira e segunda camada. De forma sucinta, o usuário deve fazer uma seleção para a primeira camada (parâmetro k da função step) e para a segunda camada (número de árvores). O modelo multi-camadas disponível contém o modelo de regressão linear múltipla na primeira camada e o modelos de árvores de regressão (random forests) na segunda camada.

Na aba "**MODELO**", são apresentados dois campos de seleção para o usuário.

11. No primeiro campo "**OPÇÃO DA 1ª CAMADA: QUANTIDADE DE STEP**" o usuário deve definir qual o valor do parâmetro k da função step que é aplicada na primeira camada linear do modelo. O parâmetro k ajusta a intensidade da penalização do modelo de regressão linear múltipla. A opção "**k = 2**" apresenta uma penalização mais flexível permitindo um número maior de variáveis no modelo final. A opção "**k = 5**" é mais restritiva e, em geral, seleciona um número menor de variáveis preditoras.



12. No segundo campo, "**OPÇÃO DA 2ª CAMADA: QUANTIDADE DE ÁRVORES**", o usuário deve definir quantas árvores de regressão serão criadas na função **randomforest,** referente ao modelo não-linear na segunda camada. A quantidade de árvores de regressão do modelo impacta na avaliação das variáveis preditoras e no tempo de execução. A opção padrão "**100 árvores**" é mais rápida, mas pode ser inviável para um número grande de variáveis preditoras. A opção "**500 árvores**" é mais demorada mas, em geral, garante que todas as variáveis preditoras sejam consideradas no modelo.



13. Após fazer a seleção dos dois parâmetros do modelo híbrido multi-camadas o usuário deve pressionar o botão "**EXECUTAR**".

14. O usuário pode aguardar o término da execução do modelo na aba "**MODELO**" até a mensagem de término aparecer ou pode imediatamente avançar para a terceira aba "**RESULTADOS**".

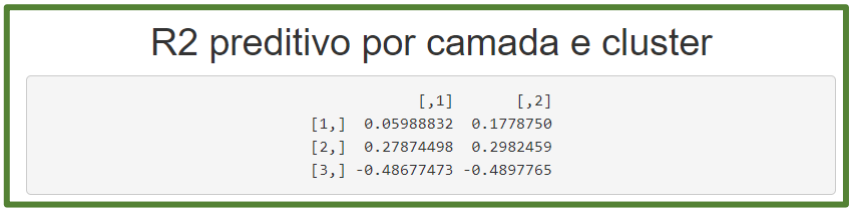
RESULTADOS

Na aba "**RESULTADOS**" o usuário pode visualizar os resultados do modelo híbrido. A depender do número de variáveis preditoras selecionadas (principalmente na segunda camada), a execução do modelo pode durar entre 1 e 2 minutos.

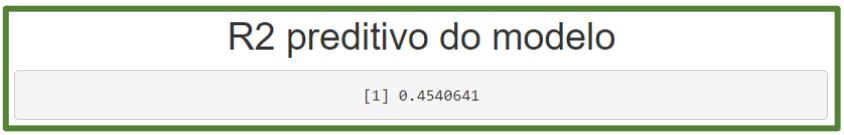
15. A terceira aba apresenta seis informações que resumem o modelo ajustado: (1) o coeficiente de determinação preditivo (R^2) por camada e cluster; (2) o coeficiente de determinação preditivo (R^2) do modelo completo; (3) a matriz de pesos dos modelos lineares e árvores de regressão na composição final; (4) os modelos lineares ajustados em cada cluster; (5) os modelos não-lineares ajustados em cada cluster; e a (6) base de dados.

16. Após a execução do modelo, os resultados aparecerão embaixo de suas respectivas descrições.

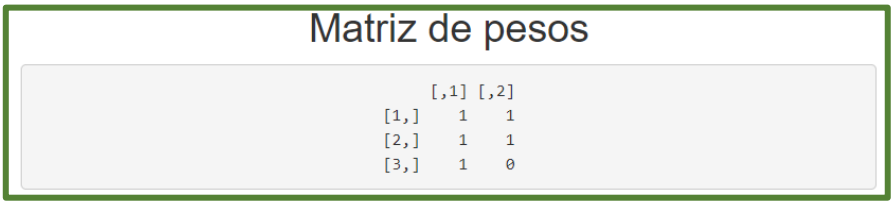
17. O campo "**R^2 preditivo por camada e cluster**" apresenta o R^2 preditivo para cada uma das duas camadas (coluna) e para cada cluster (linhas, no exemplo, 3 clusters). Importante frisar que o valor da segunda camada (segunda coluna) é cumulativo. No exemplo apresentado, para o segundo cluster a primeira camada apresentou um R^2 preditivo igual a 27,87%, enquanto a segunda camada apresentou um R^2 preditivo igual a 29,82%, ou seja, um ganho de aproximadamente 2%.



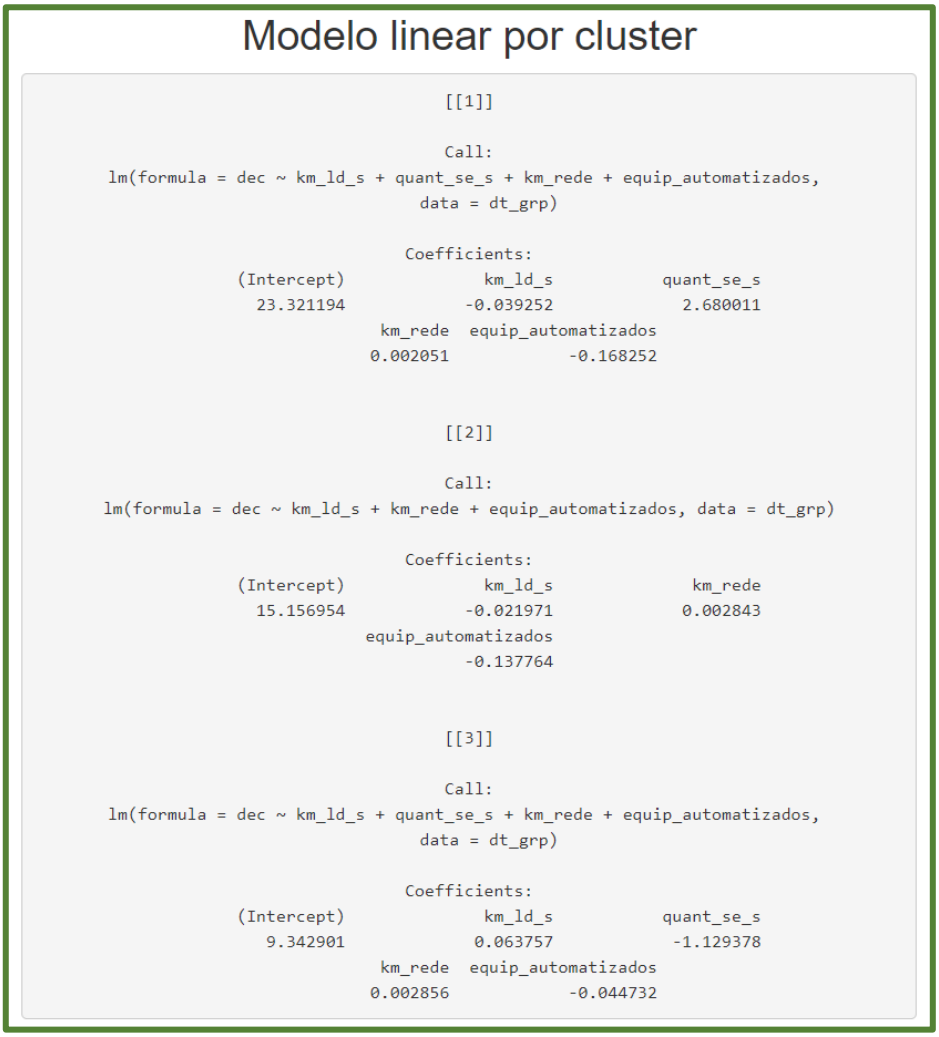
18. O campo "**R^2 preditivo do modelo**" apresenta o R^2 preditivo para o modelo considerando todas as camadas e todos os clusters, ou seja, um resumo do modelo.



19. O campo "**Matriz de pesos**" indica se o uso da segunda camada foi efetivo para cada cluster. Caso a segunda camada seja utilizada, a primeira e segunda colunas da linha (cluster) possuem o número 1 (exemplo clusters 1 e 2); caso o uso da segunda camada não seja necessário, a matriz apresentará o número 1 na primeira coluna e 0 na segunda (exemplo cluster 3).



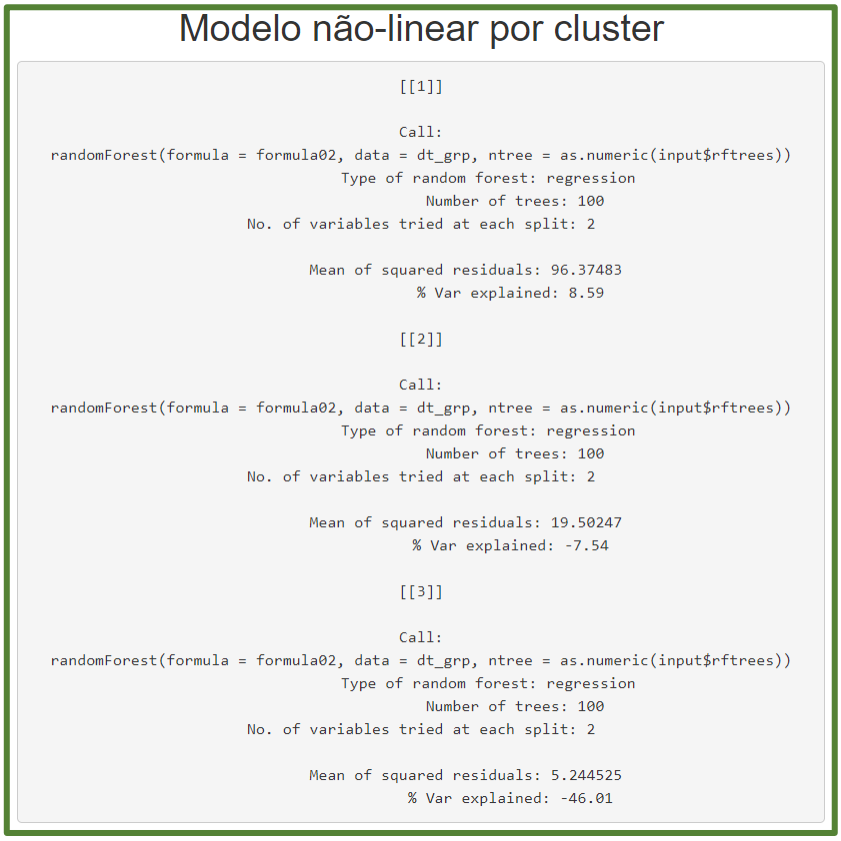
20. O campo "**Modelo linear por cluster**" apresenta as variáveis significativas e os coeficientes do modelo linear para cada cluster.



Modelo linear por cluster

```
[[1]]

Call:
lm(formula = dec ~ km_ld_s + quant_se_s + km_rede + equip_automatizados,
    data = dt_grp)

Coefficients:
       (Intercept)             km_ld_s          quant_se_s
         23.321194          -0.039252            2.680011
         km_rede  equip_automatizados
         0.002051           -0.168252


[[2]]

Call:
lm(formula = dec ~ km_ld_s + km_rede + equip_automatizados, data = dt_grp)

Coefficients:
       (Intercept)             km_ld_s              km_rede
         15.156954          -0.021971             0.002843
equip_automatizados
         -0.137764


[[3]]

Call:
lm(formula = dec ~ km_ld_s + quant_se_s + km_rede + equip_automatizados,
    data = dt_grp)

Coefficients:
       (Intercept)             km_ld_s          quant_se_s
          9.342901           0.063757           -1.129378
         km_rede  equip_automatizados
         0.002856           -0.044732
```

21. O campo “**Modelo não-linear por cluster**” apresenta as informações referentes a segunda camada não-linear (random forest) para cada cluster.



22. O campo “**Base de dados**” apresenta a base de dados que foi utilizada para construir o modelo.

Base de dados

	dsc_conj_n	codigo	grupos3	km_ld_s	quant_se_s	km_rede
1	ARACAGI	15092	1	4.337599	1	1035.3276
2	ABAETE 2	15096	1	56.941862	1	1446.8924
3	AGUAS FORMOSAS	15098	1	44.623968	1	2810.2069
4	ALMENARA	15100	1	41.048479	2	3619.1449
5	AIMORES	15102	1	11.968503	1	1756.0611
6	BARBACENA 2	15109	1	72.146603	0	4235.1184
7	BARAO DE COCAIS 1	15111	1	15.786787	1	987.9915
8	BOCAIUVA	15112	1	46.553446	1	3805.5163
9	BURITIS 2	15132	1	42.987898	1	1398.6023
10	BURITIS 1	15133	1	50.191697	1	1881.8900
11	BRASILIA DE MINAS	15134	1	115.314311	1	2835.1827
12	BRASILANDIA 2	15139	1	315.870073	7	8748.1102
13	BERILO	15140	1	33.023487	1	1811.1781
14	CARLOS CHAGAS 2	15144	1	47.862194	1	1511.8932
15	CARLOS CHAGAS 1	15145	1	48.902481	1	2346.0718
16	CAETE 1	15149	1	25.753816	1	683.6642
17	CARATINGA 1	15150	1	98.351867	2	3384.5470
18	CI SANTA LUZIA	15153	1	33.505031	2	519.3684
19	CORINTO 1	15154	1	139.047406	1	2420.8049
20	CAPELINHA 1	15155	1	91.596493	2	4932.1745

	equip_protecao	equip_automatizados	total_clientes	area_km_quad	estradas_km
1	341	8	7001	1263.4393	645.4
2	471	18	12218	1901.6853	1112.5
3	912	24	20714	3449.8286	4746.4
4	1122	27	29983	4945.3867	8735.3
5	760	12	13409	1957.8772	2561.4
6	1542	52	35189	2886.8678	11843.0
7	464	18	18929	698.2197	1195.7
8	1326	39	30494	7133.3808	5481.6
9	471	14	3641	3615.0763	557.7
10	557	13	10489	5386.5070	2363.1
11	996	25	21899	3113.1640	6181.4
12	2321	93	32533	22675.9787	20788.7
13	514	15	11386	1618.7952	3161.9
14	508	9	7570	2254.4578	2567.2
15	772	9	13086	3653.9794	3383.3

23. Após conferir os resultados do modelo, o usuário pode salvar todas as informações em um arquivo em formato ".RData" ao pressionar o botão "**SALVAR MODELO**". Além de permitir ao usuário salvar o modelo para uso posterior, esse formato de arquivo também foi escolhido para transferir todas as informações necessárias para a interface do Simulador/Otimizador. O usuário também pode realizar o download da base de dados utilizada com o valor da variável resposta calculada pelo modelo em formato ".xlsx" ao pressionar o botão "**SALVAR EXCEL**".


CEMIG
DADOS
MODELO
RESULTADOS
SALVAR MODELO
SALVAR EXCEL
CEMIG P&D ANEEL
R2 PREDITIVO POR CAMADA E CLUSTER
[,1] [,2]
[1,] -5.3118278 -5.5242728
[2,] -0.1406527 -0.2787434
[3,] -0.2197946 -0.3544114
[4,] -0.2316543 -0.5784694
R2 PREDITIVO DO MODELO
[1] -1.520501
MATRIZ DE PESOS
[,1] [,2]
[1,] 1 0
[2,] 1 0
[3,] 1 0
[4,] 1 0
MODELO LINEAR POR CLUSTER
[[1]]
Call:
lm(formula = log_valor_arrec ~ consumo_faturado_bt + consumo_faturado_at,

MÓDULO DO SIMULADOR E OTIMIZADOR

OBJETIVO DO APLICATIVO: A interface computacional do simulador e otimizador foi desenvolvida com o objetivo de permitir ao analista fazer análises exploratórias de possíveis cenários com base nos modelos estatísticos-computacionais desenvolvidos na interface de modelos híbridos. Por se tratar de uma ferramenta computacional para análise de modelos já ajustados, sugere-se que a mesma seja utilizada somente após o desenvolvimento do modelo estatístico-computacional desejado.

INSTRUÇÕES DE USO:

- Para acessar o módulo do simulador e otimizador, será necessário chamar a função "shiny_simul_otim", como descrito abaixo:

```
> shiny_simul_otim()
```

- Após a execução da função, o aplicativo Shiny com o módulo do simulador e otimizador será aberto automaticamente na aba "**DADOS**", como apresentado na figura a seguir.

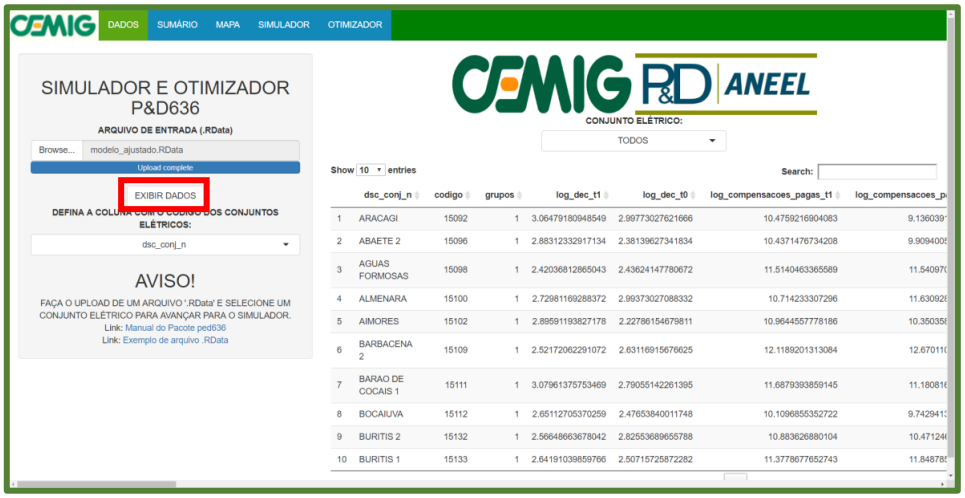
- Para configurar um arquivo de entrada, a ser utilizado pela interface, é importante que o mesmo siga a estrutura definida na interface do modelo híbrido multicamadas:
  - O arquivo deve estar em formato .RData;
  - O arquivo deve conter todas as informações geradas na interface do modelo híbrido multicamadas: R^2 preditivo por camada e cluster; R^2 preditivo do modelo; Matriz de pesos; Modelo linear por cluster; Modelo não-linear por cluster; Base de dados;
  - O usuário precisa ter conhecimento de qual configuração de conjuntos elétricos a base está relacionada (3º ou 4º CRTP).

- Com a base de dados em formato ".RData" estruturada corretamente, basta clicar em "**BROWSE**" no campo "**ARQUIVO DE ENTRADA**" e selecionar o arquivo a ser utilizado,

como mostra a figura a seguir. Feito isso, a base de dados será carregada para a interface.

- Quando a base de dados é carregada surge uma barra azul com a informação "**Upload complete**".

- Para avançar o usuário deve pressionar o botão "**EXIBIR DADOS**". Ao selecionar o botão, a base de dados utilizada na construção do modelo híbrido será apresentada.



- Em seguida o usuário deve indicar no campo "**DEFINA A COLUNA COM O CÓDIGO DOS CONJUNTOS ELÉTRICOS:"** qual das colunas contém a informação com o código dos conjuntos elétricos. Ao carregar os dados, por padrão, a interface seleciona a primeira coluna da base de dados.

	dsc_conj_n	codigo	grupos	log_dec_t1	log_dec_t0	log_compensacoes_pagas_t1	log_compensacoes_p
1	ARACAGI	15092	1	3.06479180948549	2.99773027621666	10.4759216904083	9.136039
2	ABAETE 2	15096	1	2.88312332917134	2.38139627341834	10.4371476734208	9.909400
3	AGUAS FORMOSAS	15098	1	2.42036812865043	2.43624147780672	11.5140463365589	11.54097
4	ALMENARA	15100	1	2.72981169288372	2.99373027088332	10.714233307296	11.63092
5	AIMORES	15102	1	2.89591193827178	2.22786154679811	10.9644557778186	10.35035
6	BARBACENA 2	15109	1	2.52172062291072	2.63116915676625	12.1189201313084	12.67011
7	BARAO DE COCAIS 1	15111	1	3.07961375753469	2.79055142261395	11.6879393859145	11.18081
8	BOCAIUVA	15112	1	2.65112705370259	2.47653840011748	10.1096855352722	9.742941
9	BURITIS 2	15132	1	2.56648663678042	2.82553689655788	10.883626880104	10.47124
10	BURITIS 1	15133	1	2.64191039859766	2.50715725872282	11.3778677652743	11.84878

- Após realizar esses passos, o usuário já tem acesso às abas "**SUMÁRIO**" e "**MAPA**". As duas abas permitem fazer uma análise exploratória da base de dados que foi carregada na interface. Caso o usuário tente avançar para as abas sem realizar as etapas previamente necessárias, uma mensagem de erro será apresentada.

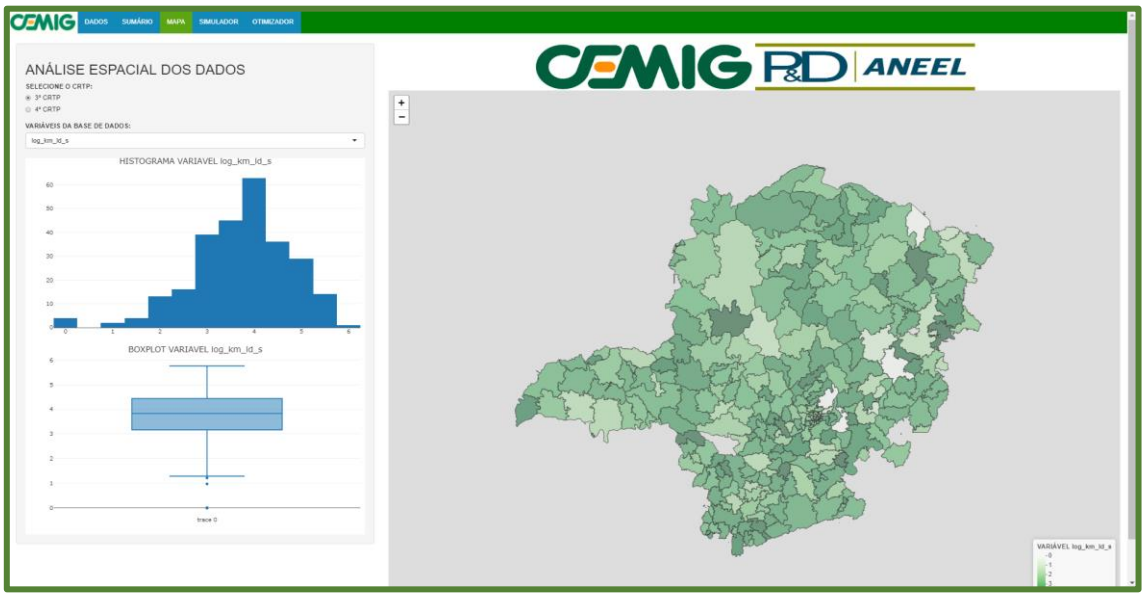
SUMÁRIO

Na aba "**SUMÁRIO**" o usuário tem acesso à uma visualização dos dados por cluster tal qual exibido na última aba da interface de modelos híbridos. A primeira informação disponibilizada é a variável que foi utilizada na base de dados para fazer a regionalização dos dados. O usuário pode alterar o cluster usando o menu de botões no campo "**SELECIONE O CLUSTER:**" no menu lateral na esquerda. As informações apresentadas são: R^2 preditivo do modelo; R^2 preditivo do cluster 1ª camada; R^2 preditivo do cluster 2ª camada; Matriz de pesos por camada e cluster; Modelo linear por cluster; Modelo não-linear por cluster.

MAPA

Na aba "**MAPA**" o usuário tem acesso à uma análise espacial da base de dados. O usuário deve selecionar o CRTP referente aos dados (3º ou 4º) no campo "**SELECIONE O CRTP:**" e a variável que deseja visualizar no campo "**VARIÁVEIS DA BASE DE DADOS**" no menu lateral à esquerda.

- Para visualizar os dados o usuário deve selecionar o CRTP referente à base de dados carregada e selecionar uma variável **NUMÉRICA** da base de dados. Variáveis com texto não geram mapas nem gráficos.



- Após realizar as seleções adequadas, a interface exibe o histograma e boxplot da variável selecionada assim com o mapa dos conjuntos elétricos coloridos em uma escala

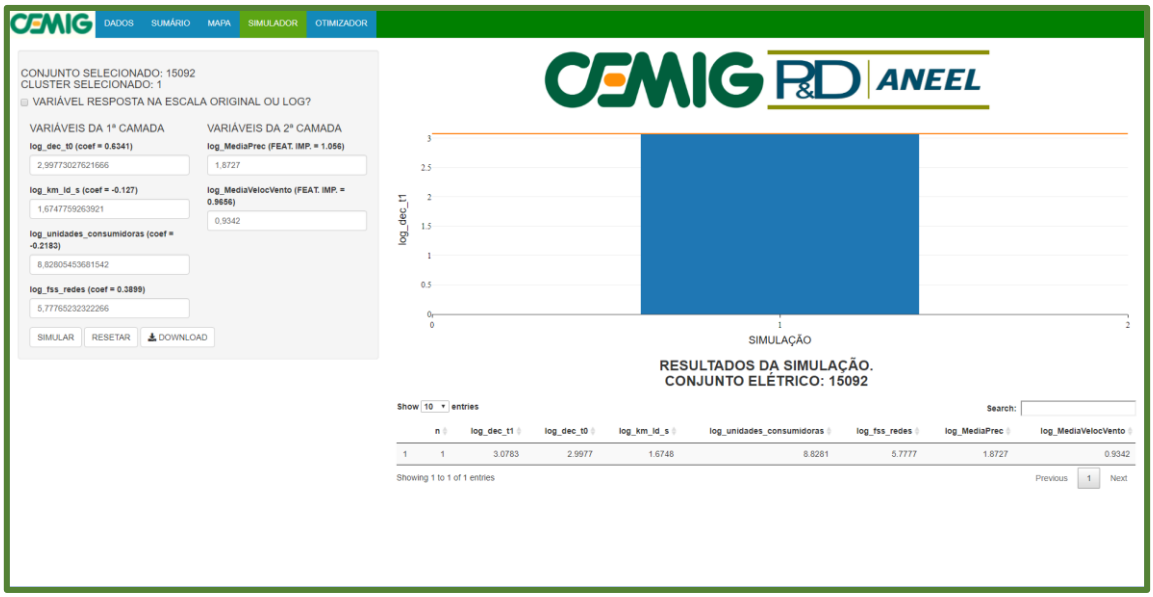
de cor variando de amarelo (para valores pequenos) para vermelho (para valores grandes). Todos os gráficos e mapas são interativos, o usuário pode passar o mouse em cima dos gráficos para verificar os valores ou o nome do conjunto elétrico e o seu valor (no mapa).

SIMULADOR

O usuário deve selecionar um conjunto elétrico na aba "**DADOS**" para acessar as abas "**SIMULADOR**" e "**OTIMIZADOR**". Ao selecionar um conjunto elétrico, ele ficará destacado em cor cinza. O objetivo do simulador é que o usuário altere as variáveis preditoras da 1ª e 2ª camadas e observe o seu impacto na variável resposta. O usuário pode fazer repetidas simulações e posteriormente guardar os resultados em uma planilha em formato Excel.

- Caso o usuário tente avançar para a aba "**SIMULADOR**" ou "**OTIMIZADOR**" sem selecionar um conjunto elétrico será apresentada uma mensagem de erro.

- A aba "**SIMULADOR**" inicia com algumas informações preenchidas para facilitar o uso. Ela se divide em duas partes: no menu à esquerda o usuário pode identificar o conjunto elétrico selecionado, o cluster a que ele pertence, modificar valores e executar as simulações; à direita o usuário tem acesso aos resultados das simulações.

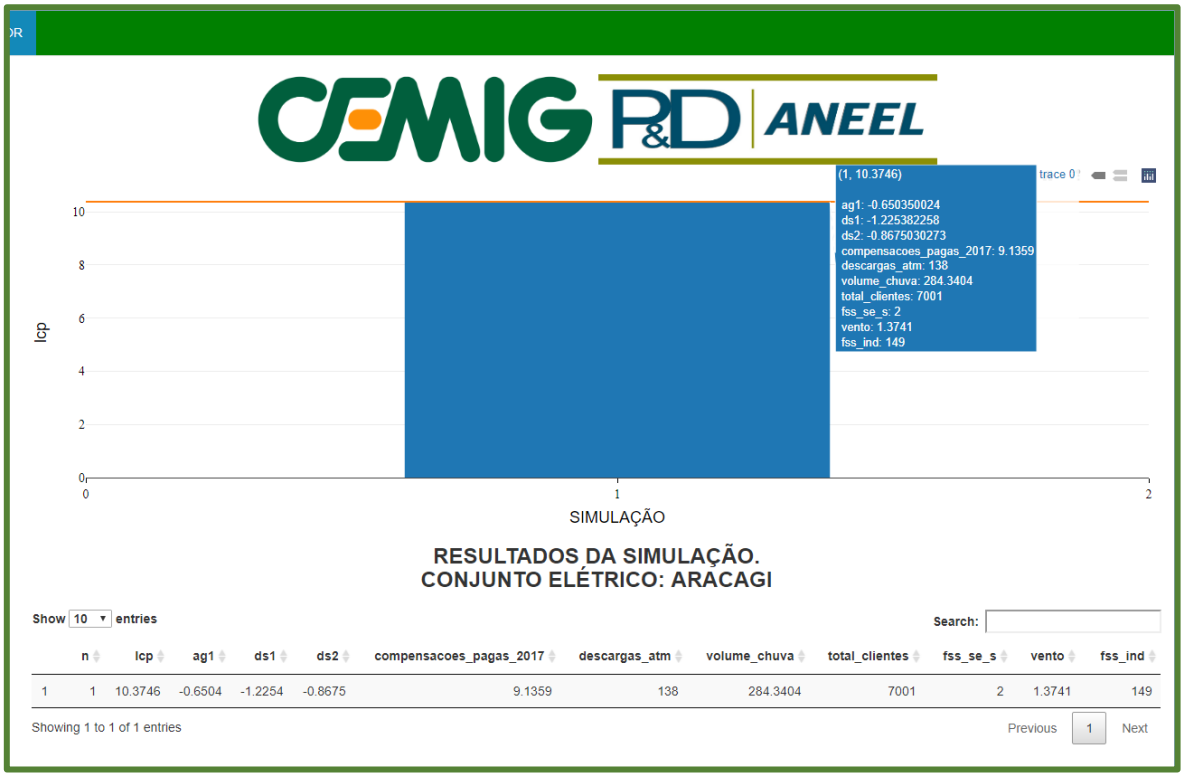


- A primeira opção do usuário é a caixa de seleção "**VARIÁVEL RESPOSTA NA ESCALA ORIGINAL OU LOG?**". Caso a variável resposta no simulador esteja na escala log e o usuário deseje visualizar os valores na escala padrão, pode selecionar a caixa e o valor convertido será apresentado entre parênteses.

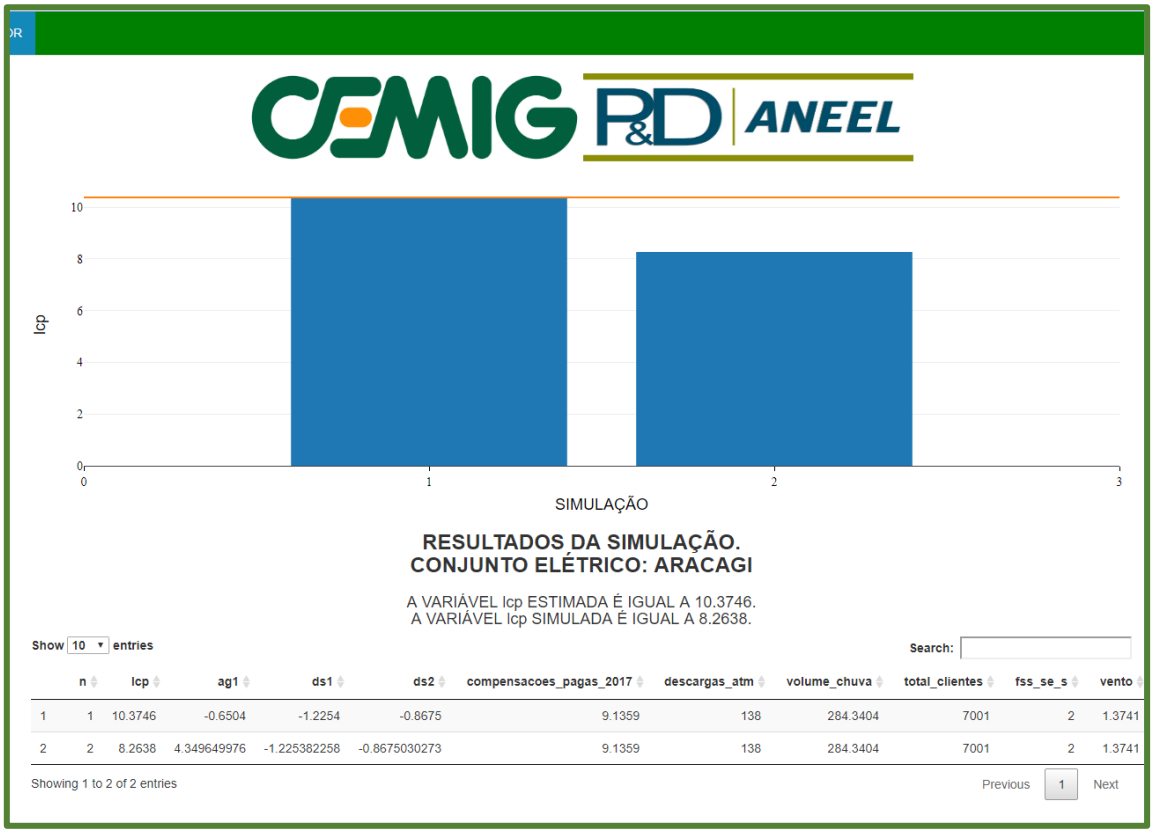
- Em seguida o usuário é apresentado às variáveis preditoras significativas da 1ª e 2ª camada do modelo híbrido. Para as variáveis da 1 ª camada é apresentado entre parênteses o valor do seu coeficiente; para as variáveis da 2ª camada é apresentado o valor do seu *feature importance*. As variáveis apresentadas nessa aba mudam de acordo com o cluster do conjunto elétrico selecionado pelo usuário na aba "**DADOS**". A caixa de seleção de cada variável inicia com o valor atual do conjunto elétrico selecionado.

- Para realizar as simulações, o usuário deve alterar os valores das variáveis de interesse e pressionar o botão "**SIMULAR**". Os resultados da simulação são apresentados à direita da interface e serão descritos a seguir. Além desse botão o usuário tem acesso a mais dois: o botão "**RESETAR**" que reinicia a interface; o botão "**DOWNLOAD**" no qual o usuário pode fazer o download dos resultados das simulações em formato Excel.

- No início das simulações, à direita, o usuário visualiza um gráfico de barras que inicia com uma barra e uma linha horizontal vermelha. Ambos, representam o valor estimado pelo modelo híbrido para a variável resposta no conjunto elétrico selecionado. O gráfico é interativo, o usuário pode posicionar o cursor em cima da coluna para visualizar o valor da variável resposta estimada e os valores das variáveis preditoras que geraram a coluna.

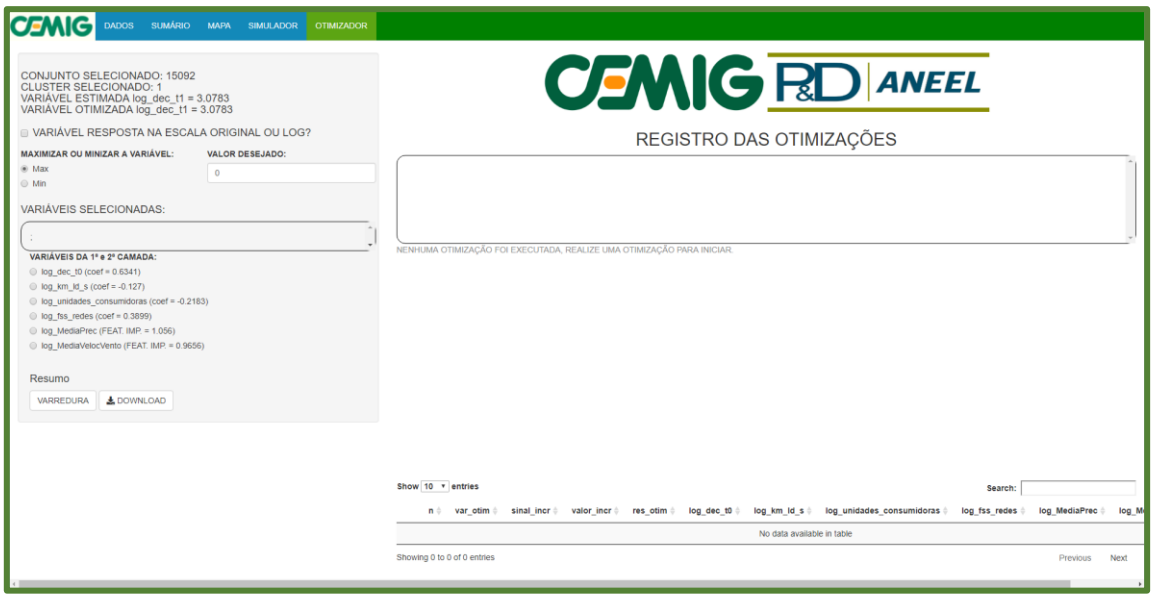


- Ao apertar o botão “**SIMULAR**” são realizadas três alterações na interface:
  - Surge uma nova barra com altura proporcional ao novo valor simulado para a variável resposta.
  - O valor da variável resposta estimado originalmente pelo modelo e o valor simulado são exibidos abaixo do gráfico.
  - Mais abaixo, surge uma nova linha no registro de simulações guardando os valores de todas as variáveis preditoras que geraram a simulação (O registro de simulações inicia com os valores iniciais do conjunto elétrico simulado. Ao apertar o botão “**DOWNLOAD**” essa tabela é baixada no computador do usuário em formato Excel).

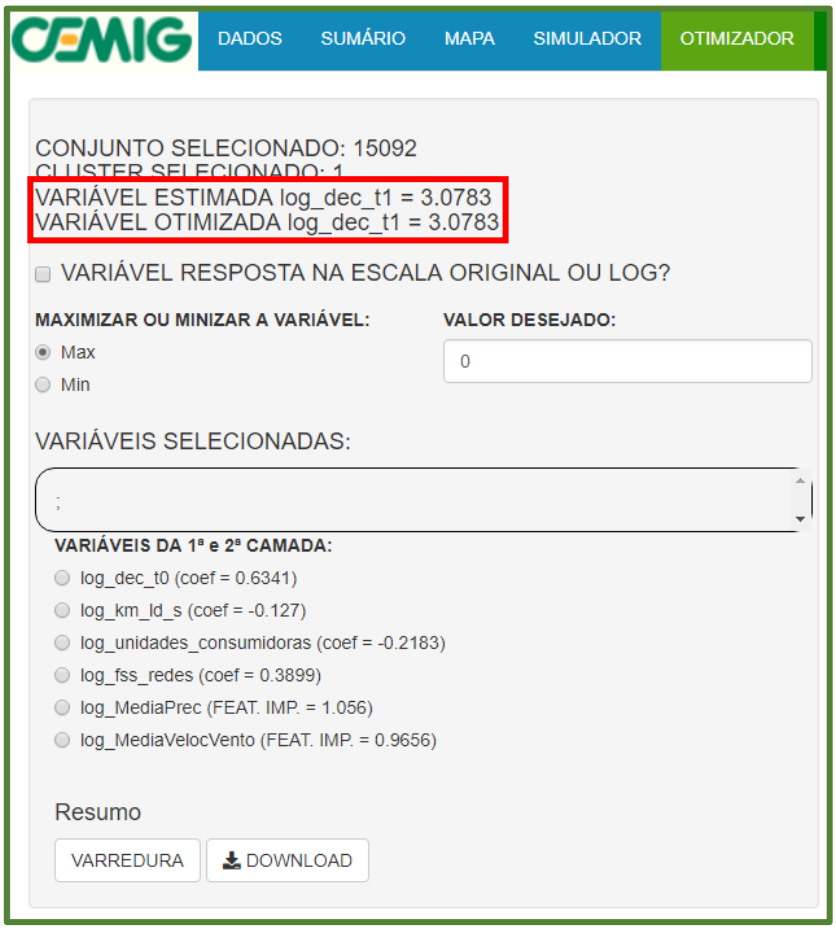


OTIMIZADOR

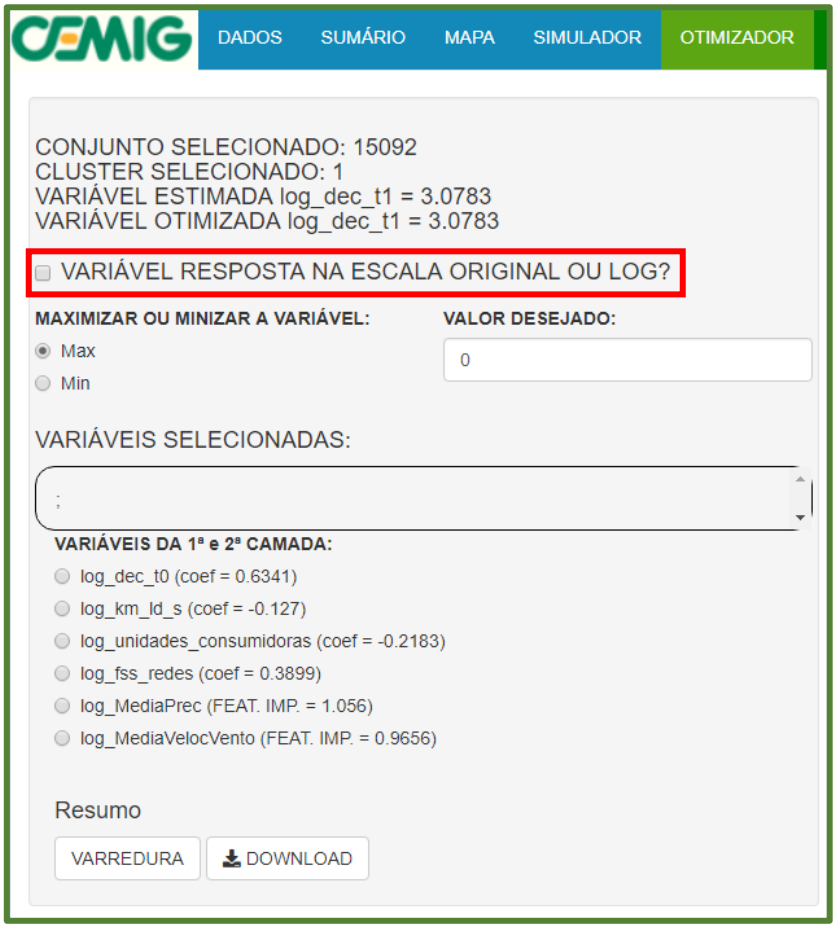
Na aba "**OTIMIZADOR**" o usuário pode aplicar o algoritmo de otimização (*grid search*-adaptado) para buscar a combinação de variáveis preditoras que se aproximam o máximo possível do valor objetivo definido para a variável resposta. Para tanto, o usuário deve preencher uma sequência de campos que serão descritos a seguir. O menu lateral à esquerda contém o conjunto elétrico selecionado, o seu cluster e os campos que o usuário deve preencher; à direita são exibidos os resultados das otimizações.



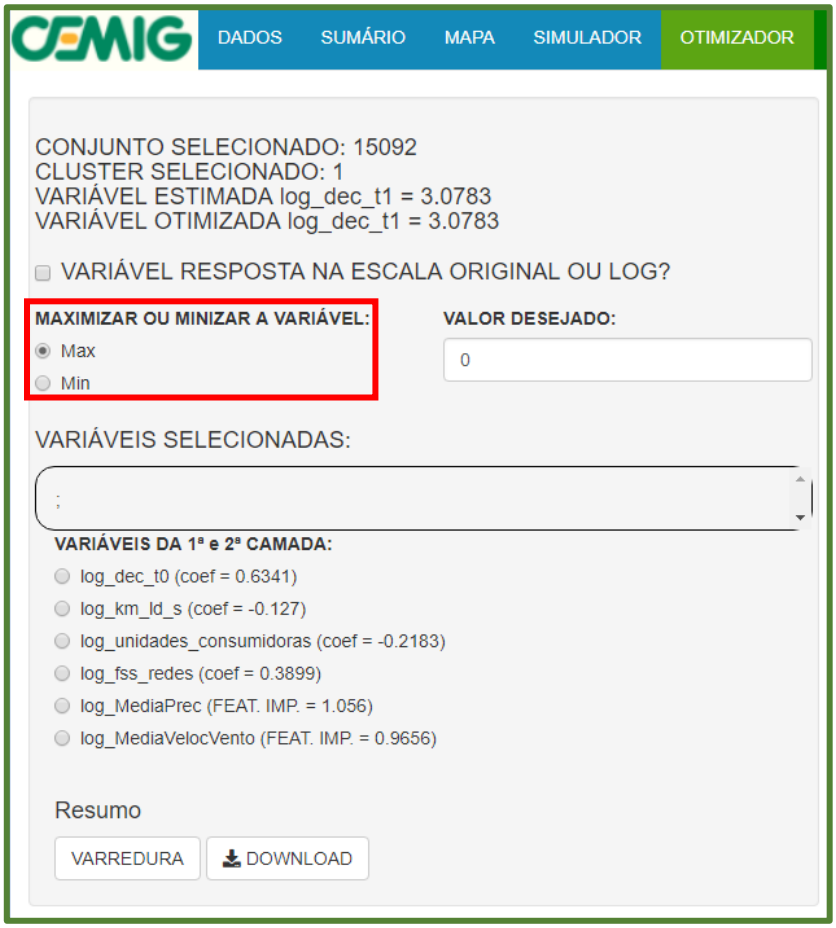
- No topo do menu lateral (à esquerda) são exibidos o valor estimado pelo modelo híbrido da variável resposta e o último valor da otimização (ao abrir o otimizador os valores apresentados são iguais).



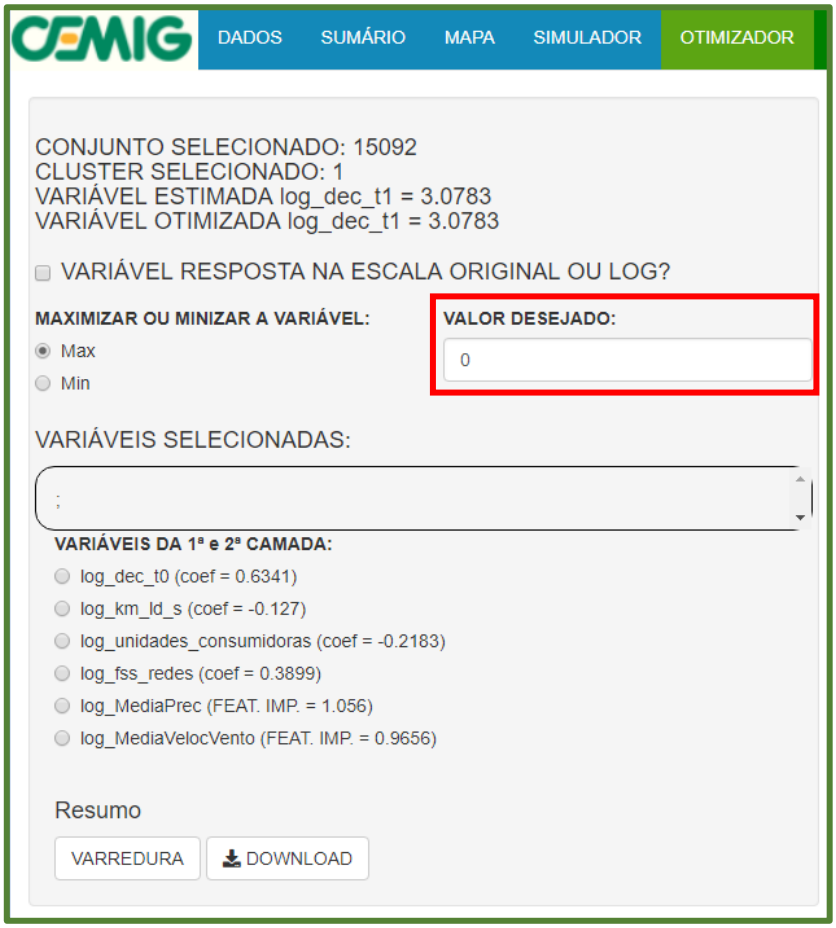
- Tal qual no simulador, caso a variável resposta esteja na escala log o usuário tem a opção de visualizar o valor na escala original ao selecionar a caixa de seleção "**VARIÁVEL RESPOSTA NA ESCALA ORIGNAL OU LOG?**".



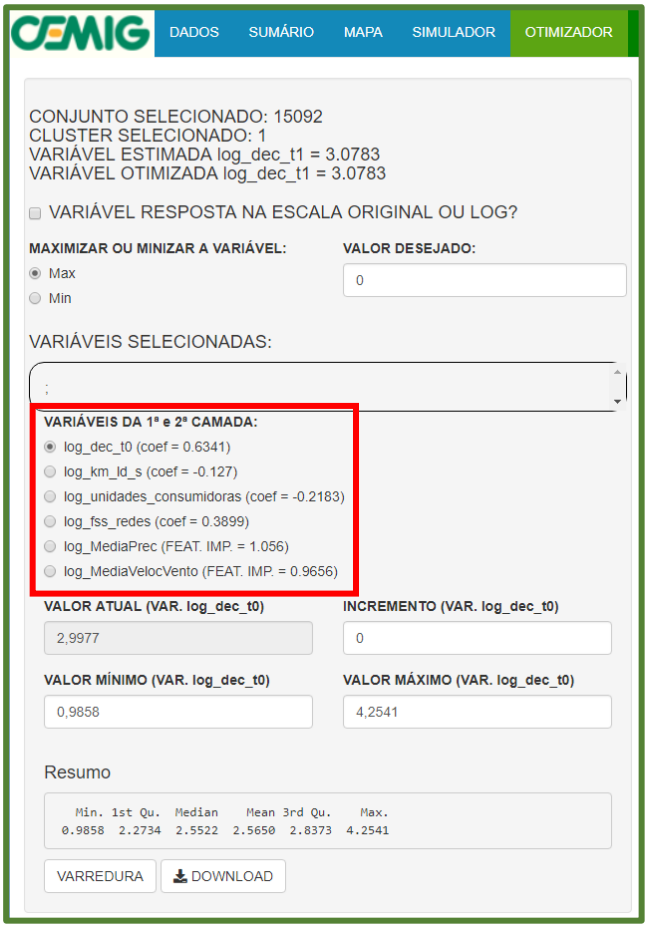
- Em seguida, o usuário deve indicar se o objetivo é de maximizar ou minimizar a variável resposta. Para tanto, deve selecionar a opção desejada no campo "**MAXIMIZAR OU MINIMIZAR A VARIÁVEL?**".



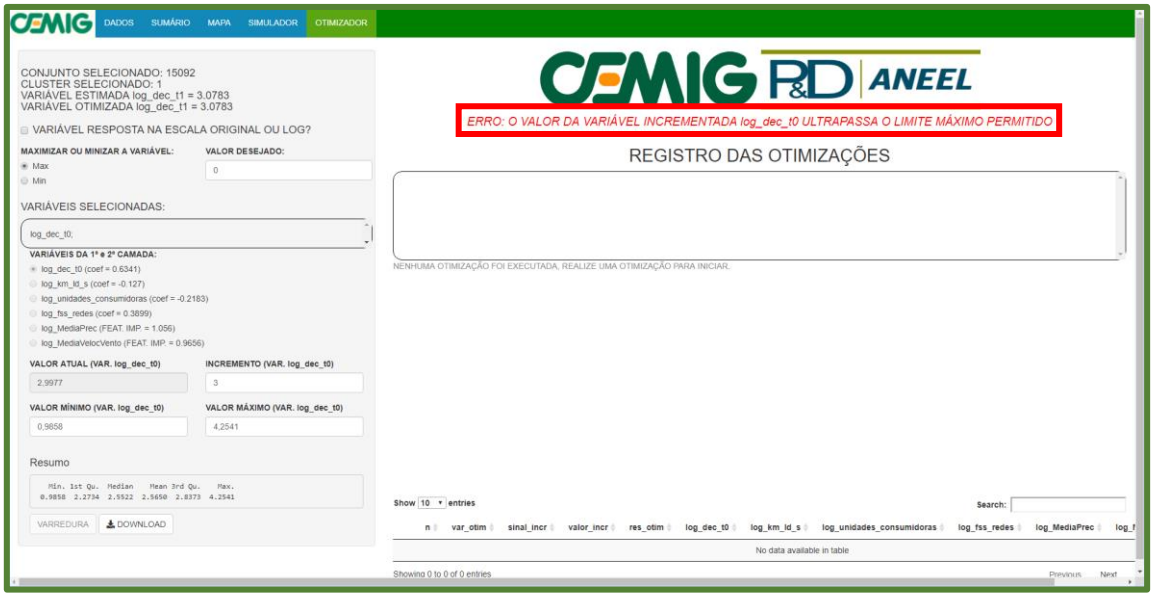
- O usuário deve então indicar qual valor mínimo ou máximo que o otimizador deve buscar alcançar. O valor pode ser inserido no campo "**VALOR DESEJADO:**". Caso o usuário tenha selecionado maximizar, ele deve inserir um valor maior do que o estimado; caso contrário, deve inserir um valor menor do que o estimado.



- Em seguida o usuário deve indicar quais variáveis deseja que o otimizador faça a varredura em busca da solução ótima. O otimizador apresenta todas as variáveis preditoras significativas das 1ª e 2ª camadas: as variáveis da 1ª camada exibem o seu coeficiente entre parênteses; as variáveis da 2ª camada exibem o seu valor de *feature importance*.



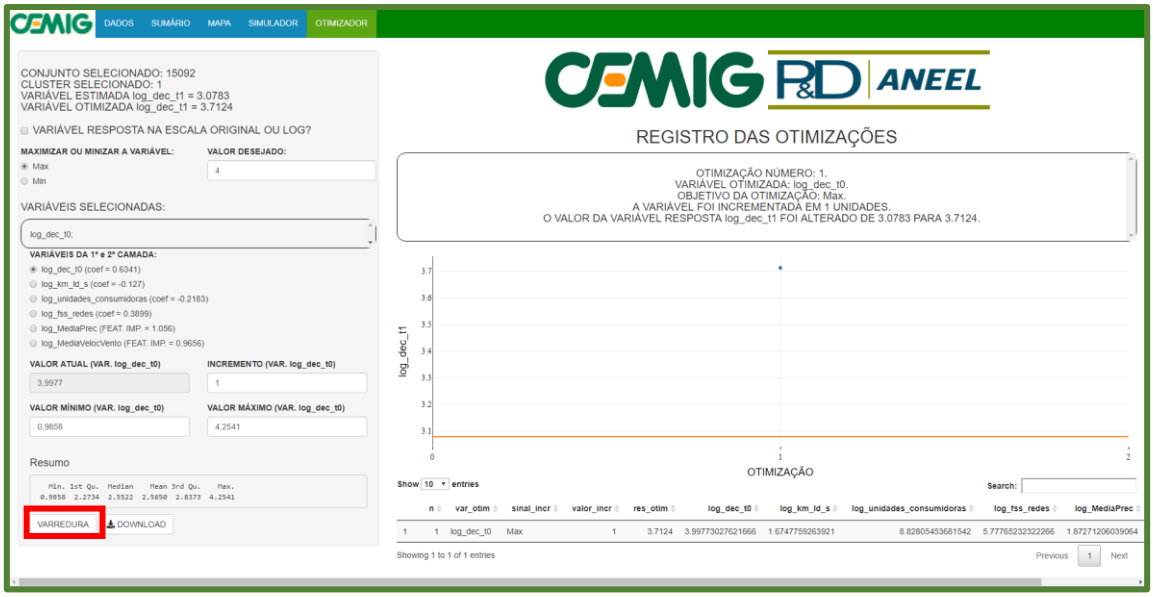
- Ao selecionar uma variável surgem quatro campos dos quais três devem ser preenchidos: “**VALOR INICIAL**”, “**INCREMENTO**”, “**VALOR MÍNIMO**” e “**VALOR MÁXIMO**”. O primeiro campo é o único que não pode ser alterado, ele apresenta o valor atual/inicial da variável selecionada (serve como parâmetro de referência para definir os demais campos). O campo “**INCREMENTO**” deve ser preenchido com o valor da variação que o otimizador deve aplicar na variável selecionada (atentar para inserir o valor com o sinal – positivo ou negativo – adequado). Os campos “**VALOR MÍNIMO**” e “**VALOR MÁXIMO**” delimitam os limites que a variável preditora selecionada pode assumir. Caso o usuário insira um incremento positivo, somente o campo “**VALOR MÁXIMO**” deve ser alterado; caso o usuário insira um incremento negativo, somente o campo “**VALOR MÍNIMO**” deve ser alterado. O campo “**VALOR MÁXIMO**” inicia preenchido com o maior valor encontrado para a variável selecionada dentre todos os conjuntos elétricos na base de dados. O campo “**VALOR MÍNIMO**” inicia preenchido com o menor valor encontrado para a variável selecionada dentre todos os conjuntos elétricos na base de dados. Se o usuário alterar o campo “**INCREMENTO**” sem alterar o limite correspondente para um valor adequado, todos os botões que o permitem avançar na simulação são congelados até que seja corrigido e um erro explicando o contexto é exibido.



- Além disso, ao selecionar a variável é apresentado um resumo (1º e 3º quartis, média, mediana, mínimo e máximo) dos valores de todos os conjuntos elétricos logo abaixo no campo "**RESUMO**". Isso deve ajudar o usuário a definir limites factíveis para o conjunto selecionado.

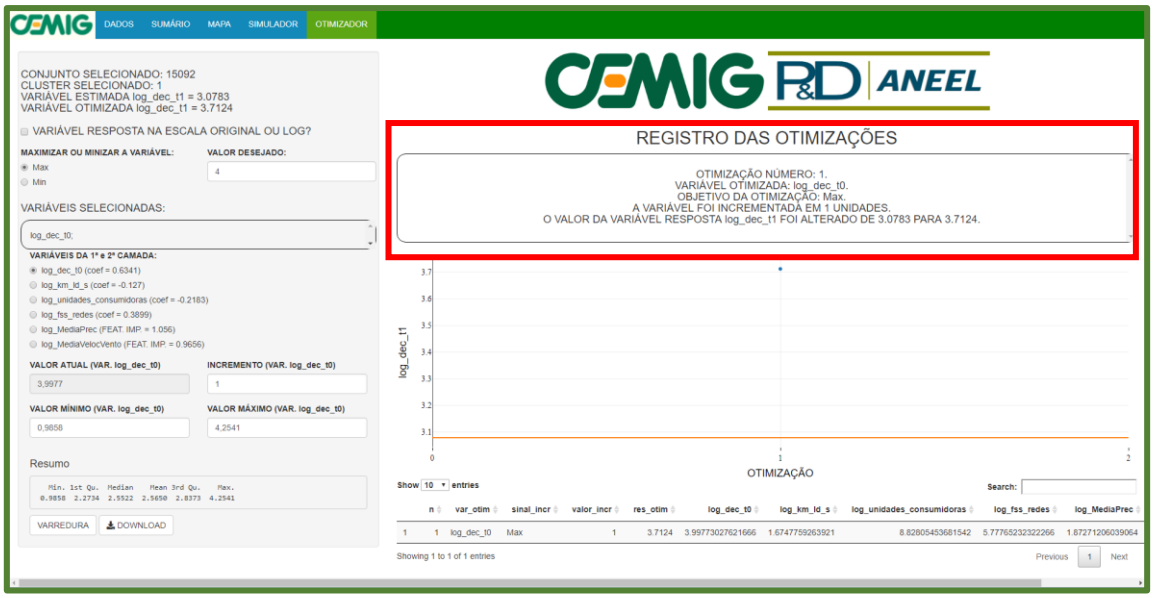
- Ao selecionar uma variável e preencher os campos corretamente a variável escolhida é exibida na caixa "**VARIÁVEIS SELECIONADAS**". Esse campo tem por objetivo indicar para o usuário quais variáveis serão utilizadas durante a otimização.

- Após inserir o objetivo da simulação (maximização ou minimização), o valor desejado para a variável resposta e todas as variáveis desejadas (com seus respectivos incrementos e limites), o usuário pode prosseguir e pressionar o botão "**VARREDURA**". Ao fazer a varredura, o otimizador buscará dentre as variáveis preditoras selecionadas (avaliando os seus incrementos e seus limites) qual variável deve ser alterada uma única vez no valor do incremento indicado tal que o valor da variável resposta fique o mais próximo do valor desejado.

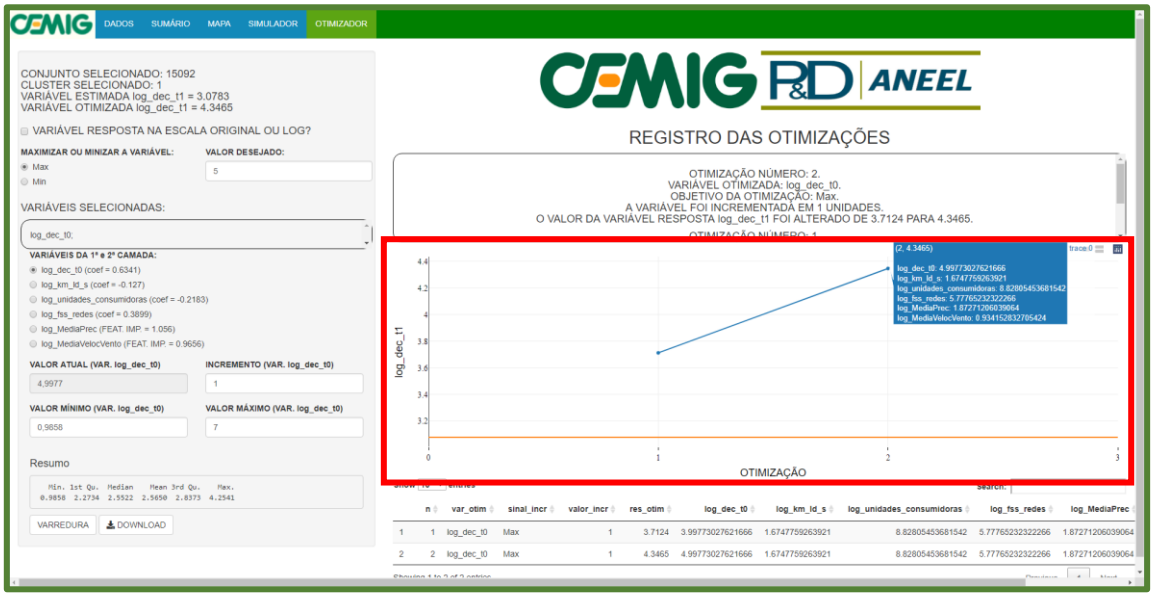


- O resultado da otimização é apresentado em três campos:

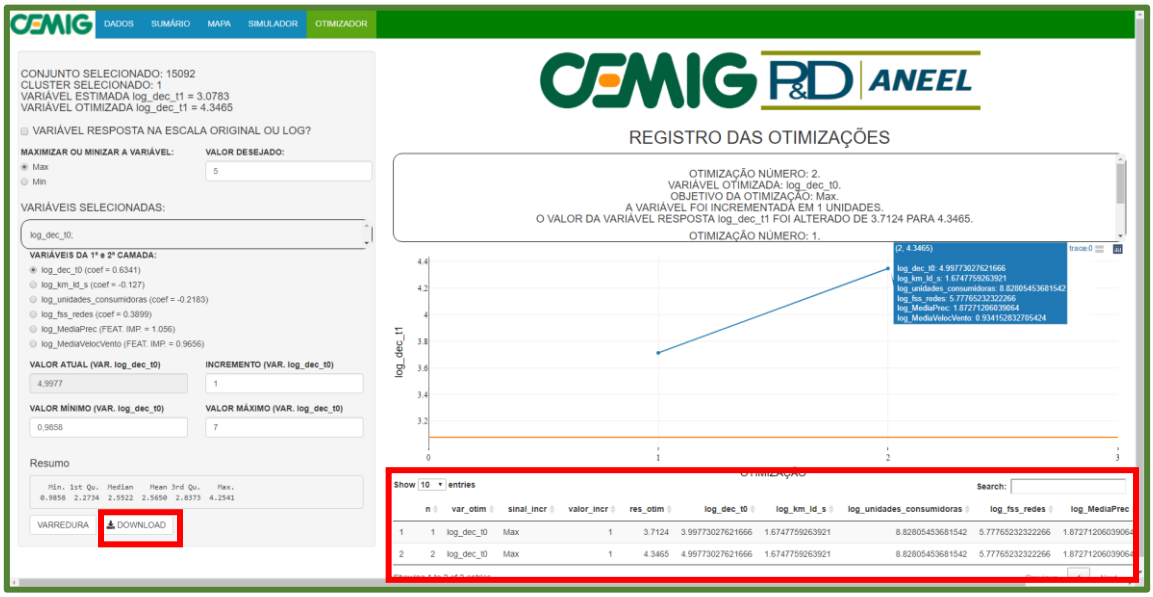
- No campo “**REGISTRO DAS OTIMIZAÇÕES**” ficam os registros de cada otimização que foi realizada: “**OTIMIZAÇÃO NÚMERO**” indica o número da otimização realizada; “**VARIÁVEL OTIMIZADA**” indica qual variável preditora foi escolhida para esse passo de otimização; “**OBJETIVO DA OTIMIZAÇÃO**” indica se a otimização é de maximização ou minimização; “**A VARIÁVEL FOI INCREMENTADA EM X UNIDADES**” indica o valor do incremento que foi aplicado na variável preditora; “**O VALOR DA VARIÁVEL RESPOSTA Y FOI ALTERADO DE X PARA Y**” indica o quanto a variável resposta foi alterada.



- Abaixo é exibido um gráfico de pontos e retas com os valores das otimizações. Também é exibido uma reta com o valor da variável resposta estimada pelo modelo. O gráfico é interativo, ao posicionar o cursor em cima do ponto, são exibidos o valor otimizado da variável resposta e os valores das variáveis preditoras que geraram tal resultado.



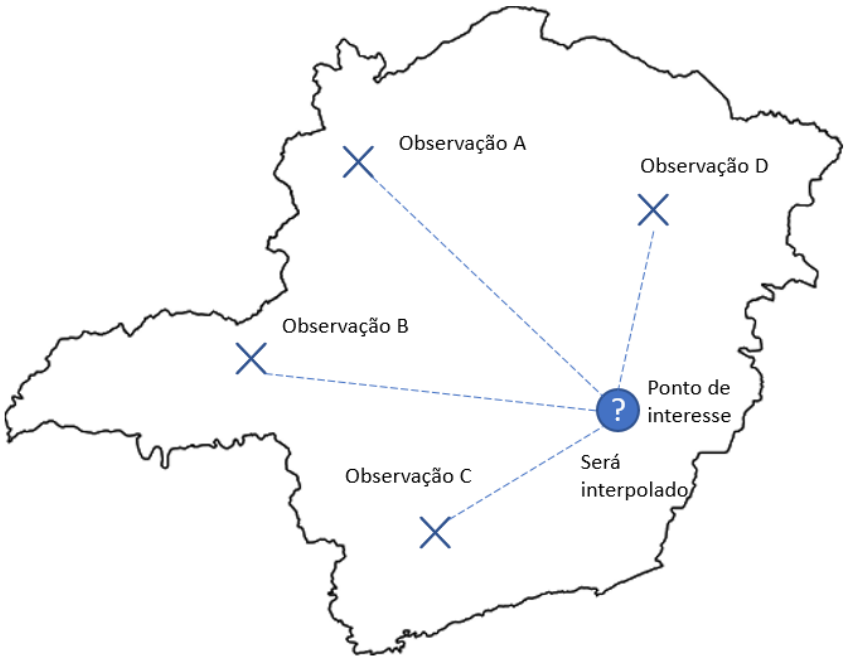
- Por fim, abaixo é apresentada uma tabela contendo o resumo de todas as otimizações realizadas. Essa tabela pode ser baixada no computador do usuário em formato Excel ao pressionar o botão "**DOWNLOAD**".



APÊNDICE

Contextualização sucinta sobre a metodologia da interpolação espacial e o parâmetro IDP.

O método de interpolação na interface é o IDW (Inverse Distance Weighting), no qual o valor de uma variável de interesse em um determinado ponto em um mapa é calculado a partir de uma soma ponderada de todas as demais observações observadas em outros pontos desse mapa. Os pesos são definidos em função da distância entre o ponto de interesse e os pontos com as observações. Valores coletados em pontos mais próximos ao ponto de interesse têm maior peso na ponderação e na definição de qual será o resultado da interpolação. A imagem a seguir ilustra a metodologia.



De forma sucinta, o peso atribuído a cada observação é uma função da sua distância ao ponto de interesse. A taxa da queda dos pesos em função da distância é definida pelo parâmetro IDP (IDW Power Coefficient). Quanto maior o IDP, maior a taxa de decaimento dos pesos com a distância, ou seja, a interpolação será definida a partir dos resultados observados nos pontos mais próximos. Quanto menor o IDP, menor a taxa de decaimento, permitindo que a interpolação seja definida a partir de resultados observados nos pontos mais distantes.

Portanto, para a execução da interpolação na interface, o código desenvolvido busca identificar qual é o valor do IDP que minimiza o erro do resultado da interpolação, atingindo os melhores resultados. Para tanto, o usuário executará os ajustes dos limites do parâmetro IDP visíveis na aba "**MODELO**".