UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

BRUNO CRISTIANO GOMES

A estatística como ferramenta de apoio à auditoria: Uma discussão a respeito da capacidade de identificação de exceções em dados de empresas de energia elétrica

Belo Horizonte

2021

BRUNO CRISTIANO GOMES

A estatística como ferramenta de apoio à auditoria: Uma discussão a respeito da capacidade de identificação de exceções em dados de empresas de energia elétrica

Versão final da monografia de especialização apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito à obtenção do título de **Especialista em Estatística.**

Orientador: Prof. Dr. Thiago Rezende dos Santos

Belo Horizonte

2021

Gomes, Bruno Cristiano.

G633e A estatística como ferramenta de apoio à auditoria [manuscrito]: uma discussão a respeito da capacidade de identificação de exceções em dados de empresas de energia elétrica. / Bruno Cristiano Gomes. — 2021.

xiii,75.f. il.

Orientador: Thiago Rezende dos Santos.

Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências 72-75.

1. Estatística. 2. Auditoria. 3. Estatística aplicada à auditoria. I. Santos, Thiago Rezende dos.II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatísitca .III.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 220ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE BRUNO CRISTIANO GOMES.

Aos vinte e dois dias do mês de fevereiro de 2021, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Bruno Cristiano Gomes**, intitulado: "*A estatística como ferramenta de apoio à auditoria: Uma discussão a respeito da capacidade de identificação de exceções em dados de empresas de energia elétrica*", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Thiago Rezende dos Santos – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 22 de fevereiro de 2021.

Prof. Thiago Rezende dos Santos (Orientador)
Departamento de Estatística / UFMG

Prof. Cristiano de Carvalho Santos
Departamento de Estatística / UFMG

Prof. Victor Schmidt Comitti
Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais

RESUMO

Objetivo: Consiste na avaliação da capacidade do uso de técnicas e métodos estatísticos para identificação de exceções no apoio a trabalhos de auditoria. **Método**: Tipo de pesquisa realizada, quanto aos objetivos, descritiva e, quanto aos procedimentos, documental. Foram utilizados dados das distribuidoras CEMIG, COELBA e ELETROPAULO, publicados pela ANEEL, no período de janeiro de 2017 a julho de 2020. As técnicas utilizadas foram estatísticas exploratórias, regressão linear, modelos ARIMA para séries temporais, lei de Benford e Análise de Componentes Principais. **Resultados**: A análise exploratória dos dados, por meio da apuração da média, mediana, desvio padrão e moda, se mostrou eficaz na identificação de exceções. A aplicação da correlação nos dados de Consumo, Receita, Número de UC's e Tarifas possibilitou identificar padrões que necessitam uma análise dos microdados segregados por clientes. O maior benefício da Lei de Benford foi dar indícios de elementos que não necessariamente são valores extremos, podendo ser muito útil no direcionar de amostras a serem realizadas. Para distância de Cook os elementos mais influentes estavam mais concentrados no ano de 2020, decorrentes do impacto da pandemia. Em relação à Análise de Componentes Principais, o maior benefício foi o direcionamento da análise para as variáveis que tem maior capacidade de explicar a variabilidade da base de dados. **Conclusão**: as técnicas estatísticas utilizadas foram capazes de identificar exceções e auxiliar o direcionamento para aprofundamento de análises.

Palavras-chave: Auditoria; Estatística; Estatística aplicada à auditoria; *Littlewood*; *Benford law*; Energia; CEMIG;

ABSTRACT

Objective: Assess the ability to use statistical techniques and methods to identify exceptions in support of audit work. **Method**: Type of research carried out, as to the objectives, descriptive, and, as to the procedures, documentary. Data from the companies CEMIG, COELBA and ELETROPAULO, published by ANEEL, from January 2017 to July 2020 were used. The techniques used were exploratory statistics, linear regression, ARIMA models for time series, Benford's law and Principal Component Analysis. **Results**: The exploratory analysis of the data, by calculating the mean, median, standard deviation and mode, proved to be effective in identifying exceptions. The application of the correlation in the data of Consumption, Revenue, Number of UC's and Tariffs made it possible to identify patterns that need an analysis of the microdata segregated by customers. The greatest benefit of Benford's Law was to give evidence of elements that are not necessarily extreme values, which can be very useful in directing samples to be performed. For Cook's distance, the most influential elements were more concentrated in the year 2020, due to the impact of the pandemic. In relation to Principal Component Analysis, the greatest benefit was to direct the analysis towards the variables that have the greatest capacity to explain the variability of the database. **Conclusion**: the statistical techniques used were able to identify exceptions and assist in directing further analysis.

Keywords: Audit; Statistic; Statistics applied to the audit; Littlewood; Benford law;

AGRADECIMENTOS

Ao professor orientador Thiago Rezende e demais professores do curso de especialização em Estatística.

Aos meus pais, Suely e Antônio, meu irmão, Érico, pela referência de vida e motivação para sempre buscar o conhecimento.

À minha linda esposa Elziane que sem o seu apoio esta conquista definitivamente não seria possível.

LISTA DE TABELAS

LISTA DE FIGURAS E GRÁFICOS

SUMÁRIO

INTRODUÇÃO

A pouco conhecida Lei de *Littlewood*, "*Littlewood's law*", também denominada de "*Law of miracles*", de *John Edensor Littlewood*, estabelece que existe uma probabilidade de 1 em 1 milhão de eventos que um milagre aconteça. Conceituando milagre como um evento raro e, extrapolando "evento" para atos e fatos que ocorrem nas organizações, é provável que eventos raros permeiem as organizações, segundo essa lei.

Não é a intenção comprovar a probabilidade de ocorrência de eventos extremos, mas sim de buscar ferramentas que auxiliem nessa identificação. A Estatística, enquanto área que tem dados como matéria-prima, tem papel fundamental neste apoio, uma vez que "*transforma dados em informações úteis para os tomadores de decisão (...) faz com que você passe a conhecer os riscos associados à tomada de decisões no contexto empresarial e permite que você compreenda e minimize a variação no processo de tomada de decisão"* (Levine, 2014). Percebe-se, então, a importância da Estatística no uso de dados e seus reflexos na tomada de decisões e, combinada com outras disciplinas, como a computação, é capaz de ultrapassar fronteiras, aperfeiçoando ainda mais essa tomada de decisões.

Conforme a pesquisa "*The Future of Jobs Report*" (2018, p.9) feita pelo do "*World Economic Forum*", a auditoria junto com a contabilidade, são profissões que sofrerão uma crescente redução por serem susceptíveis a avanços tecnológicos e processos de automação, muito desses avanços provenientes de uma interação multidisciplinar. Por este motivo, a auditoria, enquanto área que deve identificar exceções e agregar valor às operações de uma organização, segundo o Instituto dos Auditores Internos - IIA, tem que mostrar mais sua relevância para as organizações e também interagir com outras disciplinas.

Sendo assim, suportada por técnicas e métodos estatísticos, aliada a infraestruturas computacionais, a atividade de auditoria pode vislumbrar um horizonte com um enorme potencial de inovação, de crescimento, aumentando sua relevância. Dentre os principais benefícios, pode-se citar técnicas que identifiquem padrões, antecipe riscos, exponha exceções, ou seja, benefícios advindos dessa interdisciplinaridade e que podem potencializar o trabalho dos auditores.

O presente trabalho tem o objetivo de, a partir de bases de dados publicadas na ANEEL – Agência Nacional de Energia Elétrica de empresas de energia elétrica, aplicar técnicas e métodos estatísticos e avaliar sua eficácia na detecção de possíveis exceções. Pretende-se aplicar técnicas simples nas bases de dados, como as medidas de tendência central, e também mais rebuscadas, como regressão e séries temporais, com a finalidade de identificar possíveis desvios, orientando, assim, o trabalho dos auditores e a tomada de decisões.

Definição de Auditoria

O Instituto dos Auditores Internos – IIA define a atividade de auditoria como:

> *A auditoria interna é uma atividade independente e objetiva de avaliação e consultoria, criada para agregar valor e melhorar as operações de uma organização. Ela auxilia a organização a atingir seus objetivos a partir da aplicação de uma abordagem sistemática e disciplinada à avaliação e melhoria da eficácia dos processos de gerenciamento de riscos, controle e governança.*

Araújo (1998, p. 15) simplifica esse conceito, trazendo uma definição mais prática: "*auditoria é o processo de confrontação entre uma situação encontrada e um determinado critério, ou, em outras palavras, é a comparação entre o fato ocorrido e o que deveria ocorrer*".

Ambos conceitos se complementam, mas, para o presente estudo, será dado destaque à passagem da definição do IIA em que a auditoria é *"criada para agregar valor e melhorar as operações de uma organização"*. No entanto, em função deste excerto, pode surgir o seguinte questionamento: De que forma a auditoria pode agregar valor e melhorar operações?

Tipos de Auditoria

Boynton (2002, p.31) inicia a resposta a esse questionamento separando a auditoria em 3 tipos:

- Auditoria de demonstrações financeiras;

- Auditoria de *compliance;*
- Auditoria operacional.

A auditoria de demonstrações financeiras envolve avaliação de evidências a respeito dos relatórios financeiros de uma entidade, para emissão de parecer se sua apresentação está adequada, de acordo com os princípios contábeis. A auditoria de *compliance*, por sua vez, envolve obtenção e avaliação de evidências para determinar se certas atividades financeiras ou operacionais de uma entidade obedecem a condições, regras ou regulamentos a elas aplicáveis. *Boynton* conclui com a auditoria operacional que, assim como as outras, envolve obtenção e avaliação de evidências a respeito da eficiência e eficácia das atividades operacionais de uma entidade, como: gestão de estoques, logística, arrecadação, saúde e segurança.

Todos os tipos de auditoria citados, como se pode ver, têm em comum a obtenção e avaliação de evidências. Neste sentido, o *Public Company Accounting Oversight Board* – PCAOB, órgão criado pelo Congresso Norte-Americano para supervisionar auditorias de empresas que negociam ações, orienta, *no Auditing Standard 15 – AS 15*, o que pode ser entendido como **evidência de auditoria**:

> *Evidência de auditoria é toda informação, obtida por meio de procedimentos de auditoria ou outras fontes e são usadas pelo auditor chegar nas conclusões pelas quais sua opinião é baseada. (...) consiste em informações que suportem e corroborem as afirmações da administração a respeito das demonstrações financeiras ou controles internos sobre relatórios financeiros e informações que contradizem tais afirmações.*

Como se viu a evidência de auditoria deve corroborar as conclusões dos auditores, contradizendo informações, ou seja, pode ser entendida como exceções identificadas a respeito de fatos contábeis ou financeiros, por exemplo.

O PCAOB, ainda no AS 15, desenvolve a resposta detalhando 7 tipos principais procedimentos para se obter evidencias de auditoria:

1. **Inspeção**: Envolve o exame físico de documentos, internos ou externos, em papel, eletrônico ou outro meio. Inspeção de registros e documentos fornece

evidência de auditoria em variados graus de confiança, dependendo de sua natureza, origem e, nos casos de registros internos, na efetividade dos controles que os produzem. Um exemplo de inspeção usado em testes de controle é a inspeção de registros para evidenciar de autorizações de acesso.

2. **Observação**: Consistem em observar um processo ou procedimento ser performado por outros, ex.: o auditor observa a contagem de estoque por um empregado da companhia. Observação pode fornecer evidência sobre a performance de um processo ou procedimento, mas, neste caso, a evidencia é limitada ao tempo e local da observação e também limitada pelo fato de que o ato de ser observado pode afetar como o processo ou procedimento é realizado.
3. **Entrevista**: pode ser realizada durante a auditoria, além de outros procedimentos de auditoria. Podem variar de entrevistas formais por escrito a informais orais. Avaliar as respostas às perguntas é parte integrante do processo de consulta
4. **Confirmação**: Representa uma forma particular de evidência de auditoria obtida pelo auditor por meio de um terceiro, de acordo com os padrões do PCAOB.
5. **Recálculo**: consiste em se verificar a precisão dos cálculos matemáticos de documentos ou registros. Pode ser realizado manual ou eletronicamente.
6. ***Reperformance***: A repetição do desempenho envolve a execução independente de procedimentos ou controles que foram originalmente executados pelo pessoal da empresa.
7. **Procedimentos analíticos**: Os procedimentos analíticos consistem em avaliações de informações financeiras feitas por um estudo de relações plausíveis entre dados financeiros e não-financeiros. Os procedimentos analíticos também abrangem a investigação de diferenças significativas dos valores esperados.

Estes procedimentos não são isolados e podem ser *performados* em conjunto, como é o caso da entrevista que, se feita em conjunto com uma inspeção, permite esclarecer dúvidas no momento da análise.

Por fim, o *Internacional Standards of Auditing* 315 - ISA 315, conclui a resposta apresentando as “assertivas” que são os tipos de distorções que o auditor pode identificar nas suas análises (p.264). Essas assertivas, podem ser, dentre outras (p. 298):

- **Existência** — Existência física de ativos, dívidas;
- **Totalidade** — Verificação se todas transações que deveriam ser registradas, foram registradas, ou seja, se a base de dados está completa, íntegra;
- **Exatidão** — Apurar se os cálculos ou valores dados foram registrados/calculados corretamente;
- ***Cutoff*** — se transações ou eventos foram registrados nos períodos corretos.

Isto posto, respondendo ao questionamento inicial “*De que forma a auditoria pode agregar valor e melhorar operações?*” E a resposta: Por meio da aplicação de procedimentos de auditoria que permitam ao auditor coletar evidências confiáveis, com base nas assertivas, que, quando analisadas e tratadas, agreguem valor aos processos e permitam as empresas atingirem seus objetivos.

Auditoria e fraudes

A norma 1210.A2 do IIA estabelece que

> Os auditores internos devem possuir conhecimento suficiente para avaliar o risco de fraude e a maneira com o qual é gerenciado pela organização, porém, não se espera que possuam a especialização de uma pessoa cuja principal responsabilidade seja detectar e investigar fraudes.

Observa-se que o auditor deve considerar o risco de fraudes nos trabalhos de auditoria, apesar de não ser esperado que seja responsável por identificar fraudes. O Instituto na “Declaração de Posicionamento do IIA” (2019) afirma que o auditor deve ter conhecimento suficiente para:

- Identificar alertas vermelhos que indiquem que pode ter havido fraude;
- Compreender as características da fraude, as técnicas usadas para cometê-la e os diversos esquemas e cenários de fraude;

- Avaliar a eficácia dos controles de prevenção ou detecção de fraudes.

O documento acrescenta que não é responsabilidade direta da auditoria interna impedir que fraudes ocorram na empresa, essa é uma responsabilidade da administração como primeira linha de defesa. No entanto, os escândalos recentes, como Wirecard (DW, 2020) e no IRB Brasil (Folha, 2020), colocaram luz nas atribuições e responsabilidades dos auditores na identificação de fraudes.

Segundo *Brian Fox* (LEMMON), a profissão de auditor pode se tornar obsoleta num futuro não muito distante se essa responsabilidade na identificação de fraudes não existir. Pondera, porém, que os auditores não devem encontrar 100% das fraudes, mas devem encontrar distorções materiais, não importando se em função de erro ou fraude, concluindo que "*se é distorção material os auditores devem identificar*".

Dessa forma, a identificação de fraudes passa, num primeiro momento, pela identificação de exceções que, com uma análise aprofundada, pode resultar em fraudes ou distorções materiais.

Exceções em auditoria

Conforme dito anteriormente, a finalidade do trabalho dos auditores é, e deve ser, *agregar valor e melhorar as operações de uma organização*. Todavia, no dia-a-dia da atividade a busca incessante é por erros, sejam eles: intencionais ou não.

Neste estudo, utilizar-se-á como conceito de exceções de auditoria, eventos extremos ou não que necessitem aprofundamento, afim de se caracterizar erro ou fraude. As exceções identificadas podem dar indícios para, numa posterior análise mais aprofundada, caracterizar um erro intencional (1), erro não-intencional (2) ou apenas um evento natural (3):

1. Erros de digitação, de cálculo, de interpretação podem ser exemplos erros **não intencionais.**

2. Em relação aos **intencionais**, Nigrini (2020), destaca alguns padrões mais comuns:
    - Números arredondados;
    - Limítrofes;
    - Duplicados;
    - Valores de crescimentos rápidos;
    - Números que não sigam a lei do primeiro dígito;
    - O*utliers*.

3. Os **eventos naturais**, por fim, não são erros nem fraudes, mas ocorrências pouco frequentes, que chamam atenção principalmente de quem não está no dia-a-dia do negócio, mas que precisam de um maior esclarecimento. Um exemplo de evento natural é o estorno de pagamento, porque no meio da relação de receitas de uma empresa podem existir valores negativos, decorrentes de estornos, uma situação completamente normal.

No entanto, deve-se ter em mente que, por mais que sejam utilizados auditores experientes ou técnicas avançadas para identificação de erros ou fraudes, é pouco provável que será possível identificar todas exceções existentes nas bases de dados. Não obstante, é fundamental, para quem deseja ter sucesso nessa empreitada, além de pessoal e técnicas adequadas, buscar novos conhecimentos, fazer inter-relações e "*pensar com a cabeça do criminoso"* (GOMES, p. 79, 2019), se for este o caso.

PROBLEMATIZAÇÃO

A auditoria não é uma atividade que sobrevive isolada, isto é, ela precisa de processos corporativos para que possa ser executada, como: financeiro, suprimentos recursos humanos, TI. Esses processos são os ambientes pelos quais a auditoria frequenta para atingir seu objetivo de aumentar e proteger o valor organizacional.

Alvin Toffler (1984, p. 181) cunhou o termo “Sobrecarga Informativa” em que explica que quando os indivíduos estão mergulhados em situação de mudança rápida e irregular ou em um contexto carregado de novidades, sua precisão preditiva despenca. Para compensar isso, os indivíduos precisam processar muito mais informações do que antes:

> *(…) In the words of psychologist George A. Miller of Rockefeller University, there are "severe limitations on the amount of information that we are able to receive, process, and remember." By classifying information, by abstracting and "coding" it in various ways, we manage to stretch these limits (…)*

Isto é, segundo *Toffler*, existem limites para a quantidade de processamento dessas informações e sugere classificar as informações, abstraindo-as e "codificando-as" de várias maneiras, conseguimos estender esses limites”. Dessa forma, tendo em vista a atual dinâmica da economia, concorrência, redução de custos, é visível a mudança rápida e contínua dos processos corporativos. Esses processos geram um excesso de dados e informações desafiando os auditores a tirar o máximo proveito. Como consequência, também evolui, na mesma proporção, novas formas e tipos de erros, sejam fraudulentos ou não, que precisam cada vez mais da *expertise* dos auditores e de técnicas e métodos para identificação.

No contexto de auditoria, a etapa de planejamento de trabalhos, que tem os planos anuais como produto, deve acompanhar essa dinâmica. Esses planos, têm como passo inicial a realização da avaliação de riscos que geram uma priorização dos processos a serem auditados na empresa. Nesse processo, são avaliados riscos passados que, sem as técnicas adequadas, podem não refletir os desafios vigentes ou futuros a serem enfrentados pelas empresas. Isso pode diminuir a relevância da

atividade de auditoria por ser mais reativa do que proativa ou ainda “atacar” riscos irrelevantes.

É também uma prática da auditoria “contemporânea” o uso de técnicas de amostragem que se justificam *quando o custo e tempo de se fazer um exame em 100% dos dados são maiores que as consequências adversas de possivelmente emitir um parecer errôneo* (*Boynton*, 2002, p.456). Contudo, a auditoria é cada vez mais desafiada a identificar fraquezas materiais e fraudes (LEMON, 2020) e a análise amostral, ou métodos muito manuais, podem não ser tão efetivos neste contexto. Dessa forma, a demanda por trabalhos que utilizem uma amostragem mais assertiva ou que verifiquem 100% da população dos dados de uma forma mais automatizada é cada vez mais uma necessidade.

Outro ponto a se considerar, mencionado anteriormente, são as enormes bases de dados que os processos geram diariamente. Estas bases contêm informações cujos auditores precisam extrair exceções sob pena de não serem eficazes no seu trabalho. Apesar da auditoria se utilizar de técnicas estatísticas, pode ser que as utilizadas atualmente possam não responder na qualidade e velocidade necessárias para as empresas atingirem seus objetivos.

Pelos exemplos acima, fica clara a necessidade de adaptação e evolução da atividade dos auditores nesse contexto de mudanças rápidas e disruptivas pelas quais passam a sociedade e, consequentemente, também as organizações.

OBJETIVO

Objetivo geral

O objetivo deste trabalho consiste na avaliação da capacidade do uso de técnicas e métodos estatísticos para identificação de exceções no apoio a trabalhos de auditoria.

Objetivos específicos

1. Avaliar a capacidade das técnicas estatísticas exploratórias, regressão linear, modelos ARIMA para séries temporais, lei de Benford e Análise de Componentes Principais para identificação de exceções;

2. Identificar quais técnicas estatísticas são mais eficazes na identificação de exceções.

JUSTIFICATIVA

A importância deste trabalho reside no fato de que a atividade de auditoria precisa ter formas para acompanhar um ambiente cada vez mais complexo, mitigando a probabilidade e impacto de riscos. Para isso, deve utilizar técnicas e métodos eficazes para o cumprimento de sua missão de "*Aumentar e proteger o valor organizacional, fornecendo avaliação (assurance), assessoria (advisory) e conhecimentos (insights) objetivos baseados em riscos.* " (IIA – Missão da Auditoria Interna).

Deve-se salientar que o aumento dessa complexidade, pode ser acompanhado da ocorrência de erros mais frequentes e também de novos tipos de fraude. Sendo assim, os auditores precisam entender profundamente os processos e terem a disposição uma variedade de métodos que lhes permitam ser efetivos na realização de suas tarefas.

Como em outras áreas, a estatística tem potencial para apoiar e muito a auditoria no atingimento destes objetivos, uma vez que, tem uma enorme gama de ferramentas e métodos para trabalhar com bases de dados, extraindo conhecimentos. Seja pelo uso das mais simples técnicas exploratórias, como valores mínimos e máximos, até a identificação de padrões pelo uso de regressões, séries temporais, a estatística coloca tudo isso à disposição como meio para direcionar um maior aprofundamento.

REFERENCIAL TEÓRICO

A identificação de exceções em bases de dados por meio de técnicas estatísticas tem sido um campo muito estudado recentemente em função da popularização do uso do *machine learning,* uma vez que valores extremos afetam diretamente os modelos. Também por este motivo, existe uma diversidade de materiais que abordam o tema e serão abordados no presente trabalho.

Deve-se salientar que as técnicas a serem estudadas podem não necessariamente identificar exceções, mas tem enorme potencial de direcionar o trabalho dos auditores,

seja no apoio a essa identificação, na análise do comportamento ou fornecendo indícios para um aprofundamento mais assertivo.

O Instituto dos Auditores Internos - IIA orienta, em seus diversos normativos, que o auditor deve identificar, analisar, avaliar e documentar informações suficientes para atingir os objetivos do trabalho (IIA 2300). Neste sentido, considerando o contexto, de que os departamentos de auditoria devem fazer mais com menos (GTAG 16, p. 3), a auditoria deve buscar novas ferramentas e tecnologias para melhorar sua efetividade. Dentre essas ferramentas cita o uso de análises estatísticas avançadas, como correlação, análise de tendências, séries temporais, lei de *Benford*. (GTAG 16, p. 20).

Uma importante ferramenta estatística que pode auxiliar o trabalho de auditores é a **análise de *cluster***. O objetivo da análise de *cluster* é identificar grupos significativos nos dados. Normalmente, nos dados esses grupos estão internamente coesos e, a análise de cluster, auxilia a encontrar grupos cujos membros tenham algo em comum que não compartilham com membros de outros grupos (BOUVEYRON, p.1). Dessa forma, aplicar a *clusterização* em grandes bases, agrupando os dados automaticamente, conforme características comuns, pode ser de grande valia para auditores.

Outras ferramentas úteis são apresentadas no livro "*Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*" (BAESENS, 2015) que traz o estado da arte para detecção de fraudes utilizando análise de dados. Os autores descrevem as técnicas necessárias para detectar fraude, apresentando aos leitores desde o básico até a metodologia avançada de reconhecimento de padrões (BAESENS, p.13). No presente trabalho, este referencial será base para, utilizando os métodos descritos, identificar exceções por meio de: estatísticas descritivas, demonstração gráfica dos dados, detecção de *outliers*, Análise de Componentes Principais - ACP e regressão linear.

Das ferramentas citadas anteriormente, a aplicação de **estatísticas descritivas ou exploratórias** é o primeiro passo para se trabalhar com base de dados. Apesar de populares e de uma aparente simplicidade, a aplicação de média, moda, mediana,

desvio padrão, demonstração gráfica dos dados, além de permitir um conhecimento imediato da base de dados, possibilita também identificar padrões ou valores extremos muitas vezes de forma mais eficaz do que técnicas mais contemporâneas. Um exemplo simples e eficaz é a identificação de valores faltantes por meio de uma simples análise de um gráfico de linhas ou identificar valores extremos por meio de um *boxplot*.

Em se tratando de grandes bases de dados, algumas vezes em função da variedade de informações pode ser necessário a utilização de técnicas que reduzam o tamanho daquela base. Desse modo, a **Análise de Componentes Principais - ACP**, pode prestar um excelente serviço no apoio às atividades de identificação de exceções.

A Análise de Componentes é uma técnica de estatística multivariada que está relacionada à ideia de redução de massa de dados, com menor perda possível da informação. Essa técnica possibilita transformar um conjunto de variáveis originais em outro conjunto de variáveis de mesma dimensão denominadas de Componentes Principais (HONGYU, 2015), sendo seu objetivo **identificar as direções** (ou componentes principais) **cuja variabilidade dos dados é máxima** (KASSAMBARA, p. 12). Dessa forma, ao invés de trabalhar com uma base imensa de variáveis, a identificação dos componentes principais pode simplificar a análise e, com isso, o analista melhorar sua eficiência.

Outra alternativa para simplificar o tamanho das bases de dados é a utilização de **técnicas de amostragem**:

> *Sampling is a process that selects units from a population of interest, in such a way that the sample can be generalized for the population with statistical confidence (…). The statistical confidence will vary based on the sampling technique used and the size. (RAMASUBRAMANIAN, 2018, p.80).*

A literatura estatística destaca formas tradicionais de amostragem, categorizando-as em probabilísticas e não-probabilísticas. Como a intenção deste estudo é a identificação de exceções, podem ser utilizadas formas probabilísticas aliadas a não-probabilísticas, sem a obrigatoriedade da extrapolação dos resultados. Contudo, será

dado foco em amostragens mais julgamentais, sem essa necessidade de extrapolar resultados a princípio, tendo apenas a finalidade de direcionar análises.

A **Lei de Newcomb-Benford – LNB** é uma ferramenta eficaz neste sentido, porque permite identificar grupos de valores que tenham diferenças entre a frequência dos dígitos da base de dados real em relação à frequência estabelecida pela LNB. Também conhecida como lei do primeiro dígito, a LNB **afirma que há uma frequência pré-definida de dígitos** (Tabela 1) numa série de dados naturalmente gerada, exemplo: extensão dos maiores rios do mundo, população de países, valor de notas fiscais (GOMES, 2013, p. 4).

Tab. 1: Frequência de dígitos esperada conforme LNB

Dígito	Frequência
1	30,103%
2	17,609%
3	12,494%
4	9,691%
5	7,918%
6	6,695%
7	5,799%
8	5,115%
9	4,576%

Fonte: Lei de Newcomb-Benford

Posto isso, caso a frequência de algum primeiro dígito de uma base de dados não siga esse padrão, é possível que haja algum erro e sugere-se aprofundar a análise naquele grupo para buscar exceções.

A **Lei dos Grandes Números – LGN** também pode ser uma ferramenta a ser utilizada para identificação de exceções em bases de dados. Conforme *Wasserman* (NA, p. 76), a Lei dos Grandes Números é um teorema que diz que a média de grandes quantidades de amostras é próxima à média da distribuição dos dados. Por exemplo, a proporção de caras de um grande número de lançamentos deve ser próxima a ½. *Leshik* (p. 137) acrescenta que a LGN "*praticamente garante que amostras muito grandes sejam altamente representativas da população da qual foram selecionadas aleatoriamente. Números pequenos não geram essa propriedade*". Sendo assim, segundo o autor, realizando-se grandes quantidades de amostras numa base de

dados, essa média será próxima à média populacional, caso esse comportamento não seja observado, deve-se aprofundar a análise

A revista *Scientific American* complementa a definição da LGN como:

> "(…) *a principle of probability called the Law of Large Numbers shows that an event with a low probability of occurrence in a small number of trials has a high probability of occurrence in a large number of trials". Scientific American. Vol. 291. No 2.* (p. 32)

Ou seja, essa lei da probabilidade mostra que eventos extremamente improváveis tem alta probabilidade de ocorrência num grande número de tentativas. Isto posto, a LGN pode ser uma forma de identificação de exceções no presente estudo. Todavia, deve-se avaliar a população de forma a planejar a seleção da amostra, uma vez que podem existir desbalanceamentos ou concentração de dados que irão afetar consideravelmente a amostra e, assim, a média não representar distribuição dos dados.

Suplementar à LGN, a "lei dos milagres de *Littlewood"*, também pode prestar um serviço inédito para auditores na motivação para identificar exceções. Essa lei estabelece que existe uma frequência para que eventos raros ou "milagres" aconteçam (DYSON, 2006):

> *John Littlewood was a famous mathematician who was teaching at Cambridge University when I was a student. Being a professional mathematician. He defined a miracle as an event that has special significance when it occurs, but occurs with a probability of* ***one in a million.*** *This definition agrees with our commonsense understanding of the word "miracle."*

Littlewood define "milagres" como eventos que ocorrem na proporção de 1:1.000.000. *Freeman Dyson* (2006), exemplifica a ocorrência dessa lei:

> *Littlewood's law of miracles states that in the course of any normal person's life, miracles happen at a rate of roughly one per month. The proof of the law is simple. During the time that we are awake and actively engaged in living our lives, roughly for eight hours each day, we see and hear things happening at a rate of about one per second. So the total number of events that happen to us is about 30,000 per day, or about a million per month. With few exceptions, these events are not miracles because they are insignificant. The chance of a miracle is about one per million events.* ***Therefore we should expect about one miracle to happen, on the average, every month.***

A lei afirma que eventos raros ou milagres são aqueles que tem probabilidade de ocorrência de 1 em 1 milhão e exemplifica que eventos com essa probabilidade ocorrem pelo menos uma vez ao mês na vida de uma pessoa: tomando por base que as pessoas ficam acordadas por volta de 8hr por dia e que, a cada segundo fatos ocorrem com a pessoa, considerando o período de 1 mês, a quantidade de ocorrências é próxima a 1 milhão.

Deve-se enfatizar que não se pretende discutir do que é considerado “evento” para essa lei, tampouco qual a frequência da ocorrência de milagres. Mas, sim, avaliar se a Lei de dos Milagres pode ser capaz de identificar exceções ou, pelo menos, ser um motivador para isso.

Isto posto, considerando as empresas como pessoas jurídicas, impactadas por uma enorme quantidade de fatos e atos a cada segundo, é possível que eventos raros ocorrerão, seja por erro, fraude ou uma exceção normal. Desse modo, a avaliação da capacidade da identificação desses eventos, utilizando a lei dos grandes números ou a de *Littlewood,* não deve ser menosprezada, mesmo que seja usada dar motivação aos auditores da possibilidade de existirem exceções.

A **análise de regressão**, por sua vez, é uma ferramenta que também pode ser utilizada para identificar exceções em bases de dados por meio de uma análise mais preditiva ou ainda pela sua capacidade de gerar novos *insights* para o negócio. Pela regressão, exceções podem ser identificadas, por exemplo, por meio do cálculo **distância de Cook**, que detecta valores influentes que alteram a reta de regressão. Em relação a *insights*, segundo Cascarino (2017, p.94), as técnicas de regressão podem apoiar o julgamento e recomendação dos auditores, apoiando a lucratividade da companhia, por ex. Cascarino exemplifica que um auditor que esteja auditando a estratégia da alta administração que prevê que a aquisição de alguma nova planta pode aumentar a lucratividade da companhia, pode, por meio de um modelo de regressão, encontrar uma correlação negativa entre o tamanho da planta e as receitas da companhia.

Por fim, as **séries temporais** também podem apoiar a identificação de comportamentos discrepantes. Segundo Morettin e Toloi (2018, p.1), uma série temporal é qualquer conjunto de observações ordenadas no tempo e, a modelagem de uma série temporal, permite:

1. Fazer previsões de valores futuros da série, seja de curto ou longo-prazo;
2. Descrever o comportamento da série, por meio de verificação de tendências, ciclos e variações sazonais;
3. Buscar periodicidades relevantes nos dados.

No contexto de energia elétrica, a utilização das séries temporais, como ferramenta de combate perdas das distribuidoras de energia, pode ser de grande utilidade, uma vez que é possível mapear o comportamento passado e prever o consumo futuro de energia dos clientes. Variações relevantes, nessa análise, podem dar indícios de perdas, seja por irregularidades ou não.

Seja qual o objetivo, Morettin e Toloi (2018, p. 4) salientam que os modelos de séries temporais devem ser simples e parcimoniosos, ou seja, o número de parâmetros envolvidos deve ser o menor possível, e, se possível, sua utilização não deve apresentar dificuldades às partes interessadas em utilizá-los.

METODOLOGIA

Delimitado o tema, objetivos e a justificativa, apresenta-se os procedimentos metodológicos a serem utilizados neste trabalho. No presente estudo, o tipo de pesquisa que irá ser realizada, quanto aos objetivos será a **descritiva**. Segundo Andrade (2002), apud *Beuren* (p. 81, 2012), "*a pesquisa descritiva preocupa-se em observar os fatos, registrá-los, analisá-los, classificá-los e interpretá-los, e o pesquisador não interfere neles. Assim, os fenômenos do mundo físico e humano são estudados, mas não são manipulados pelo pesquisados"*

Neste sentido, considerando que serão utilizados dados das distribuidoras **Companhia Energética de Minas Gerais - CEMIG, Companhia de Eletricidade do Estado da Bahia - COELBA e ELETROPAULO, atual Enel Distribuição São Paulo,** publicados pela ANEEL, com o objetivo de avaliar a aplicação de técnicas estatísticas para avaliar exceções, a **pesquisa descritiva** é a mais adequada comparativamente à exploratória e à explicativa. Isso porque, a pesquisa exploratória é aplicável quando se há pouco conhecimento sobre o assunto e busca-se entendê-lo com maior profundidade (BEUREN, p. 80), muitas vezes incorporando características inéditas. Quanto a explicativa, a intenção é identificar fatores que determinam ou contribuem para a ocorrência dos fenômenos, ou seja, explicar a razão e o porquê das coisas (BEUREN, p. 82).

Em relação aos procedimentos, a **pesquisa documental** é a que mais se relaciona ao presente trabalho científico. A escolha desse tipo de pesquisa justifica-se porque "*visa selecionar tratar e interpretar a informação bruta, buscando extrair dela algum sentido e introduzir lhe algum valor, podendo, desse modo, contribuir com a comunidade científica a fim de que outros possam voltar a desempenhar futuramente o mesmo papel*" (Silva e Grigolo *apud* Beuren, 2013, p. 89).

Deve-se ressaltar, no entanto, os tipos de documentos a serem utilizados não são provenientes de uma fonte primária, mas de uma secundária ou de "segunda mão". Gil (1999) *apud* Beuren (p.90) define documentos de segunda mão como aqueles que de alguma forma já foram analisados, como: relatórios de pesquisa, relatórios de

empresas, entre outros. Sendo assim, os dados que serão utilizados, disponibilizados pela ANEEL, são oriundos, primariamente, das bases de dados dos equipamentos e sistemas das companhias, e, secundariamente, do site da ANEEL.

Conhecendo a base de dados

Para este trabalho, serão **utilizados dados de 2017 a julho de 2020 das distribuidoras CEMIG**, COELBA e ELETROPAULO, disponibilizados mensalmente pela Agência **Nacional de Energia Elétrica - ANEEL**[1]. A escolha dessas empresas se deu porque:

- São as 3 com maiores quantidades de unidades consumidoras no mês de janeiro de 2020,
- Têm a maior receita bruta no mesmo período,
- Estão distribuídas em 3 estados da federação: Minas Gerais, São Paulo e Bahia.

A ANEEL disponibiliza esse *dataset* com informações de "Número de Unidades consumidoras faturadas", "consumo faturado de energia (MWh)" e "tarifas médias" por região, empresa e "classe de consumo - mensal e anual" a partir de 2003.

Ressalta-se que os dados foram obtidos de fontes secundárias, ou seja, foram coletados no site da agência reguladora ANEEL, que são oriundos das empresas, provenientes de seus equipamentos e sistemas informatizados.

Outra observação importante a se fazer é que os dados de 2020, até o momento da realização deste trabalho, haviam sido apurados até julho, diferentemente dos anos

[1] Disponível no site: https://www.aneel.gov.br/relatorios-de-consumo-e-receita - Consumidores, Consumo, Receita e Tarifa Média – Região, Empresa e Classe de Consumo: http://relatorios.aneel.gov.br/_layouts/xlviewer.aspx?id=/RelatoriosSAS/RelSAMPRegiaoEmp.xlsx&Source=http%3A%2F%2Frelatorios%2Eaneel%2Egov%2Ebr%2FRelatoriosSAS%2FForms%2FAllItems%2Easpx&DefaultItemOpen=1

de 2017, 2018 e 2019 que os dados são até dezembro. Dessa forma, eventuais comparações a serem feitas ano-a-ano devem considerar este fato.

A seguir, será dado detalhamento de cada variável que integra a base de dados. A variável número de "**Unidades Consumidoras – UC's**", segundo a ANEEL (p. 12, 2014):

> Conjunto composto por instalações, equipamentos elétricos, condutores e acessórios, incluída a subestação, quando do fornecimento em tensão primária, caracterizado pelo recebimento de energia elétrica em apenas um ponto de entrega, com medição individualizada, correspondente a um único consumidor e localizado em uma mesma propriedade ou em propriedades contíguas.

Em vista disso, diferente do pensamento comum, as UC's podem ou não estarem vinculadas a um domicílio, mas sim a um ponto de entrega, por exemplo: Postes de iluminação pública, iluminação de uma quadra de esportes, empresas.

O "**consumo faturado de energia**" é a energia consumida pela UC cobrada pela distribuidora com base na diferença de leituras entre o início e o final do período, que normalmente é de 30 dias. Na base de dados, esse consumo é expresso em MegaWatt hora - MWh, mas, na conta de energia, está na base de quilowatt hora - kWh. A ANEEL ainda acrescenta:

> Os dados de receita faturada e receita faturada com tributos (PIS, ICMS e Cofins) se referem às tarifas de aplicação homologadas, incluem o adicional de bandeiras tarifárias e não contêm COSIP/CIP (contribuição para custeio do serviço de iluminação pública).

A variável "**Tarifa média**", por sua vez, é o "preço" cobrado que incide sobre o montante de energia consumida e é a média dos valores dos diversos tipos de clientes existentes de uma mesma classe de consumo. Pode-se citar, como exemplo, os consumidores residenciais que tem pelo menos 4 valores diferentes de tarifas: Residencial, Residencial Baixa Renda, Rural, Tarifa Branca (CEMIG). Sendo assim, os valores devem ser analisados somente como referência para o estudo.

"**Classes de consumo**", por fim, são divisões/categorias utilizadas para classificar os consumidores de energia elétrica. A base de dados detalha os dados pelos seguintes tipos de classe de consumo, cada uma com sua especificidade: Comercial, Consumo

próprio, Iluminação Pública, Industrial, Poder Público Residencial, Rural, Rural Irrigante e Serviço público: Tração elétrica, 'Água, Esgoto e Saneamento' e Aquicultor.

Dito isso, pretende-se proceder a análise das bases de dados utilizando as ferramentas mencionadas, iniciando-se pelos dados consolidados ano-a-ano. Caso sejam identificadas exceções, será feita análise na base de dados detalhada por classes de consumo, visando a identificação de possíveis respostas para as exceções.

Ferramentas Estatísticas

Dentre as ferramentas que serão utilizadas, dentro das análises exploratórias, a **correlação** é de suma importância para identificação de padrões. Por meio dela, é possível **medir a força relativa de uma relação linear entre duas variáveis numéricas**. Os valores para o coeficiente de correlação se estendem desde -1, para uma correlação negativa perfeita, até +1 para uma correlação positiva perfeita (LEVINE, p.117). Esse coeficiente é medido por meio da fórmula:

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{(N-1)s_x s_y}$$

Pela fórmula, verifica-se que o coeficiente de correlação (r) é a medida da covariância padronizada, ou seja, é dividida pelo produto do desvio padrão dos dados. Essa padronização é importante porque, diferente da covariância, a correlação independe da unidade de medida dos dados.

Deve-se ressaltar que a correlação, por si só, não consegue provar a existência de efeito de causalidade, ou seja, que a variação no valor de uma variável tenha causado a variação na outra variável (LEVINE, p117). Desse modo, faz-se necessário aprofundar a análise para identificar o que realmente causou a correlação.

Outra ferramenta a ser explorada, a **Análise de Componentes Principais - ACP** tem uma importante função na redução da dimensionalidade dos dados do *dataset*, uma vez que permite, com pouca perda de informação, criar componentes que **resumem a variação total das *p* variáveis resposta**.

Os **métodos de análise multivariada**, como o ACP, têm como objetivo tomar *p* variáveis resposta ($X_1, X_2, \ldots, X_n$) e, a partir delas, encontrar combinações lineares para produzir variáveis latentes ($Z_1, Z_2, \ldots Z_n$). A ordem de importância dessas variáveis é tal que Var(Z_1) ≥ Var(Z_2) ≥ Var(Z_3) ... ≥ Var(Z_p), sendo Var(Z_j) a variância do j-ésimo componente, ou seu autovalor (SILVA, p.44).

Desse modo, seja X a matriz de *n* observações e *p* variáveis. O primeiro componente principal é uma combinação linear tal que:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p.$$

Sendo a_{ij}o coeficiente associado à importância da j-ésima variável resposta em Z_1.A determinação dos coeficientes é feita por meio da técnica de autovalores e autovetores da matriz de covariância ou de correlação, calculando:

$$(\boldsymbol{R} - \boldsymbol{\lambda}_j \boldsymbol{I})a_j = 0$$

Sendo R é a matriz de correlação das *p* variáveis resposta, λ é o *j-ésimo* autovalor dessa matriz e $\boldsymbol{a_j}$ é o *j-ésimo* autovetor associado, isto é, o conjunto de coeficientes do *j-ésimo* componente principal. Tem se então que $\boldsymbol{Z_1 = X_{a1}}$ são os valores (escores) do primeiro componente (SILVA, p.44).

A **Lei de Newcomb-Benford**, por sua vez, tem na literatura importantes resultados na identificação de fraudes em bases de dados. Por meio de uma frequência esperada dos dígitos, a LNB permite identificar variações que podem ser indícios de fraude ou erro no dataset.

A frequência esperada dos dígitos pode ser apurada por meio do seguinte logaritmo:

FRANK BENFORD

The frequency of first digits thus follows closely the logarithmic relation

$$F_a = \log\left(\frac{a+1}{a}\right), \qquad (1)$$

where F_a is the frequency of the digit a in the first place of used numbers.

TABLE II

OBSERVED AND COMPUTED FREQUENCIES

Natural Number	Number Interval	Observed Frequency	Logarithm Interval	Observed − Computed	Prob. Error of Mean
1	1 to 2	0.306	0.301	+0.005	±0.008
2	2 to 3	0.185	0.176	+0.009	±0.004
3	3 to 4	0.124	0.125	−0.001	±0.004
4	4 to 5	0.094	0.097	−0.003	±0.003
5	5 to 6	0.080	0.079	+0.001	±0.002
6	6 to 7	0.064	0.067	−0.003	±0.002
7	7 to 8	0.051	0.058	−0.007	±0.002
8	8 to 9	0.049	0.051	−0.002	±0.002
9	9 to 10	0.047	0.046	+0.001	±0.003

Figura 1 – Retirado do *paper* de Frank Benford's "The Law of Anomalous Numbers"*, 1938, Domínio público*

Dessa forma, por meio do cálculo do logaritmo de cada dígito, apura-se a frequência percentual esperada dos dígitos. No entanto, caso uma base de dados tenha frequência dos dígitos diferente da prevista da LNB pode ser indício de erro ou fraude, para bases de dados naturalmente geradas.

Para se avaliar a conformidade à LNB, Nigrini (2012) apresenta uma tabela com gradações de conformidade para verificar se a **diferença observada** em relação à Lei é adequada ou não para o primeiro, segundo, dois primeiros e três primeiros dígitos.

Digits	Range	Conclusion
First Digits	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity
Second Digits	0.000 to 0.008	Close conformity
	0.008 to 0.010	Acceptable conformity
	0.010 to 0.012	Marginally acceptable conformity
	Above 0.012	Nonconformity
First-Two Digits	0.0000 to 0.0012	Close conformity
	0.0012 to 0.0018	Acceptable conformity
	0.0018 to 0.0022	Marginally acceptable conformity
	Above 0.0022	Nonconformity
First-Three Digits	0.00000 to 0.00036	Close conformity
	0.00036 to 0.00044	Acceptable conformity
	0.00044 to 0.00050	Marginally acceptable conformity
	Above 0.00050	Nonconformity

Tab. 1.1 - Mean Absolute Deviation para LNB. Fonte Nigrini (p.160, 2012)

Essas diferenças são denominadas pelo autor como MAD – *Mean Absolute Deviation*, traduzido como "Média dos Desvios Absolutos". Segundo Nigrini, os MAD são semelhantes ao "*Mean Absolute Percentage Error*" que é usado em análise de séries temporais para medir acurácia do erro em percentagem. Assim, para considerar conformidade total à LNB, o MAD deve estar entre 0.000 a 0.006, quando avaliamos o primeiro dígito. Caso seja maior que 0.015, é considerado não conforme. Salienta-se que, caso o MAD seja não conforme, não necessariamente deve-se considerar que há fraude ou erro na base de dados, apenas que deve haver um maior aprofundamento na análise.

Os **modelos de regressão**, conforme dito no referencial teórico, é de grande uso para geração de *insights* para auditores. A modelagem das variáveis dependentes e independentes numa equação de regressão permite, a partir das estimativas dos parâmetros, determinar como uma variável independente (X) exerce, ou parece exercer, influência sobre outra variável (Y), dependente (CECON, p.122):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Sendo,

β_0 = intercepto de Y para a população
β_1 = inclinação da população

ε_i = erro aleatório em Y para observação *i*
Y_i = variável resposta para obs. *i*
X_i = variável independente para obs. *i*

Essa modelagem pode ocorrer quando uma única variável independente numérica, X, é utilizada para prever a variável dependente numérica Y, caracterizando a regressão linear simples. Por outro lado, quando diversas variáveis independentes são utilizadas para prever uma única variável dependente numérica, seria o caso de um modelo de regressão múltipla. (LEVINE, p. 482)

No entanto, existem limitações. Isso porque, uma vez que a regressão depende do comportamento históricos para modelagem, o histórico pode reduzir o poder de detecção em relação a variados tipos de fraudes que utilizarem novos mecanismos ou métodos ainda não conhecidos. E, sendo assim, esses tipos de fraude não estarão incluídos no banco de dados histórico de casos de fraudes, a partir do qual o modelo preditivo aprendeu (BAESENS, p.21).

Ainda nesse contexto de regressão linear, a **distância de Cook** também pode prestar um serviço importante quando o objetivo é identificar exceções. A distância de Cook identifica os valores influentes que alteram os resultados da análise de regressão (KASSAMBARA, p. 51), por meio da distância entre os coeficientes calculados com e sem a *i* ª observação:

$$D_i = \frac{e_i^2}{ps^2}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{(\boldsymbol{b}-\boldsymbol{b}_{(i)})'\mathbf{X}'\mathbf{X}(\boldsymbol{b}-\boldsymbol{b}_{(i)})}{ps^2}.$$

Notação:

D_i	distância de Cook
e_i	$_i$ ° resíduo
h_i	$_i$ ° elemento diagonal de
p	número de parâmetros do modelo, incluindo a constante
s^2	quadrado médio do erro
b	vetor do coeficiente
$b_{(i)}$	vetor de coeficientes calculados depois de excluir a i a observação
X	matriz de planejamento

A regra geral é que uma observação tem grande influência se a distância de Cook exceder 4/(n – p – 1), onde n é o número de observações e p o número de variáveis preditoras (KASSAMBARA, p. 51, 2012). A distância de Cook é importante porque demonstra que não são todos *outliers* ou valores extremos serão considerados influentes na análise de regressão linear, mas apenas aqueles que ultrapassarem o limite. Isso pode trazer um maior foco no trabalho do auditor.

A aplicação de modelos de **séries temporais** para identificação de comportamentos em dados de consumo de energia elétrica é uma das formas que as empresas de energia elétrica utilizam para identificarem irregularidades.

Uma série temporal pode ser expressa da seguinte forma $Y(t) = [Y_{1(t)}, Y_{2(t)}, \dots, Y_{r(t)}]$, em que cada componente de Y, denota uma variável e diferentes momentos de tempo, ou seja, *r* corresponde ao número de variáveis no tempo *t*.

Conforme dito anteriormente, **as séries temporais tentam prever o comportamento futuro de uma variável a partir de seu comportamento passado,** isto é, objetiva prever valores futuros de determinada variável utilizando dados históricos, ao invés de construir modelos de causa e efeito (MARGARIDO, p. 11, 2020). Desse modo, após a modelagem de uma série temporal em determinada base de dados, é possível comparar as previsões feitas com os valores reais da série e, assim, identificar possíveis exceções, ou ainda, antecipar futuras exceções.

Modelo mais conhecido de série temporal, o **Modelo Autorregressivo Integrado de Média Móvel - ARIMA**, segundo Margarido (p.30), procura explicar o comportamento presente e futuro de uma variável com base nos seus próprios valores passados, também denominados de parâmetros autorregressivos (AR) e seu próprio erro presente e passados, chamados de parâmetros de médias móveis (MA).

Matematicamente, o ARIMA é representado como: $\nabla^d \tilde{y}_t = \frac{1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p} = \frac{\theta(B)}{\phi(B)} a_t$, onde o termo $\nabla^d \tilde{y}_t$ representa a variável diferenciada e centrada em relação à sua própria média, enquanto, $\frac{\theta(B)}{\varphi(B)}$ são os

polinômios dos operadores de médias móveis de ordem (q) e autorregressivo de ordem (*p*), respectivamente.

Margarido (p.49) resume que a estratégia para se trabalhar com modelo ARIMA consiste em quatro fases: 1) identificação; 2) estimação; 3) verificação e 4) previsão. Os três primeiros passos são indicados para aqueles que objetivam unicamente fazer uma análise estrutural. O quarto passo, somente é necessário para aqueles que desejam fazer previsões da série temporal, objetivo deste trabalho.

Experimentações

Serão realizadas experimentações com **valores simulados** para fins de avaliação da capacidade das técnicas citadas anteriormente de identificar exceções, utilizando os dados consolidados e os segregados por classe de consumo.
Para os dados segregados, serão realizadas as seguintes simulações na variável **consumo** no mês de agosto de 2019:

- Classe **Residencial**: valor **duplicado** em relação a julho;
- Classe **Comercial**: valor **limítrofe**, isto é, final 9 em relação a maior grandeza a julho. Ex. Se em julho o consumo foi 380.640,75, agosto será simulado como 399.999,99
- Classe **Industrial**: valor **arredondado** para cima em relação a julho. Ex. Se em julho o consumo foi 380.640,75, agosto será simulado como 400.000,00.

Essas classes (Residencial, Comercial e Industrial) foram escolhidas porque são as que tem maiores consumos dentre as demais. Em relação aos critérios valor duplicado, limítrofe e arredondamento, são alguns padrões, definidos por Nigrini, como mais usados por fraudadores de bases de dados.

As técnicas serão aplicadas e, em seguida, será feita a análise da eficácia das ferramentas na identificação desses valores simulados. Em resumo, os tratamentos serão demonstrados a seguir:

Empresa	Tratamento	Consumo de Energia Elétrica em MWh	Mês/Ano	Classe
CEMIG-D	Duplicado	815.076,24	agosto-19	Residencial
CEMIG-D		815.076,24	julho-19	Residencial
CEMIG-D	Limítrofe	399.999,99	agosto-19	Comercial, Serviços e Outros
CEMIG-D		380.640,75	julho-19	Comercial, Serviços e Outros
CEMIG-D	Arredondamento	200.000,00	agosto-19	Industrial
CEMIG-D		185.945,89	julho-19	Industrial
COELBA		253.752,77	julho-19	Comercial, Serviços e Outros
COELBA	Limítrofe	299.999,99	agosto-19	Comercial, Serviços e Outros
COELBA		122.719,45	julho-19	Industrial
COELBA	Arredondamento	200.000,00	agosto-19	Industrial
COELBA		559.880,13	julho-19	Residencial
COELBA	Duplicado	559.880,13	agosto-19	Residencial
ELETROPAULO		779.724,01	julho-19	Comercial, Serviços e Outros
ELETROPAULO	Limítrofe	799.999,99	agosto-19	Comercial, Serviços e Outros
ELETROPAULO		218.340,27	julho-19	Industrial
ELETROPAULO	Arredondamento	300.000,00	agosto-19	Industrial
ELETROPAULO		1.289.223,70	julho-19	Residencial
ELETROPAULO	Duplicado	1.289.223,70	agosto-19	Residencial

Tab 1.2 – Simulações nos dados por classe no consumo de agosto/19

Para os dados consolidados, serão aplicados os mesmos tratamentos de julho em relação a agosto, a saber:

Empresa	Tratamento	Consumo	Mês/Ano
CEMIG-D		2.021.414,89	julho-19
CEMIG-D	Duplicado	2.021.414,89	agosto-19
COELBA		1.343.975,13	julho-19
COELBA	Arredondado para baixo	1.000.000,00	agosto-19
ELETROPAULO		2.482.869,52	julho-19
ELETROPAULO	Limítrofe	2.999.999,99	agosto-19

Tab 1.3 – Simulações nos dados consolidados consumo de agosto/19

ANÁLISE DE RESULTADOS

Para execução das análises a seguir, foi utilizado o software estatístico R que supre a necessidade de uma ferramenta informatizada de análise estatística *open source* e que abrange várias áreas do conhecimento. Adicionalmente, para algum tratamento mais rápido, também foi utilizado o Microsoft Excel.

Análise Exploratória: Média, mediana e desvio padrão

A análise exploratória envolve aplicação de medidas de tendência central para conhecimento dos dados em questão. Neste tópico, serão avaliadas a média, mediana, moda e desvio padrão, a fim de se identificar possíveis exceções.

a. Moda

O primeiro teste foi a aplicação da **moda** na base de dados. Nessa avaliação, foram identificados valores duplicados para todas empresas deste estudo:

Empresa	Consumo de Energia Elétrica em MWh	Receita de Fornecimento de Energia Elétrica	Receita de Fornecimento de Energia Elétrica com Tributos	Número de Unidades Consumidoras	Tarifa Média de Fornecimento	Tarifa Média de Fornecimento com Impostos	Mês	Ano	Mês/Ano
CEMIG-D - CEMIG D	2112124,34	1086248009	1539551197	8434516	514,29	728,91	3	2019	01/03/2019
CEMIG-D - CEMIG D	2112124,34	1086248009	1539551197	8434516	514,29	728,91	4	2019	01/04/2019

Tab. 2 – Moda nos dados consolidados das concessionárias. Autoria própria

Nos dados consolidados, conforme demonstra a tabela acima, a CEMIG D teve seus dados duplicados de março e abril de 2020. Identificada esta situação, foi necessário consultar novamente a base original da ANEEL e verificou-se que o erro foi de extração da base de dados.

Outra exceção identificada, foi na empresa COELBA. Semelhante à CEMIG D, todos os dados do mês de setembro estavam duplicados, exceto a coluna Classe. Numa nova consulta à base da ANEEL, foi identificado erro também na extração dos dados (Tab 2.1).

Empresa	Consumo de Energia Elétrica em MWh	Receita de Fornecimento de Energia Elétrica	Receita de Fornecimento de Energia Elétrica com Tributos	Número de Unidades Consumidoras	Tarifa Média de Fornecimento	Tarifa Média de Fornecimento com Impostos	Mes	Ano	Mês/Ano	Classe
COELBA - COMPAN	1295,38	703767,64	1065657,6	521	543,29	822,66	7	2019	01/07/2019	Consumo Próprio
COELBA - COMPAN	1295,38	703767,64	1065657,6	521	543,29	822,66	7	2019	01/07/2019	Iluminação Pública

Tab. 2.1 – Moda nos dados detalhados Coelba. Autoria própria

Por fim, para a empresa ELETROPAULO, único valor duplicado foi na coluna consumo de energia de 1,16 MWh. Após investigação na base de dados original, não foi identificado erro de extração na base e o valor foi uma coincidência que pode ser confirmada pela multiplicação entre tarifa média e o consumo. O resultado será exatamente o valor da Receita de Fornecimento (Tab. 2.2).

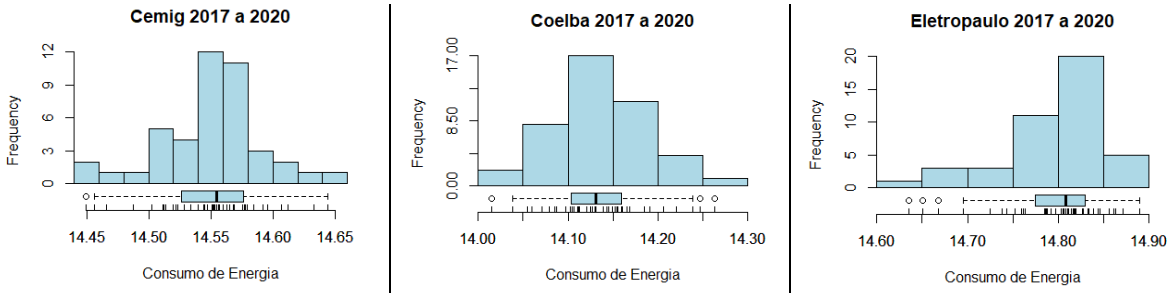
Empresa	Consumo de Energia Elétrica em MWh	Receita de Fornecimento de Energia Elétrica	Receita de Fornecimento de Energia Elétrica com Tributos	Número de Unidades Consumidoras	Tarifa Média de Fornecimento	Tarifa Média de Fornecimento com Impostos	Mes	Ano	Mês/Ano	Classe
ELETROPAULO -	1,16	248,62	327,41	1	214,33	282,25	7	2019	01/07/2019	Rural Aquicultor
ELETROPAULO -	1,16	256,64	314,16	1	221,24	270,83	4	2020	01/04/2020	Rural Aquicultor

Tab. 2.2 – Moda nos dados detalhados Eletropaulo. Autoria própria

b. Média, Mediana e Desvio padrão

Neste tópico, demonstra-se a graficamente a **média, mediana e desvio padrão** do consumo de energia, normalizado pelo logaritmo, por meio de um histograma e de um *boxplot* integrados, segregados por empresa de 2017 a 2020. A normalização foi necessária para permitir a comparação entre as empresas.

Conforme pode-se verificar na tabela 2.3, o histograma da Cemig tem um formato mais simétrico que os demais, com a média e mediana bem próximas, concentração de elementos no centro e uma grande amplitude entre os valores máximos e mínimos, demonstrada pelo *boxplot* horizontal. A Coelba também tem o histograma com um formato de sino, mais simétrico com a mediana levemente à esquerda da média. Por sua vez, a Eletropaulo tem um gráfico mais assimétrico à esquerda, com a mediana bem à direita da média.



Data	Consumo
05/2020	1.884.495

Data	Consumo
06/2020	1.221.067
12/2019	1.564.620
03/2019	1.539.320

Data	Consumo
05/2020	2.269.055
06/2020	2.304.675
04/2020	2.346.284

Tab. 2.3 – Matriz de gráficos de consumo de energia de 2017 a 2020. Autoria própria

Os gráficos de *boxplot* demonstram alguns elementos posicionados a mais de 3 desvios-padrão da média, isto é, valores extremos acima do conjunto de 99,7%.

Pelas datas que ocorreram os consumos, observa-se valores extremos para Cemig, Coelba e Eletropaulo em 2020, ocorridos em função da pandemia da Covid-19 que acarretou redução do consumo de energia, principalmente na classe comercial e serviços.

Por outro lado, a **Coelba** é a única que apresenta valores elevados de consumos na classe Residencial, em março e dezembro de 2019, de **1.564.620 e 1.539.320 MWh**, respectivamente. Esse fato é pode ser justificado porque, sendo a Coelba a distribuidora que atende o litoral nordestino, o consumo de energia tende a ser mais elevado nos períodos de início e final de ano.

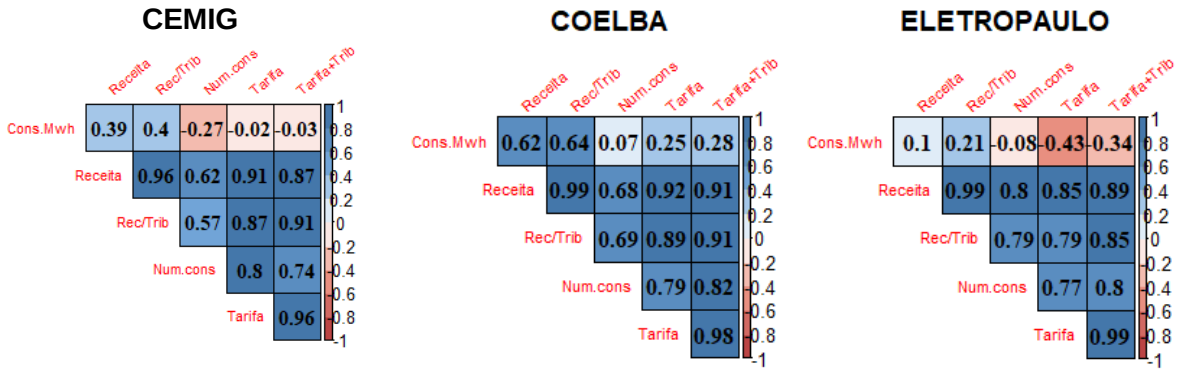
Correlação de Pearson

O coeficiente de correlação de Pearson mede a força relativa de uma relação linear entre duas variáveis contínuas. Neste estudo, como as variáveis tem unidades distintas, o cálculo do coeficiente de correlação é mais adequado que apurar a covariância.

Sendo assim, a primeira análise feita será a correlação entre as variáveis:

a. Consumo de Energia Elétrica – Consumo,
b. Receita de Fornecimento de Energia Elétrica – Receita,
c. Receita de Fornecimento com Tributos – Rec/Trib,
d. Número de Unidades Consumidoras – Num.cons,
e. Tarifa Média – Tarifa,
f. Tarifa Média com Tributos.

Conforme apresentado na tabela abaixo, em todas empresas analisadas, verifica-se correlação positiva forte entre as variáveis Receita, Receita com tributos, Número de consumidores e Tarifa.



Tab. 3.1 – Correlação entre variáveis Consumo x Receita x Receita/Trib/Núm. Consumidores x Tarifa x Tarifa/Trib.

A variável **Consumo – MWh** apresenta algumas correlações negativas que merecem destaque, a saber:

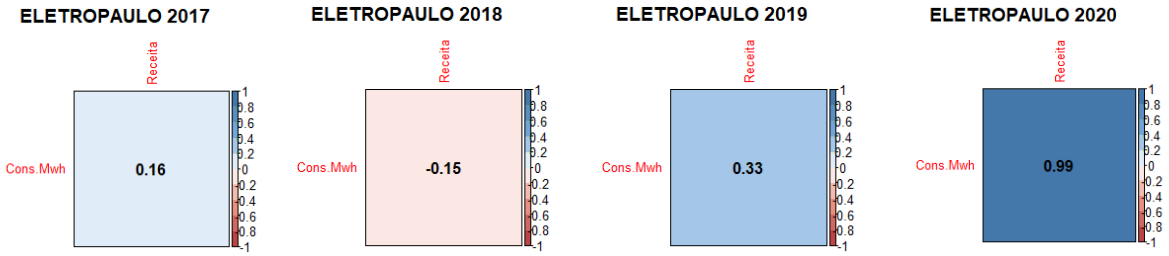
- É acompanhado pelo aumento da Receita para Cemig (0.39) e Coelba (0.62) e de apenas 0.10 para Eletropaulo, isto é, o aumento ou redução do consumo exerce pouca influência na receita da Eletropaulo;
- Negativamente correlacionado com o "Número de Consumidores" para Cemig (-0.27) e quase nulo para Eletropaulo (-0.08) e Coelba (0.07), ou seja, o aumento do número de consumidores é acompanhado pela redução do consumo para Cemig.
- Negativamente correlacionado com Tarifa para Cemig, -0.02, Eletropaulo -0.43. e positivo para Coelba 0.25, ou seja, o aumento do consumo aumenta em conjunto com o aumento da tarifa.

Dessa forma, faz-se necessário aprofundar a análise para as três exceções identificadas:

a. **Eletropaulo:** Correlação positiva de 10% entre Consumo de Energia x Receita;
b. **Coelba:** Correlação positiva 25% entre Consumo x Tarifa;
c. **Cemig:** Correlação negativa de 27% entre Consumo de Energia x Número de UC's.

a. Eletropaulo: Correlação de 0.1 entre Consumo de Energia x Receita

A primeira análise a ser feita é comparar separadamente a correlação dos anos para identificar se há algum indício baixa correlação entre Consumo e Receita da Eletropaulo.



Tab.3.2 – Correlação Eletropaulo Consumo x Receita separado por ano

Na tabela 4, somente em 2018 houve correlação negativa: -0.15. Por outro lado, nos demais anos é positiva e, em 2020, é positiva quase perfeita de 0.99. Faz-se necessário, em função disso, verificar a contribuição das classes de consumo para a correlação de -0.15, em 2018, entre consumo e receita que pode ter contribuído a de 10% do consolidado.

Na análise das 11 classes de consumidores, verifica-se que a classe que tem correlação negativa mais forte é a "Comercial, Serviços e Outros" com -0,24.

Correlação	Comercial Serviços e Outros	Consumo Próprio	Iluminação Pública	Industrial	Poder Público	Residencial	Rural	Rural Aquicultor	Serviço Público água es e san	Serviço Público tração elétrica
Consumo x Receita	-0,24	0,11	0,39	0,40	-0,04	0,19	0,45	0,99	0,68	0,27

Tab. 3.3 – Correlação entre Consumo x Receita de 2018 segregado por classe de consumo (Eletropaulo)

Desse modo, a coeficiente de correlação de -0.24 demonstra que as varáveis Consumo e Receita, para a classe de consumo "Comercial, Serviços e Outros", se movem em direções opostas.

A análise gráfica do consumo, receita e tarifa média do período pode corroborar esta situação na medida em que se observa que a tarifa média aumenta, o consumo de energia diminui juntamente com o aumento da receita:

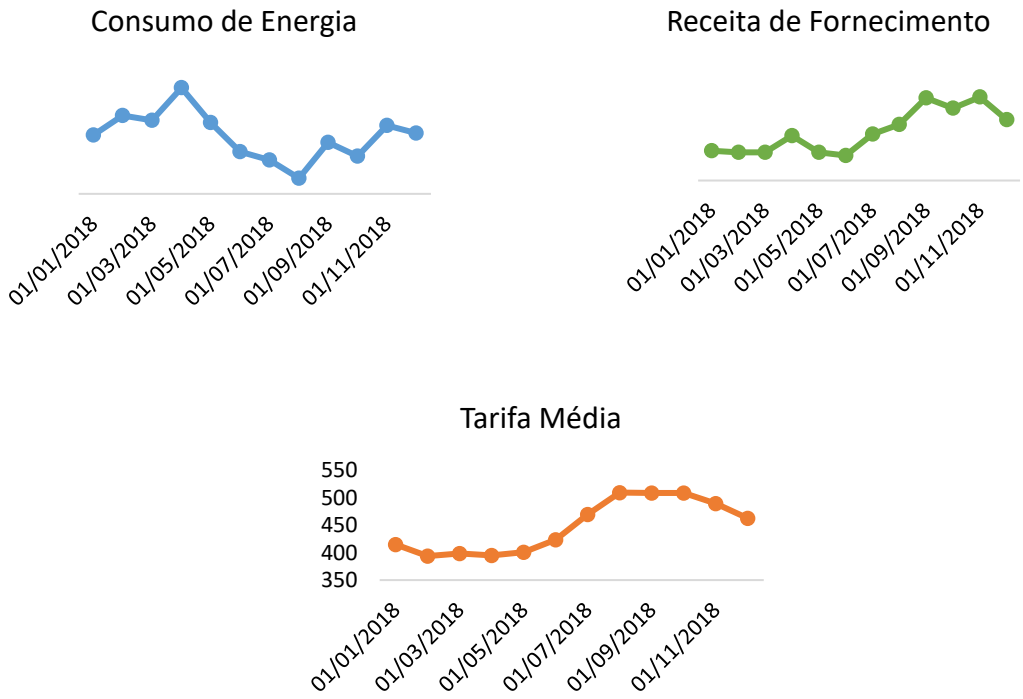
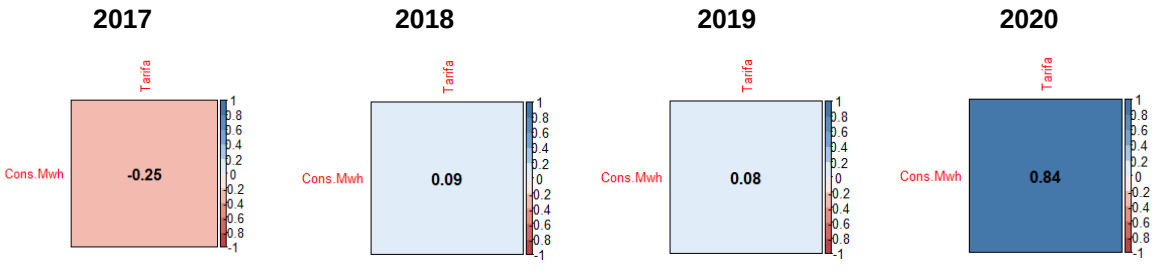


Gráfico 1 – Consumo, Receita e Tarifa no ano de 2018 (Eletropaulo)

b. Coelba: Correlação positiva 25% entre Consumo x Tarifa para Coelba;

Conforme demonstrado na Tabela 1, a Coelba, 0.25, tem a maior correlação entre consumo e tarifa, quando comparado com Cemig, 0.02, e Eletropaulo -0.43, sugerindo que o aumento do consumo é acompanhado positivamente pelo aumento da tarifa.

Neste sentido, apresenta-se a correlação ano-a-ano para uma melhor análise:



Tab. 3.4 – Correlação entre Consumo x Tarifa separado por ano (Coelba)

A partir de 2017, verifica-se que a correlação negativa entre **consumo e tarifa** se aproxima de 0 em 2018 e 2019. Em 2020, no entanto, a correlação torna-se fortemente positiva em 0.84, isto é, o aumento/redução do consumo está diretamente relacionado com o aumento da tarifa. Em função disso, estratifica-se abaixo a correlação entre classes de consumo em 2020:

Comercial Serviços e Outros	Consumo Próprio	Iluminação Pública	Industrial	Poder Público	Residencial	Rural	Rural Aquicultor	Serviço Público água es e san	Serviço Público tração elétrica
-0,43	-0,19	0,76	-0,11	-0,74	0,34	0,31	-0,74	0,42	-0,70

Tab. 3.5 – Correlação entre Consumo x Tarifa de 2020 segregada por classe de consumo (Coelba)

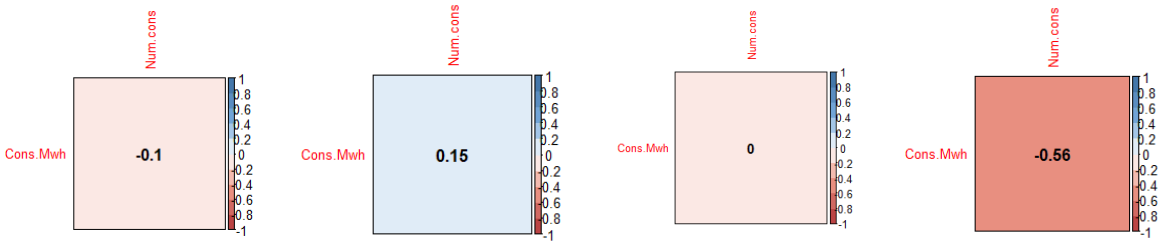
Em 2020, o aumento/redução do consumo das classes Iluminação Pública, 0.76, Residencial, 0.34, Rural, 0.31, e Serviço Público, 0.42, acompanham o aumento/redução da tarifa.

Uma das possíveis explicações para essa correlação da classe residencial pode ser o impacto que a pandemia do Covid-19 acarretou mudanças na vida das famílias. Como as pessoas ficaram mais tempo dentro de casa em função da quarentena, independente do aumento das tarifas, o consumo de energia aumentou. Por outro lado, Iluminação Pública, Rural e Serviço Público deve-se aprofundar os estudos para entender os motivos da correlação positiva.

c. Cemig: Correlação negativa de 27% entre Consumo de Energia x Número de UC's

A Cemig é uma exceção quando se compara a correlação entre **Consumo x Número de Unidades Consumidoras** com Coelba e Eletropaulo. A Cemig tem correlação negativa de 0.27, sendo a Coelba e Eletropaulo quase nulas: 0.07 e -0.08, respectivamente. Abaixo, demonstra-se a correlação segregada ano-a-ano da Cemig:

2017 **2018** **2019** **2020**



Tab. 3.6 – Correlação entre Consumo x Quantidade de UC's ano-a-ano (Cemig)

Pela tabela 6, percebe-se que somente em 2018, é que a correlação entre **consumo e quantidade de unidades consumidoras** foi positiva. Em 2020, por sua vez, a correlação é fortemente negativa, sugerindo que o aumento/diminuição do número de consumidores é acompanhado inversamente pelo consumo de energia. Faz-se necessário analisar o ano de 2020 segregado por classe de consumo:

Comercial Serviços e Outros	Consumo Próprio	Iluminação Pública	Industrial	Poder Público	Residencial	Rural	Rural Aquicultor	Serviço Público água es e san	Serviço Público tração elétrica
0,80	-0,05	0,63	0,57	0,11	-0,88	0,04	0,93	0,12	NA

Tab. 3.7 – Correlação entre Consumo x Quantidade de UC's por classes de consumo para Cemig em 2020

A classe residencial, -0.88, destaca-se demonstrando que o aumento na quantidade de consumidores tem relação inversa com o consumo de energia. No gráfico abaixo, apresenta-se a série temporal, para os meses de 2020 desta situação:

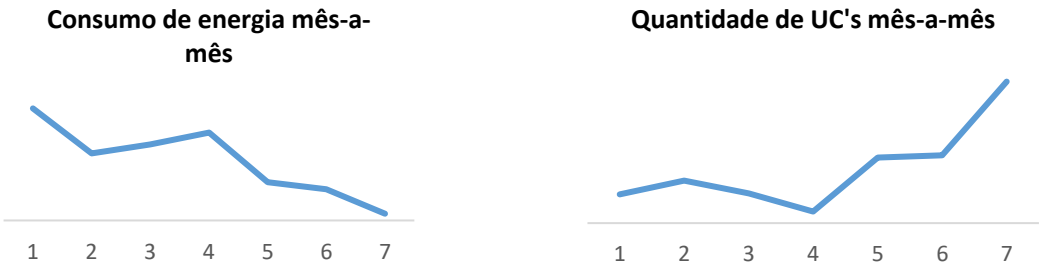


Gráfico 2 – Consumo de energia e Quantidade de UC's da Cemig em 2020 para classe residencial (Cemig)

Sugere-se aprofundar as análises para verificar a causa desta situação.

Simulação para análise exploratória

Conforme mencionado na metodologia, simulou-se consumos de energia para as distribuidoras de energia do presente estudo, de forma a se avaliar a capacidade de identificação de exceções das técnicas. A simulação foi feita no mês de agosto e considerou para a classe Residencial valores duplicados, valores limítrofes para classe Comercial e valores arredondados para classe Industrial, todos em relação ao mês de julho.

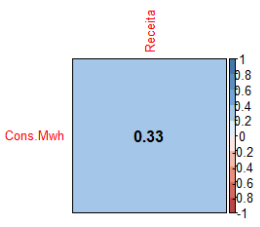
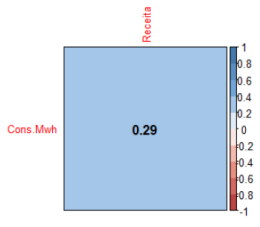
A aplicação da **moda**, nos dados segregados por classe, identificou todos valores simulados duplicados do *dataset*.

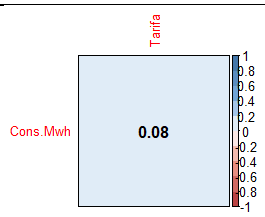
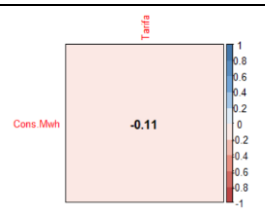
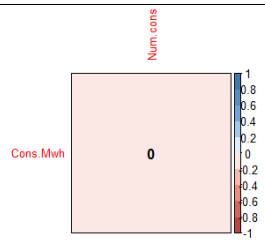
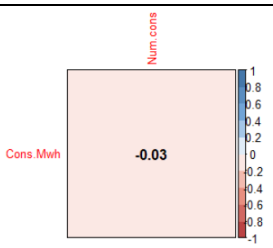
Empresa	Consumo	Mês.Ano
Cemig	815.076,24	01/08/2019
Cemig	815.076,24	01/07/2019
Coelba	559.880,13	01/07/2019
Coelba	559.880,13	01/08/2019
Eletro	1,16	01/07/2019
Eletro	1,16	01/04/2020
Eletro	1.289.223,70	01/07/2019
Eletro	1.289.223,70	01/08/2019

Tab 4 – Moda nos dados simulados de consumo

Em relação aos dados consolidados, a simulação também foi feita no consumo do mês de agosto e considerou valor duplicado para Cemig-D, **arredondado para baixo para Coelba e limítrofe para Eletropaulo**, em relação ao consumo do mês de julho.

Desse modo, a correlação entre consumo x receita foi aplicada sobre os dados simulados e os resultados apresentados o antes e o depois abaixo:

Empresa	Tratamento	Antes	Depois
Eletropaulo	Limítrofe		

Coelba	Arredondamento para baixo		
Cemig-D	Duplicação		

Tab 4.1 - Correlação entre consumo x receita nos dados simulados

Observa-se que a maior sensibilidade foi a correlação da Coelba para o ano de 2019 que saiu de uma correlação positiva 0.08 para negativa de 0.11. Deve-se observar que foi considerado o ano de 2019, apenas 12 meses, e qualquer alteração de valor será sensível, por serem poucos dados. Nesta situação, o consumo foi alterado de 1.301.238,95 MWh para 1.000.000.00 MWh.

Sendo assim, haja vista que as diferenças foram pouco relevantes, não se pode considerar que a correlação foi eficaz em identificar valores alterados.

Lei de Benford

Conforme explicado anteriormente, a Lei de *Newcomb-Benford*, popularmente conhecida como **Lei do Primeiro Dígito**, diz que, para bases de dados, naturalmente geradas e não manipuladas, existe uma padronização na frequência dos dígitos. A lei também informa que, além de ser naturalmente geradas, devem ter uma grande quantidade de elementos.

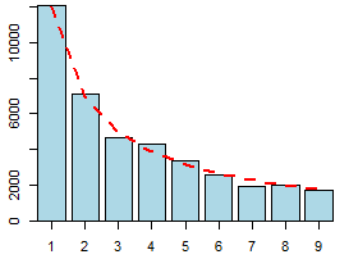
Neste sentido, aplicou-se a Lei de Benford no 1°, 2° e 3° dígitos dos consumos de energia, nos dados segregados por classe consumidora, **de todas distribuidoras**, a fim de avaliar a conformidade em relação à lei e buscar identificar exceções. **Foram utilizadas todas as distribuidoras para que houvesse grande quantidade de dados.**

Antes de apresentar a tabela com os resultados, há de se fazer duas observações a respeito de tratamentos realizados na base de dados:

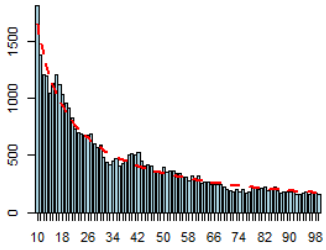
- O consumo de energia da base de dados é medido em MWh, o que traz, para consumos mais baixos, valores decimais, ex.: 0,9. O valor “0” não deve ser considerado na LNB (Nigrini, p. 86, 2012), por isso, optou-se por transformar os valores para kWh, fazendo com que esses consumos sejam considerados.

- Outro tratamento dado foi em valores negativos. Nigrini (p. 86, 2012) afirma que os sinais negativos devem ser desconsiderados, exemplificando na análise dos 2 primeiros dígitos que o valor “-34.83” fica “34”.

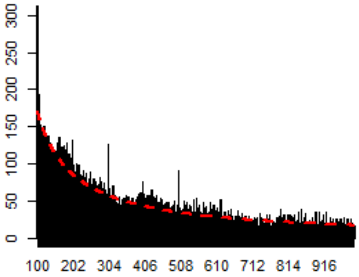
Resultados:



MAD: 0.004624401
MAD Conformity: Close conformity



MAD: 0.0009521214
MAD Conformity: Close conformity



MAD: 0.0001728957
MAD Conformity: Close conformity

Tab. 5 – Conformidade da LNB com o Consumo de Energia. Autoria própria

Tanto para o primeiro, segundo, quanto para o terceiro dígito, o MAD conclui pela estreita conformidade à lei de Benford. Apesar da conformidade, alguns dígitos tiveram variação mais significativa em relação ao padrão esperado:

Primeiro Dígito	
Dígito	absolute.diff
4	422.14
7	340.38
3	296.06
5	240.54
9	89.62

Dois primeiros dígitos	
Dígito	absolute.diff
12	177.11
10	162.07
16	156.42
14	149.1
17	127.32

Três primeiros dígitos (ordem decrescente)	
Dígito	absolute.diff
100	140.06
300	68.50
500	57.48
200	45.82
184	35.34

Tab. 5.1 – Principais diferenças para o 1°, 2° e 3° dígitos, segundo a LNB (ordem decrescente). Autoria própria

A tabela 7.1 destaca, em ordem decrescente, as principais diferenças para o 1°, 2° e 3° dígitos. Quando se verifica quais consumos iniciam pelo **1° e 2° dígitos** (tabela 1 e 2), surgem diversos valores de consumo que podem ter sua análise aprofundada, servindo como uma forma de amostragem alternativa à tradicional. Todavia, quando se verifica os **3 primeiros dígitos** (tabela 3), a amostra diminui e permite concentrar esforços em menos elementos da população, podendo aumentar a assertividade.

Este teste, analisando-se os 3 primeiros dígitos, sugere alguns indícios para que os auditores aprofundem as análises, a saber:

- Análise dos microdados de Consumo, Receita e Quantidade de UC's dos dígitos indicados;
- Análise dos microdados das Classes de consumo e períodos que mais aparecem.

a. Teste do 3° dígito da LNB para CEMIG

Empresa	Consumo de Energia Elétrica em MWh	3 Dígitos	Receita de Fornecimento de Energia Elétrica	Número de Unidades Consumidora	Tarifa Média de Fornecimento	Mês/Ano	Classe
CEMIG-D - CEMIG DISTRIBUI	100.967,32	100	41.425.549,67	12848	410,29	01/05/2017	Serviço Público (água, esgoto e saneamento)
CEMIG-D - CEMIG DISTRIBUI	184.390,35	184	65.865.735,26	681042	357,21	01/04/2017	Rural
CEMIG-D - CEMIG DISTRIBUI	184.655,09	184	64.868.021,41	681769	351,29	01/12/2017	Rural
CEMIG-D - CEMIG DISTRIBUI	184.930,62	184	64.804.880,31	690622	350,43	01/05/2018	Rural
CEMIG-D - CEMIG DISTRIBUI	200.192,36	200	99.263.164,07	72273	495,84	01/01/2019	Industrial
CEMIG-D - CEMIG DISTRIBUI	200.281,58	200	99.126.902,16	29793	494,94	01/06/2019	Industrial
CEMIG-D - CEMIG DISTRIBUI	200.068,44	200	100.293.672,40	687158	501,3	01/08/2019	Rural
CEMIG-D - CEMIG DISTRIBUI	300.779,63	300	186.954.025,60	772846	621,56	01/07/2020	Comercial, Serviços e Outros
CEMIG-D - CEMIG DISTRIBUI	3.006,60	300	1.455.752,18	749	484,19	01/05/2018	Consumo Próprio

Tab. 5.2 – Teste 3° dígito da LNB para CEMIG. Autoria própria

Utilizando-se como base os dígitos que apresentaram maiores divergências em relação à LNB, a classe que mais apareceu foi a Rural, 4 aparições, segunda pela Industrial, 2. Em relação ao consumo, a classe "Comercial, Serviços e Outros" supera os demais com uma medição de 300.799 MWh. Por outro lado, a classe "consumo próprio" teve indicação do dígito 300, com o menor consumo que as demais. Para os períodos, existem elementos mais no ano de 2017 e 2019 e, em relação aos meses, o mês de maio aparece 3 vezes.

b. Teste do 3° dígito da LNB para COELBA

Empresa	Consumo de Energia Elétrica em MWh	3 Dígitos	Receita de Fornecimento de Energia Elétrica	Número de Unidades Consumidora	Tarifa Média de Fornecimento	Mês/Ano	Classe
COELBA - COMPANHIA DE E	100.370,55	100	46.686.915,68	13754	465,15	01/01/2020	Industrial
COELBA - COMPANHIA DE E	100.286,26	100	31.635.930,68	17512	315,46	01/07/2019	Iluminação Pública
COELBA - COMPANHIA DE E	100.519,75	100	23.720.099,96	18432	235,97	01/03/2019	Rural Irrigante
COELBA - COMPANHIA DE E	100.390,24	100	21.088.244,80	13240	210,06	01/11/2017	Rural Irrigante
COELBA - COMPANHIA DE E	1.007,49	100	389.478,52	513	386,58	01/09/2019	Rural Aquicultor
COELBA - COMPANHIA DE E	50.066,47	500	18.967.246,38	214582	378,84	01/04/2019	Rural
COELBA - COMPANHIA DE E	50.017,01	500	21.211.859,55	210297	424,09	01/05/2020	Rural

Tab. 5.3 – Teste 3° dígito da LNB para COELBA. *Autoria própria*

A Coelba tem grande frequência do dígito 100 em relação ao consumo de energia. As classes que mais aparecem são Rural Irrigante (2), Rural (2), Rural Aquicultor (1), Industrial (1) e Iluminação pública (1). Dos 7 períodos apresentados, o ano de 2019 aparece 4 vezes. Interessante notar que as variações da classe rural (Rural, Irrigante e Aquicultor) aparecem 70% das vezes.

c. Teste do 3° dígito da LNB para ELETROPAULO

Empresa	Consumo de Energia Elétrica em MWh	3 Dígitos	Receita de Fornecimento de Energia Elétrica	Número de Unidades Consumidora	Tarifa Média de Fornecimento	Mês/Ano	Classe
ELETROPAULO - ELETROPAU	1.008.637,24	100	456.757.845,60	403108	452,85	01/02/2019	Comercial, Serviços e Outros
ELETROPAULO - ELETROPAU	100.058,33	100	50.014.218,57	15838	499,85	01/09/2018	Poder Público
ELETROPAULO - ELETROPAU	100.701,01	100	47.172.753,28	15673	468,44	01/02/2020	Poder Público
ELETROPAULO - ELETROPAU	100.172,75	100	37.975.512,72	15887	379,10	01/02/2018	Poder Público
ELETROPAULO - ELETROPAU	100.440,86	100	36.671.874,66	15892	365,11	01/01/2017	Poder Público
ELETROPAULO - ELETROPAU	100.052,03	100	33.639.227,58	15934	336,22	01/05/2017	Poder Público
ELETROPAULO - ELETROPAU	10.069,79	100	4.225.137,04	107	419,59	01/03/2020	Serviço Público (tração elétrica)
ELETROPAULO - ELETROPAU	1.009,53	100	403.895,19	540	400,08	01/02/2020	Rural
ELETROPAULO - ELETROPAU	1.008,76	100	393.852,08	510	390,43	01/09/2018	Rural
ELETROPAULO - ELETROPAU	1.008,02	100	366.562,49	469	363,65	01/07/2018	Rural
ELETROPAULO - ELETROPAU	1.000,43	100	340.335,27	523	340,19	01/04/2019	Rural
ELETROPAULO - ELETROPAU	3.005,24	300	796.768,96	294	265,13	01/08/2017	Consumo Próprio

Tab. 5.4 – Teste 3° dígito da LNB para ELETROPAULO. Autoria própria

A Eletropaulo também tem o dígito 100 aparecendo com grande frequência. Para a classe “Serviço Público (tração elétrica)”, a pequena quantidade de “Unidades Consumidoras”, 107, pode permitir ao auditor um teste aprofundado, apesar dessa classe se referir a transporte ferroviário. Por outro lado, a classe “Comercial, Serviços e Outros” tem que ser analisada com prioridade, porque a receita auferida supera as demais com sobra R$ 456 milhões. Em relação às classes, a “Poder Público” e novamente a “Rural” podem ter uma análise complementar, porque aparecem com mais frequência, 5 e 4 vezes, respectivamente.

Um dos benefícios da aplicação da LNB é que a o resultado sugere valores que nem sempre são valores extremos (máximos ou mínimos), o que, pelos métodos tradicionais exploratórios, não seriam consideradas exceções. De toda forma, a Lei de Newcomb-Benford é apenas um indicativo de risco elevado de fraude ou erro e a conformidade ou não da base de dados precisa ser acompanhada sempre de uma análise mais aprofundada.

Simulação Lei de Benford

Conforme mencionado na metodologia, simulou-se consumos de energia para as distribuidoras de energia do presente estudo, de forma a se avaliar a capacidade de identificação de exceções das técnicas. A simulação foi aplicada no mês de agosto e considerou, em relação ao mês de julho:

- Para a classe Residencial valores **duplicados**,
- Valores **limítrofes** para classe Comercial,
- Valores **arredondados** para Industrial.

Os três primeiros dígitos com as maiores diferenças absolutas entre a frequência esperada e a frequência real, foram:

Empresa	Classe	Tratamento	Valor de Consumo de Simulado	Dígito da LNB	Posição	Diff. Absoluta (LNB – Real)
Eletropaulo	Industrial	Arredondamento	300.000,00	300	2	69,49676
Cemig	Industrial	Arredondamento	200.000,00	200	4	47,81673
Coelba	Industrial	Arredondamento	200.000,00	200	4	47,81673
Eletropaulo	Residencial	Duplicado	1.289.223,70	128	8	30,47312
Coelba	Comercial	Limítrofe	299.999,99	299	90	15,69524
Cemig	Residencial	Duplicado	815.076,24	815	154	11,8109
Coelba	Residencial	Duplicado	559.880,13	559	376	6,11579
Eletropaulo	Comercial	Limítrofe	799.999,99	799	741	1,61315
Cemig	Comercial	Limítrofe	399.999,99	399	873	1,78215

Tab 6 – Diferenças para os três primeiros dígitos

Observa-se que a lei foi eficaz para identificação de valores arredondados da classe industrial e para valores duplicados, da classe Residencial. Para os demais, duplicados e limítrofes, tiveram posições elevadas que provavelmente não seriam amostrados. Interessante notar que o valor de maior consumo 1.289.223 MW foi indicado como suspeito, correspondendo à maior receita.

A lei do primeiro dígito, face ao exposto, sugere elementos e classes para que sejam direcionadas as análises, sendo bastante útil numa amostragem julgamental pelos auditores.

Regressão linear – Distância de Cook

A distância de Cook mostra a influência de um elemento no modelo de regressão. Esses valores influentes podem ser considerados exceções e, desse modo, analisados com maior profundidade pelos auditores.

Nesse contexto, ajustamos uma regressão linear nos dados de cada distribuidora, tendo como dependente a variável "Consumo" e independente "Quantidade de UC's". Em seguida, calculamos a distância de Cook, a fim de identificar os valores influentes para, posteriormente, aprofundar a análise.

Deve-se salientar que a aplicação de uma regressão múltipla, considerando uma variável mês como dummy, ajustaria melhor o modelo. Contudo, optou-se por realizar o procedimento mais simples de uma regressão linear.

a. Ajuste da regressão e distância de Cook para CEMIG

$$Consumo = 4.055x10^{6} - 0.233\ x\ Quantidade\ de\ UC's$$

Coeficientes:

	Estimate	Std. Error	t	value Pr(>|t|)
(Intercept)	4.055e+06	1.116e+06	3.632	0.000774 ***
Quantidade de UC's	-2.334e-01	1.325e-01	-1.761	0.085628

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC: 1101.846

Multiple R^2: 0.07035, R^2 Ajustado: 0.04767

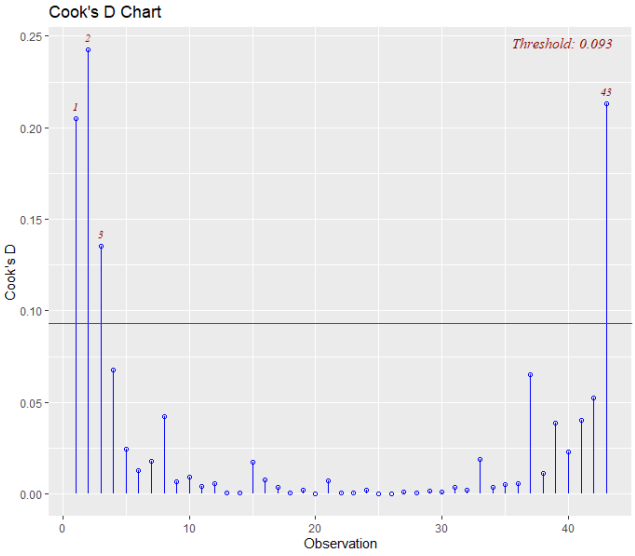
p-value: 0.08563

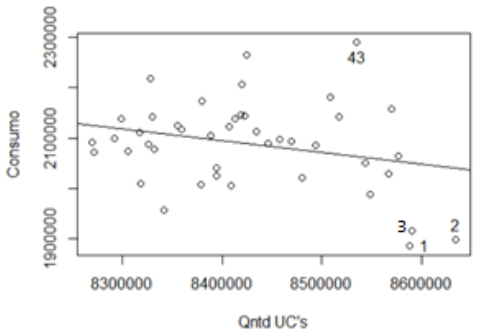
O valor estimado de $\beta 0$ é igual a $4.055x10^{6}$ e o valor de $\beta 1$ a inclinação foi de $-2.334x10^{-1}$, demonstrando que o aumento da quantidade de UC's reduz o consumo. O intercepto foi estatisticamente significativo, uma vez que o *p valor* foi de 0.000774

(menor que nível de significância de α = 0,05), mas o vetor quantidade de UC's não: α = 0,08. O AIC de 1101.846 traduz que o modelo não explica o comportamento das variáveis, assim como o R^2 que também demonstra baixa correlação.

Como o R^2 teve uma explicabilidade muito baixa, seria desnecessário aplicar a distância de Cook sobre os dados. No entanto, optou-se por aplicar para avaliar os resultados.

Distância de Cook





id.	Consumo	Número de UC's	Distância de Cook	Período
2	1897799.	8.634.326	0.243	01/07/2020
43	2291129.	8.534.441	0.213	01/11/2019
1	1884495.	8.588.419	0.205	01/05/2020
3	1916394.	8.589.829	0.135	01/06/2020

Gráfico 3.1 – Distância de Cook para CEMIG (2017 a 2020). Fonte: Autoria própria

No gráfico acima, a distância de Cook é apresentada com limite de 0.093, sendo, este limite, extrapolado pelos elementos 1,2,3 e 43. Esses elementos estão concentrados no ano de 2020 que pode ter ocorrido em função da redução de consumo, devido à pandemia. O item 43, por outro lado, ocorreu em 11/2019, decorre de aumento de consumo e, na análise por classes, a maior influência é da classe Residencial

Apesar dessa identificação, conforme visto anteriormente, o R^2 ajustado do modelo de regressão teve baixa capacidade de explicação da variabilidade dos dados: 0.047. Desse modo, será aplicada a distância de Cook sobre os resíduos da série temporal para, em seguida, comparar os resultados.

Distância de Cook sobre os resíduos da série temporal CEMIG

Um modelo de regressão em séries temporais (ARIMA) será ajustado sobre os resíduos da série temporal em relação ao número de consumidores, afim de se identificar novas possíveis exceções na base de dados analisada. Em seguida, será feita análise dos valores extremos indicados pela distância de *Cook*.

No primeiro momento, avaliou-se a normalidade dos resíduos, por meio do teste de Shapiro:

```
        Shapiro-Wilk normality test

data:  a$residuals
W = 0.98361, p-value = 0.7879
```

Por meio do resultado do teste de Shapiro-Wilk de 0.98, não se rejeita a hipótese de normalidade dos erros, ao nível de 5% de significância. Em seguida, ajustou-se um modelo de série temporal sobre os resíduos, utilizando a função *auto.arima* do pacote *forecast* do R

```
Series: rl.cemig$residuals
ARIMA(0,1,1) with drift

Coefficients:
          ma1     drift
      -0.5913  8060.098
s.e.   0.1243  1883.291

sigma^2 estimated as 8.26e+08:  log likelihood=-489.96
AIC=985.92   AICc=986.55   BIC=991.13
```

O modelo resultou em um ARIMA(0,1,1) e, na análise do gráfico dos resíduos, já é possível identificar alguns elementos extremos:

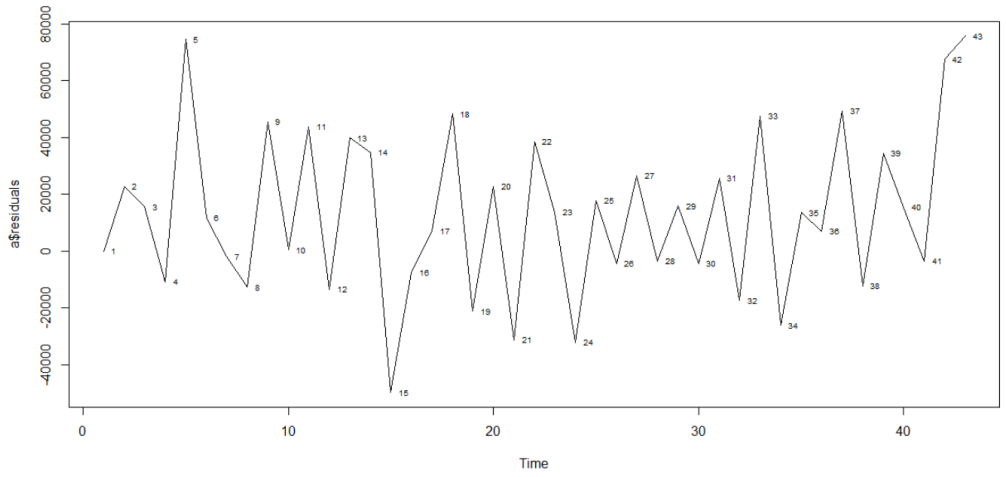


Gráfico 4 – Resíduos série temporal CEMIG – ARIMA(0,1,1)

A distância de Cook aplicada sobre a regressão entre as variáveis resíduos e número de consumidores, como variável independente, gerou as seguintes exceções:

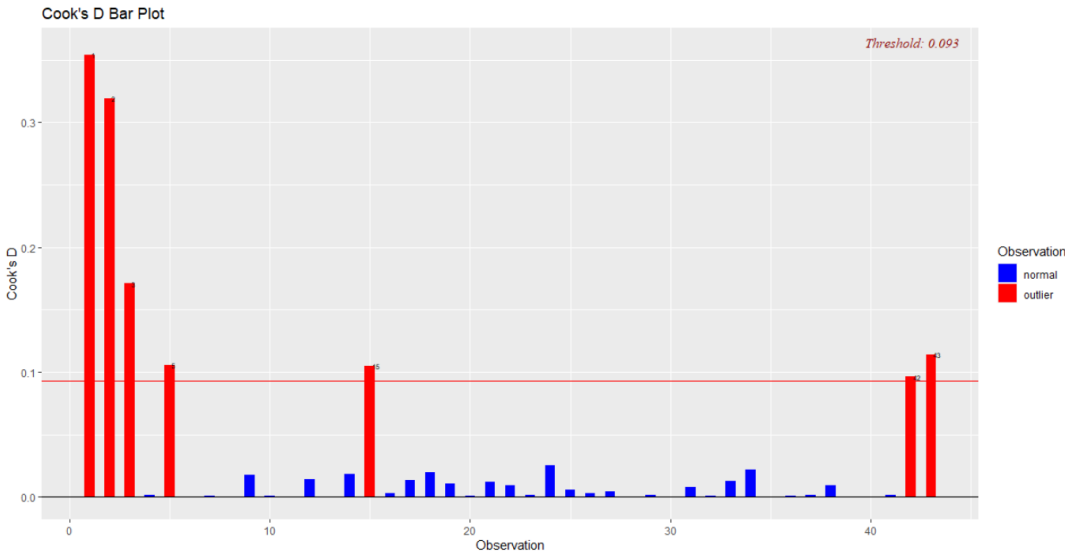


Gráfico 5 – Distância de Cook sobre os resíduos CEMIG

Diferente da anterior, os novos elementos extremos indicados pela distância de Cook foram o 1,2,3,5 e o 15.

id.	Resíduo	Número de UC's	Distância de Cook	Período
1	-165	8.588.419	0,3530	01/05/2020
2	22678	8.634.326	0,3190	01/07/2020
3	15574	8.589.829	0,1710	01/06/2020
5	74557	8.548.161	0,1060	01/04/2020
15	-49850	8.271.274	0,1050	01/02/2017

Tab 7 – Maiores distâncias de Cook sobre resíduos para CEMIG

Observa-se que, com exceção do elemento 15, os demais coincidem com o período da pandemia do coronavírus, abril a julho de 2020.

b. Ajuste da regressão e distância de Cook para COELBA

$$Consumo = 1.045\ x\ 10^6 + 5.537\ x\ 10^2\ x\ Quantidade\ de\ UC's$$

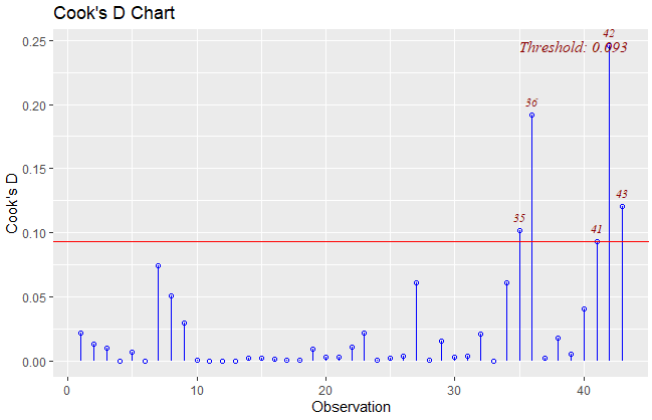
Coeficientes:

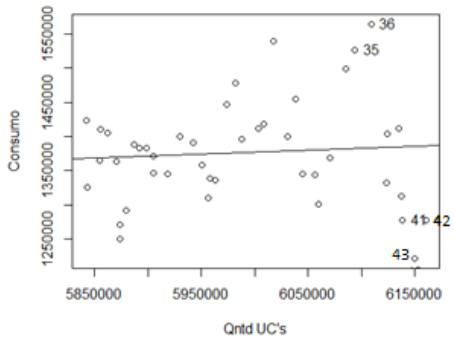
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.045e+06	7.005e+05	1.492	0.143
Quantidade UC's	5.537e-02	1.170e-01	0.473	0.639

AIC: 1092.285

Multiple R^2: 0.005432, R^2 Ajustado: -0.01883
F-statistic: 0.2239 on 1 and 41 DF, p-value: 0.6386

O valor estimado de $\beta 0$ é igual a $1.045x10^6$ e o valor de $\beta 1$ a inclinação foi de $+5.537x10^{-2}$, demonstrando que o aumento da quantidade de UC's aumenta o consumo de energia. No entanto, nenhum dos vetores foram estatisticamente significativos, uma vez que o p valor de ambos foi superior ao nível de significância de α = 0,05. O AIC de 1092.285 traduz que o modelo não explica o comportamento das variáveis, assim como o R2 que também demonstra baixa correlação.





id.	Consumo	Número de UC's	Distância de Cook	Período
42	1221067.	6.149.815	0.246	01/06/2020
36	1564620.	6.108.894	0.192	01/12/2019
43	1276933.	6.160.249	0.121	01/07/2020
35	1526445.	6.093.491	0.102	01/11/2019
41	1277081.	6.138.083	0.0931	01/05/2020

Gráfico 7 – Distância de Cook para COELBA (2017 a 2020). Autoria própria

No gráfico da Coelba, a distância de Cook tem como limite de 0.093, sendo extrapolado pelos elementos 41, 35, 36, 43 e 42, em ordem crescente. Os valores 41, 42 e 43 ocorreram em meados de 2020 e provavelmente ocorreram em função da redução de consumo ocorrido devido à pandemia. Os itens 35 e 36, por outro lado, são devido a aumento de consumo do final do ano, puxado pela classe Residencial.

Apesar dessa identificação, o R^2 ajustado do modelo de regressão teve baixíssima explicação da variabilidade dos dados: -0.018. Desse modo, será aplicada a distância de Cook sobre os resíduos da série temporal e, em seguida, comparar os resultados.

Como o R^2 teve uma explicabilidade muito baixa, seria desnecessário aplicar a distância de Cook sobre os dados. No entanto, optou-se por aplicar para avaliar os resultados.

Distância de Cook sobre os resíduos da série temporal (COELBA)

Um modelo de regressão será ajustado sobre os resíduos da série temporal em relação ao número de consumidores, afim de se identificar novas possíveis exceções na base de dados analisada. Em seguida, será feita análise dos valores extremos indicados pela distância de *Cook*.

No primeiro momento, avaliou-se a normalidade dos resíduos, por meio do teste de Shapiro:

```
        Shapiro-Wilk normality test

data:  a$residuals
W = 0.97849, p-value = 0.5888
```

Por meio do resultado do teste de Shapiro-Wilk de 0.97, não se rejeita a hipótese de normalidade dos erros, ao nível de 5% de significância. Em seguida, ajustou-se um modelo de série temporal sobre os resíduos, utilizando a função *auto.arima* do pacote *forecast* do R

```
Series: rl.coelba$residuals
ARIMA(1,0,0) with zero mean

Coefficients:
         ar1
      0.5608
s.e.  0.1279

sigma^2 estimated as 3.876e+09:  log likelihood=-535.38
AIC=1074.75   AICc=1075.05   BIC=1078.27
```

O modelo resultou em um ARIMA(1,0,0) e na análise do gráfico dos resíduos é possível identificar alguns elementos extremos.

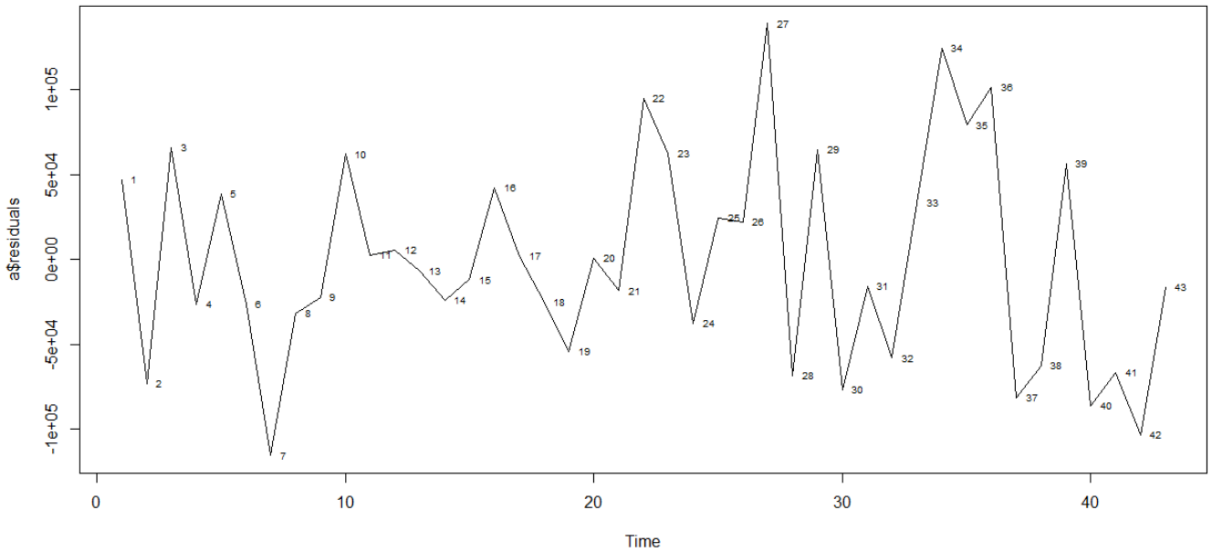


Gráfico 8 – Série Temporal dos resíduos para COELBA

A distância de Cook aplicada sobre a regressão entre as variáveis resíduos e número de consumidores, gerou as seguintes exceções:

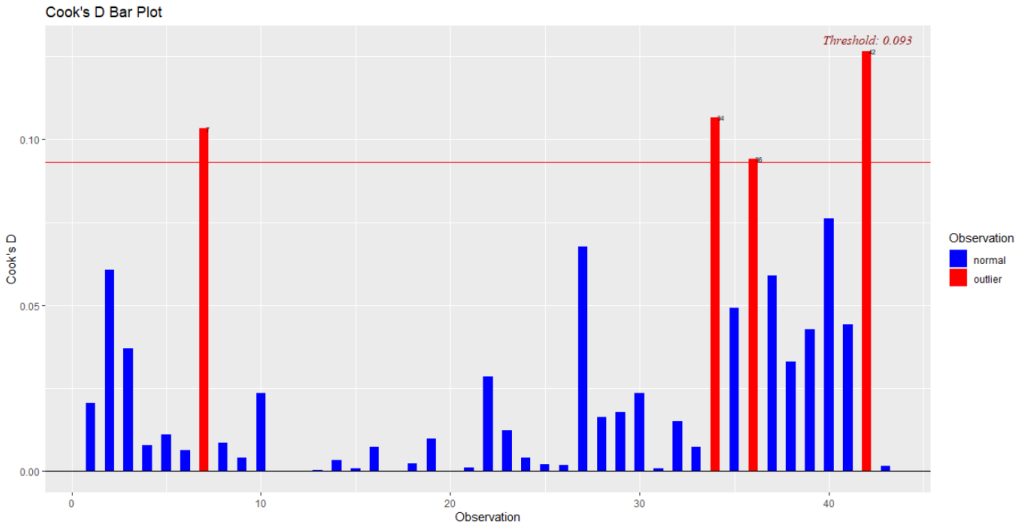


Gráfico 9 – Distância de Cook sobre os resíduos para COELBA

Diferente da anterior, os novos elementos extremos indicados pela distância de Cook foram o 7 e o 34.

id.	Resíduo	Número de UC's	Distância de Cook	Período
7	-115553	5.874.111	0,1030	01/07/2017
34	12431183	6.085.505	0,1060	01/10/2019

Tab 7.1 – Maiores distâncias de Cook sobre resíduos para COELBA

No gráfico de consumo por mês-a-mês é visível a redução do consumo de energia no meio do ano, no entanto, em 2017, a redução foi mais acentuada. Isso se deve principalmente às classes de consumo Residencial, Comercial e Industrial que tem maior peso dentre as demais e tiveram, no mês de julho, o menor consumo nos meses de 2017. Por outro lado, o mês de 10/2019, o consumo teve um dos seus maiores valores.

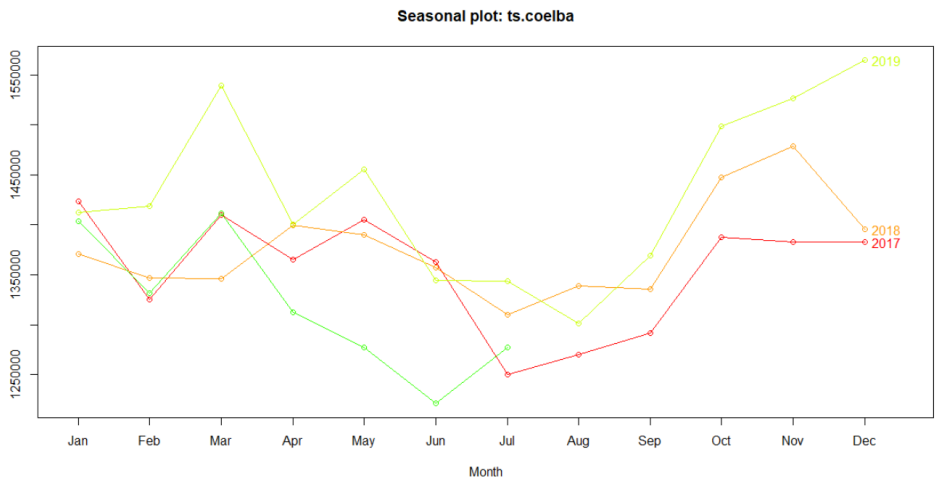


Gráfico 10 – Sazonalização consumo ano-a-ano COELBA

Considerando que a COELBA atua numa região com forte turismo nos períodos de início e final de ano, pode-se deduzir que este consumo será menor no meio do ano e maior nos demais, esclarecendo as exceções identificadas.

O gráfico da receita do período também corrobora a justificativa anterior, evidenciando a menor receita em 07/2017 e a terceira maior em 10/2019. Contudo, somente com a análise no menor nível de segregação dos dados (por UC's) é que torna possível conclusão a respeito de alguma irregularidade ou erro nos dados.

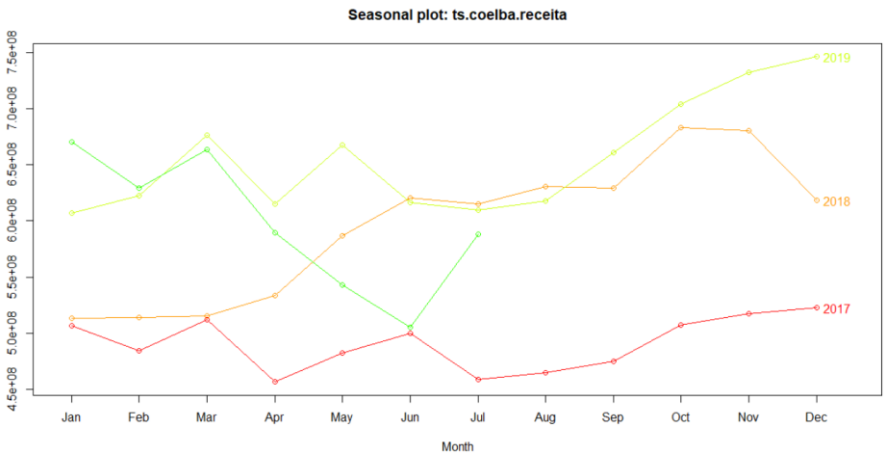


Gráfico 11 – Sazonalização receita ano-a-ano COELBA

c. Ajuste da regressão e distância de Cook para ELETROPAULO

$$Consumo = 3.313\ x\ 10^{6} - 9.042\ x\ 10^{-2}\ x\ Quantidade\ de\ UC's$$

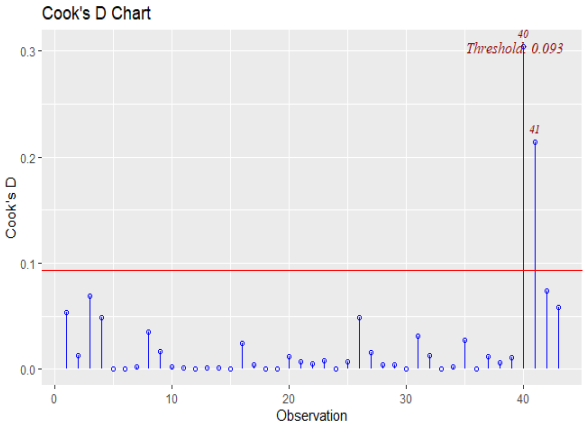
Coeficientes:

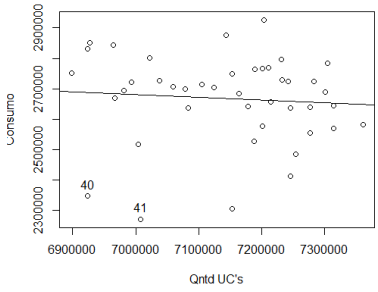
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.313e+06	1.264e+06	2.622	0.0122 *
Quantidade de UC's	-9.042e-02	1.769e-01	-0.511	0.6119

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC: 1149.408
Multiple R^2: 0.006335, R^2 Ajustado: -0.0179
F-statistic: 0.2614 on 1 and 41 DF, p-value: 0.6119

O valor estimado de β_0 é igual a $3.313x10^6$ e o valor de β_1 a inclinação foi de $-9.042x10^{-2}$, demonstrando que o aumento da "Quantidade de UC's" é acompanhado pela redução do "Consumo de energia". O intercepto foi estatisticamente significativo, apresentando valor inferior ao nível de significância de α = 0,05, porém o vetor "Quantidade de UC's" foi superior não sendo estatisticamente significante. O AIC de 1092.285 traduz que o modelo não explica o comportamento das variáveis, assim como o R^2 que também demonstra baixa correlação.





id.	Consumo	Número de UC's	Distância de Cook	Período
40	2346284	6923065	0.304	01/04/2020
41	2269055	7008106	0.214	01/05/2020

Gráfico 13 – Distância de Cook para ELETROPAULO (2017 a 2020). Autoria própria

Para Eletropaulo, por sua vez, a distância de Cook é demonstrada também com limite de 0.093, sendo extrapolada pelos elementos 40 e 41. Assim como a Cemig e Coelba,

os valores de 2020 podem ter ocorrido em decorrência da redução do consumo, acarretada pela pandemia.

A distância de Cook, conforme apresentado, tem sua importância na descoberta de quais dados são influentes a ponto de alterar mais sensivelmente a análise de regressão. Essa ferramenta, identifica não somente os valores extremos, mas os valores mais influentes, que pode trazer maior assertividade no trabalho do auditor.

Apesar dessa identificação, o R^2 ajustado do modelo de regressão teve baixíssima explicação da variabilidade dos dados: -0.017. Desse modo, será aplicada a distância de Cook sobre os resíduos da série temporal e, em seguida, comparar os resultados.

Como o R^2 teve uma explicabilidade muito baixa, seria desnecessário aplicar a distância de Cook sobre os dados. No entanto, optou-se por aplicar para avaliar os resultados.

Distância de Cook sobre os resíduos da série temporal (Eletropaulo)

Um modelo de regressão será ajustado sobre os resíduos da série temporal em relação ao número de consumidores, afim de se identificar novas possíveis exceções na base de dados analisada. Em seguida, será feita análise dos valores extremos indicados pela distância de *Cook*.

No primeiro momento, avaliou-se a normalidade dos resíduos, por meio do teste de Shapiro:

```
        Shapiro-Wilk normality test

data:  a$residuals
W = 0.982, p-value = 0.7263
```

Por meio do resultado do teste de Shapiro-Wilk de 0.98, não se rejeita a hipótese de normalidade dos erros, ao nível de 5% de significância. Em seguida, ajustou-se um modelo de série temporal sobre os resíduos, utilizando a função *auto.arima* do pacote *forecast* do R

```
Series: rl.eletropaulo$residuals
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.3935
s.e.   0.1706

sigma^2 estimated as 1.443e+10:  log likelihood=-550.41
AIC=1104.83   AICc=1105.14   BIC=1108.3
```

O modelo resultou em um ARIMA(0,1,1) e na análise do gráfico dos resíduos é possível identificar alguns elementos extremos.

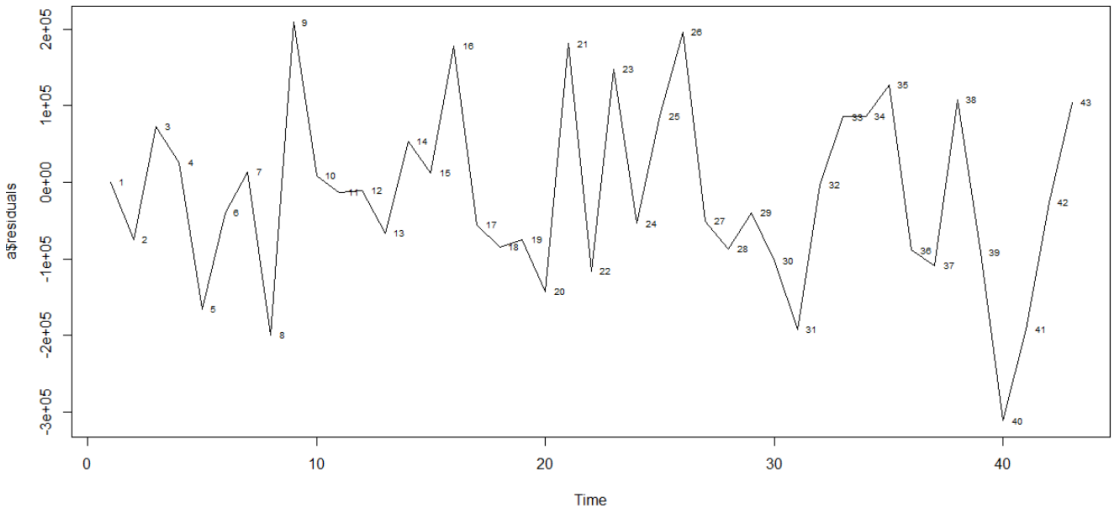


Gráfico 14 – Série temporal dos resíduos da ELETROPAULO

A distância de Cook aplicada sobre a regressão entre as variáveis resíduos e número de consumidores, gerou as seguintes exceções:

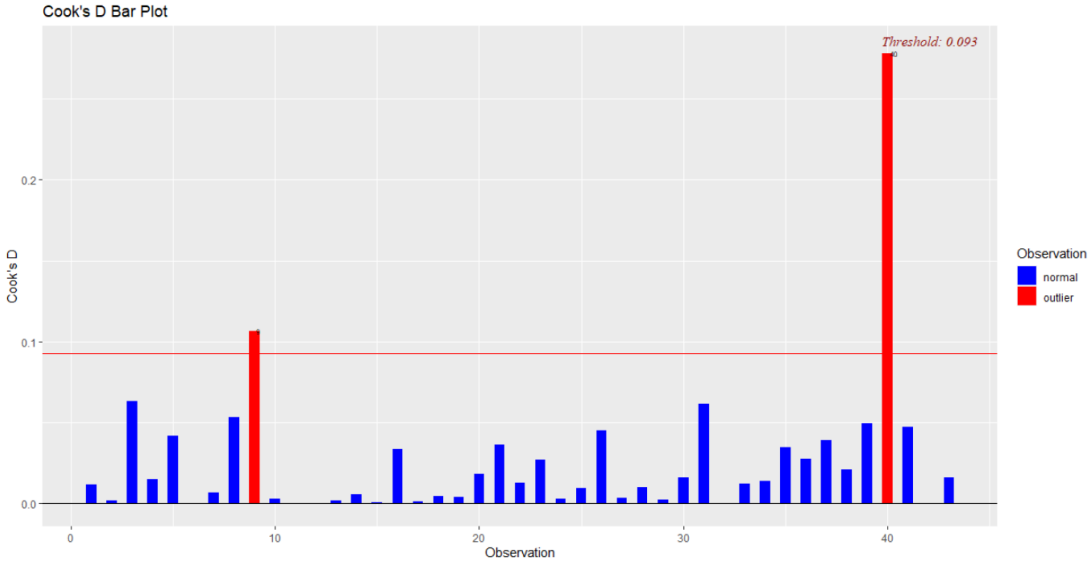


Gráfico 15 – Distância de Cook sobre os resíduos da ELETROPAULO

Os elementos extremos indicados pela distância de Cook foram o 9 e o 40:

id.	Resíduo	Número de UC's	Distância de Cook	Período
9	208639	7022538	0,106	01/09/2017
40	-311209	6923065	0,27	01/04/2020

Tab. 7.2 – Maiores distâncias de Cook dos resíduos ELETROPAULO

O período que o elemento 9 está inserido, 09/2017, é o de maior consumo da classe Residencial no ano de 2017. Comparando-se com os outros anos, no entanto, a classe Residencial, em 09/2017, não se destaca, localizando-se entre os anos de 2018 e 2019.

Gráfico 16 – Sazonalização do consumo residencial ano-a-ano ELETROPAULO

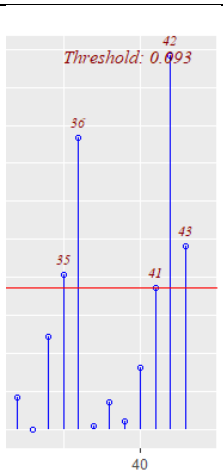
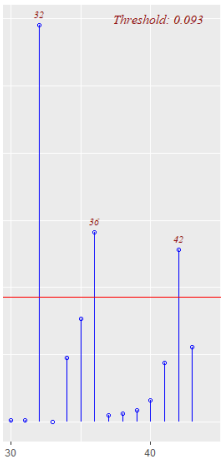
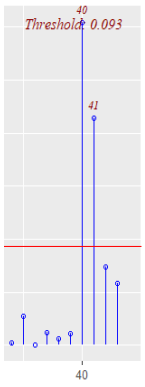
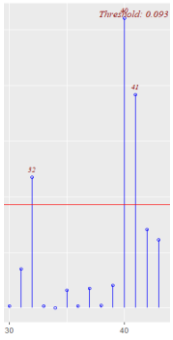
O elemento 40, por sua vez, localiza-se no período de plena pandemia causada pela COVID 04/2020. Além disso, considerando o ano de 2020 no gráfico acima, verifica-se que no mesmo período a classe residencial atingiu o consumo mínimo, corroborando a exceção identificada.

Simulação distância de Cook

Conforme mencionado na metodologia, simulou-se consumos de energia para as distribuidoras de energia do presente estudo, de forma a se avaliar a capacidade de identificação de exceções das técnicas. **A simulação foi feita no mês de agosto e considerou para a classe Residencial valores duplicados, valores limítrofes para**

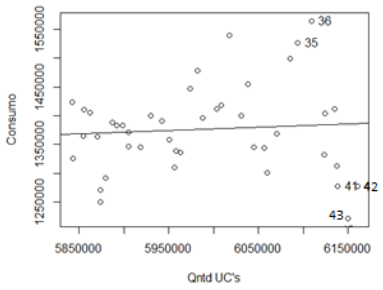
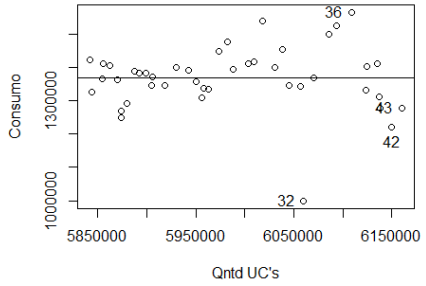
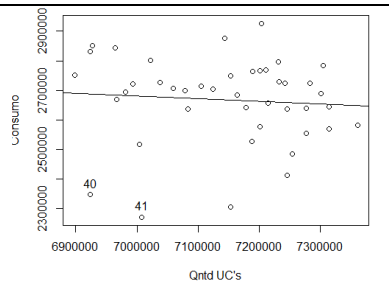
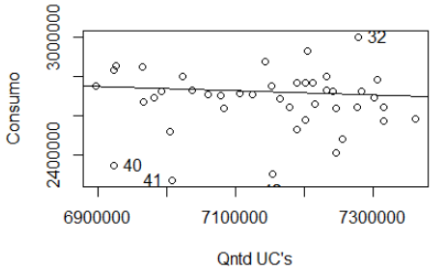
classe Comercial e valores arredondados para classe Industrial, todos em relação ao mês de julho.

Aplicando-se a distância de Cook nos dados simulados consolidados, verificou-se:

Empresa	Tratamento	Antes	Depois
Cemig	Duplicação	Sem diferenças	
Coelba	Arredondamento para baixo		
Eletropaulo	Limítrofe		Elementos 32, 40 e 41 

Antes da alteração, os elementos mais relevantes para distância de Cook da **Coelba** foram os 41, 35, 36, 43 e 42. Após a alteração do consumo de agosto, permaneceram apenas o elemento 32, 36 e 42, ou seja, o 35 e o 36 sumiram. Similar ao que ocorreu com **Eletropaulo**, que antes foram apenas os elementos 40 e 41, que sumiram, surgindo o 32.

Essa situação também pode ser observada no gráfico de regressão, observe que a linha de tendência perdeu a inclinação anterior, puxada pelo elemento 32 que é o valor de consumo 1.000.000 MWh:

Empresa	Tratamento	Antes	Depois
Coelba	Arredondamento para baixo		
Eletropaulo	Limítrofe		

Deve-se considerar que, apesar da linha de tendência ter sido afetada pelo consumo criado para a **Coelba**, a manipulação tornou o consumo de agosto, o mínimo de toda série histórica. Em relação à **Eletropaulo**, a reta de regressão deslocou um pouco para cima em função do elemento alterado 32, que passou de 2.552.967,39 MWh para 2.999.999,99, tornando o maior consumo da série histórica. Fica demonstrado, desse modo, a capacidade da distância de Cook em identificar ambas exceções.

Análise de Componentes Principais – ACP

A aplicação da Análise de Componentes Principais na base de dados de energia consumida das classes de consumo tem como objetivo identificar quais variáveis conseguem explicar melhor a variância dos dados. Esse método condensa a informação contida nas variáveis originais num conjunto menor de variáveis com perda mínima de informação.

Para identificação de exceções, a ACP pode auxiliar o auditor a direcionar seu trabalho naquelas variáveis que melhor expliquem a distribuição e dos dados. Como na base de dados da ANEEL os dados segregados por classe de consumo são correlacionados, a ACP auxilia a reduzir a quantidade de variáveis para novas variáveis que expliquem melhor a variabilidade dos dados.

a. ACP para Cemig

A aplicação da ACP nos dados de energia consumida da Cemig gerou as componentes principais com os pesos:

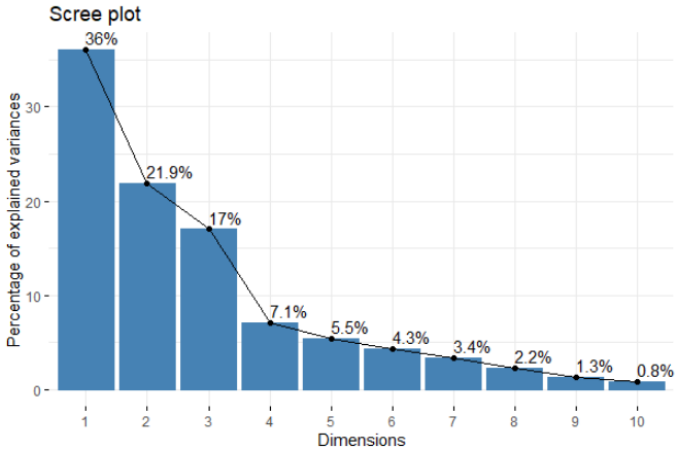


Gráfico 17 – ACP no consumo de energia para CEMIG

Verifica-se que as duas primeiras componentes principais explicam por 58% da variabilidade dos dados. Abaixo, para a Componente Principal – CP1, os autovetores das variáveis que mais contribuíram para variabilidade de 36%:

Comercial_ Serviços	Poder_ Público	Serv_P (tração)	Industrial	Consumo _Próprio	Rural_ Aqu	Iluminação _Pública	Residencial	Serv_P (água, esgoto)	Rural	Rural_Irrig
46,9%	44,5%	38,7%	36,7%	34,6%	27,5%	19,0%	14,2%	0,9%	-13,0%	-16,8%

Tab. 8 – Autovetores para CP1 da CEMIG

Com base nesses dados e no gráfico abaixo,

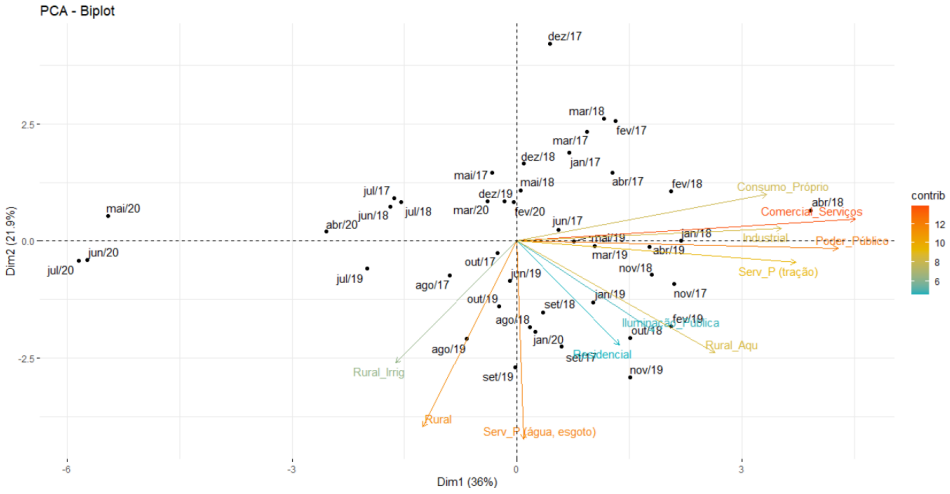


Gráfico 18 – Gráfico da ACP para CEMIG

observa-se forte correlação entre as variáveis **Comercial, Poder Público, Serviço Público (tração), Industrial e Consumo Próprio,** por seus vetores formarem ângulos agudos, refletindo os valores dos autovetores que mais contribuiram. Por outro lado, as variáveis **Rural e Rural Irrigante** foram as que menos contribuem para explicação dos dados da CP1. Desse modo, a Componente 1 pode ser caracterizada por um indicador de **Comércio, Indústria e Governo**, explicando 36% dos dados.

Em relação a contribuições individuais, os meses de **maio, junho e julho de 2020** foram os que mais contribuiram, reduzindo a explicação da CP1. O mês de abril, também, auxiliou, mas positivamente a explicabilidade da Componente 1. Por outro lado, o mês de setembro de 2019 foi o que menos contribuiu.

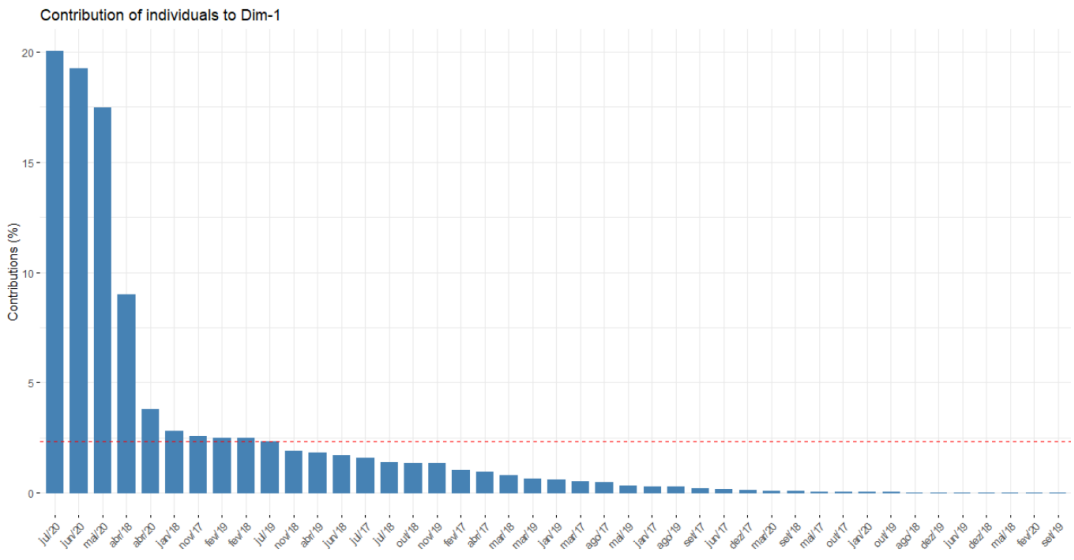


Gráfico 19 – Contribuições da CP1 da CEMIG mês-a-mês

O consumo nestes períodos de maiores contribuições, maio a julho de 2020 e abril de 2018, referem-se aos menores consumos históricos, o que auxiliou a redução da explicabilidade da PC1.

b. ACP para Coelba

As duas primeiras Componentes Principais da Coelba explicam por 60% da variabilidade dos dados:

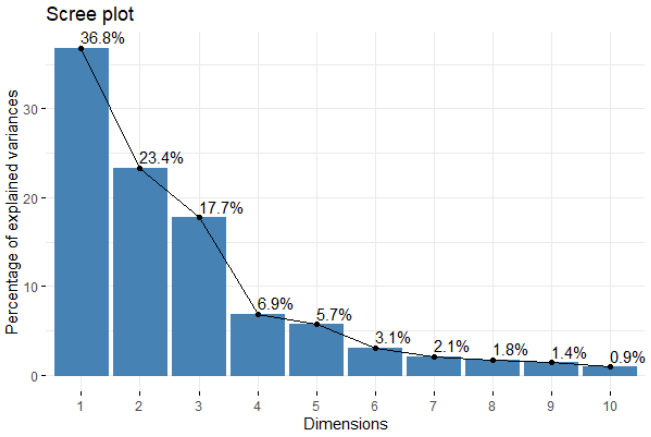


Gráfico 20 – Contribuições das componentes da COELBA

Na tabela abaixo são apresentados, para a primeira componente, os autovetores em ordem decrescente das variáveis que mais contribuíram para variabilidade de 36.8%:

Poder_ Público	Comercial_ Serviços	Consumo_ Próprio	Serv_P (água, esgoto)	Residencial	Rural_ Aqu	Serv_P (tração)	Rural	Iluminação _Pública	Industrial	Rural_Irrig
44,84%	43,55%	42,93%	37,72%	36,86%	25,51%	24,87%	9,97%	9,50%	3,23%	0,17%

Tab. 8.1 – Autovetores para CP1 da COELBA

Na figura a seguir, verifica-se forte correlação entre as variáveis **Poder Público, Comercial e Serviços, Consumo Próprio e Serviço Público (água e esgoto),** por seus vetores formarem ângulos agudos, refletindo os valores dos autovetores que mais contribuiram.

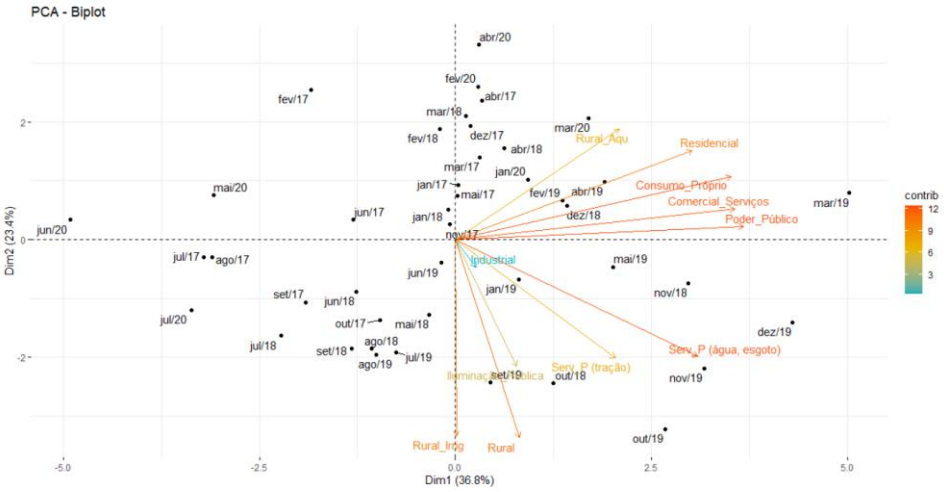


Gráfico 21 – Gráfico da ACP para COELBA

Por outro lado, a variável **Industrial** foi a que menos contribui para explicação dos dados da CP1. Desse modo, a Componente 1 pode ser caracterizada por um indicador de **Comércio e Governo**, explicando 36% dos dados.

Em relação a contribuições individuais, os meses de **maio e dezembro de 2019** foram os que mais contribuiram, aumentando a capacidade explicativa da CP1. O mês de **junho de 2020**, por outro lado, reduziu a explicabilidade da Componente 1.

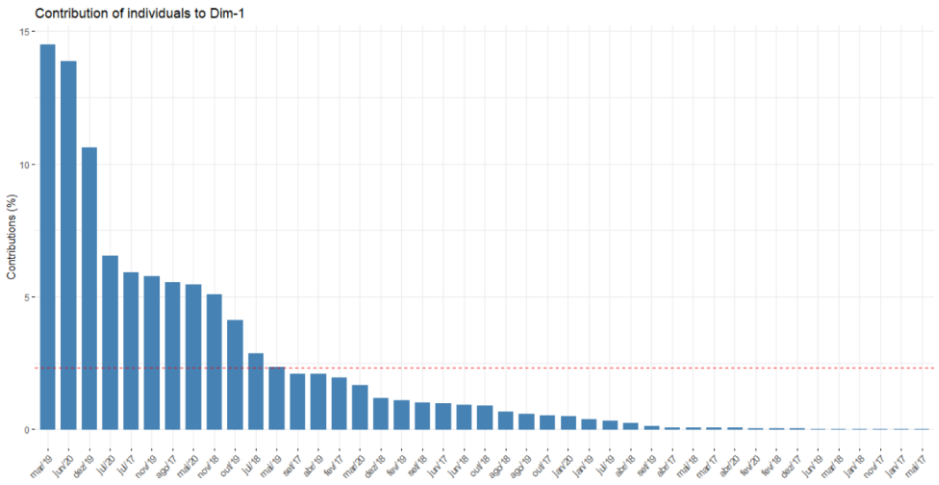


Gráfico 22 – Contribuições da CP1 da COELBA mês-a-mês

Analisando o consumo nestes períodos de maiores contribuições, março/19, dez/19, verifica-se que se refere aos maiores consumos históricos das classes características e, para jun/20, o menor consumo.

c. ACP para Eletropaulo

A Eletropaulo tem as duas primeiras componentes principais respondendo por 70% da variabilidade dos dados.

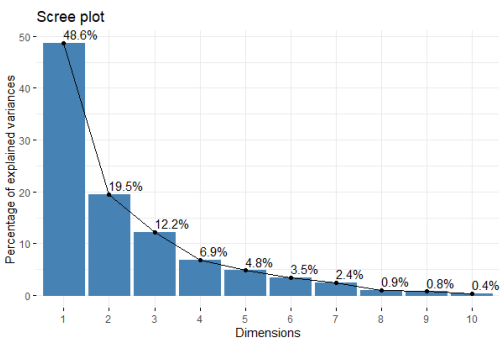


Gráfico 23 – Contribuições das componentes da ELETROPAULO

Abaixo, para a primeira componente, os autovetores das variáveis que mais contribuíram para variabilidade de 48.6%:

Industrial	Serv_P (água, esgoto)	Serv_P (tração)	Poder_Público	Comercial_Serviços	Rural_Aqu	Consumo_Próprio	Iluminação_Pública	Residencial	Rural
40,66%	40,20%	40,03%	36,43%	35,85%	33,38%	31,09%	18,50%	2,97%	-9,13%

Tab. 8.2 – Autovetores para CP1 da ELETROPAULO

Em seguida, observa-se forte correlação entre as variáveis **Industrial, Serviço Público (água e esgoto) e Seviço Público (tração elétrica),** por seus vetores formarem ângulos agudos, refletindo os valores dos autovetores que mais contribuiram.

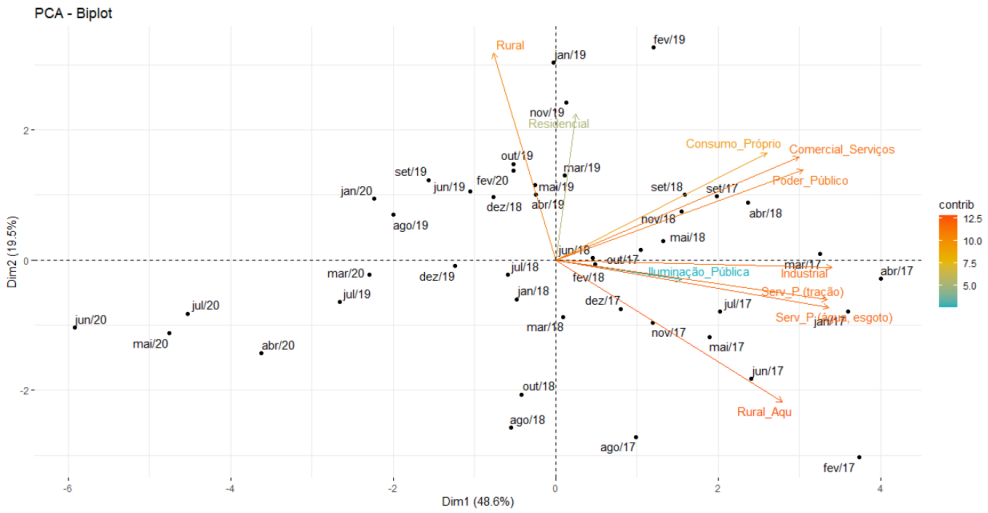


Gráfico 24 – Gráfico da ACP para ELETROPAULO

Por outro lado, a variável **Rural** foi a que menos contribui para explicação dos dados da CP1. Desse modo, a Componente 1 pode ser caracterizada por um indicador de **Indústria e Governo**, explicando quase 50% dos dados.

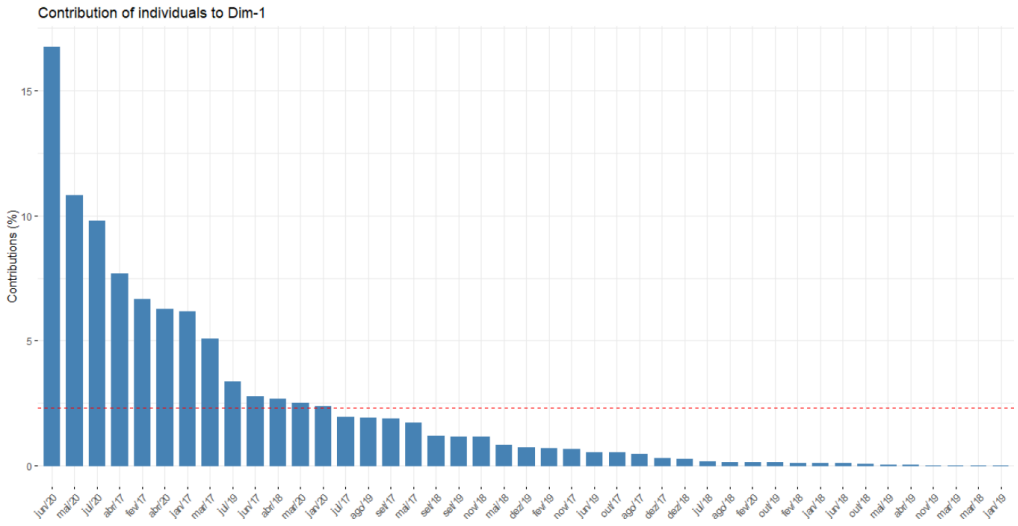


Gráfico 22 – Contribuições da CP1 da ELETROPAULO mês-a-mês

Em relação a contribuições individuais, os meses de **maio, junho e julho de 2020** foram os que mais contribuiram, reduzindo a explicação da CP1. O mês de **abril de 2017**, por outro lado, aumentou a explicabilidade da Componente 1.

Deve-se destacar que a Eletropaulo não nenhuma UC na classe rural irrigante. Este, por si só, é um fator que merece ser analisado, uma vez que todas empresas analisadas, em julho de 2020, têm UC's nesta classe: Cemig 17.835 unidades e Coelba 21.431. Muito embora a Eletropaulo estar localizada em região metropolitana com poucas características rurais que necessitem irrigação.

Em relação aos meses que mais contribuíram para a CP1, observa-se que todos estão no período da pandemia do COVID-19 que acarretou grande redução de consumo de energia. Esses consumos foram os valores mínimos para as classes características da componente Indústria e Governo.

Simulação ACP

Conforme mencionado na metodologia, simulou-se consumos de energia para as distribuidoras de energia do presente estudo, de forma a se avaliar a capacidade de identificação de exceções das técnicas. A simulação foi feita no mês de agosto e considerou para a classe Residencial valores duplicados, valores limítrofes para

classe Comercial e valores arredondados para classe Industrial, todos em relação ao mês de julho.

Para Cemig e Coelba, não foram verificadas diferenças das componentes principais anteriores. Por outro lado, para a Eletropaulo, as variáveis Serviço Público (tração) e Industrial trocaram de lugar:

Antes	Industrial	Serv_P (água, esgoto)	Serv_P (tração)	Poder_Público	Comercial_Serviços	Rural_Aqu	Consumo_Próprio	Iluminação_Pública	Residencial	Rural
	40,66%	40,20%	40,03%	36,43%	35,85%	33,38%	31,09%	18,50%	2,97%	-9,13%

Depois	Serv_P (tração)	Serv_P (água, esgoto)	Industrial	Poder_Público	Comercial_Serviços	Rural_Aqu	Consumo_Próprio	Iluminação_Pública	Residencial	Rural
	40,47%	40,39%	39,01%	36,63%	36,19%	33,44%	31,59%	18,85%	3,57%	-8,59%

Tab. 8.3 – Autovetores para CP1 da ELETROPAULO antes e após simulação

O gráfico permite verificar que o mês de agosto de 2019, cujos consumo das variáveis Residencial, Comercial e Industrial foram alterados para maior, reduziram a força do vetor industrial, tornando um pouco mais forte o vetor Serviço Público (tração). No entanto, a classificação do indicador **Indústria e Governo** se mantém, assim como o poder de 50% de explicação da componente 1.

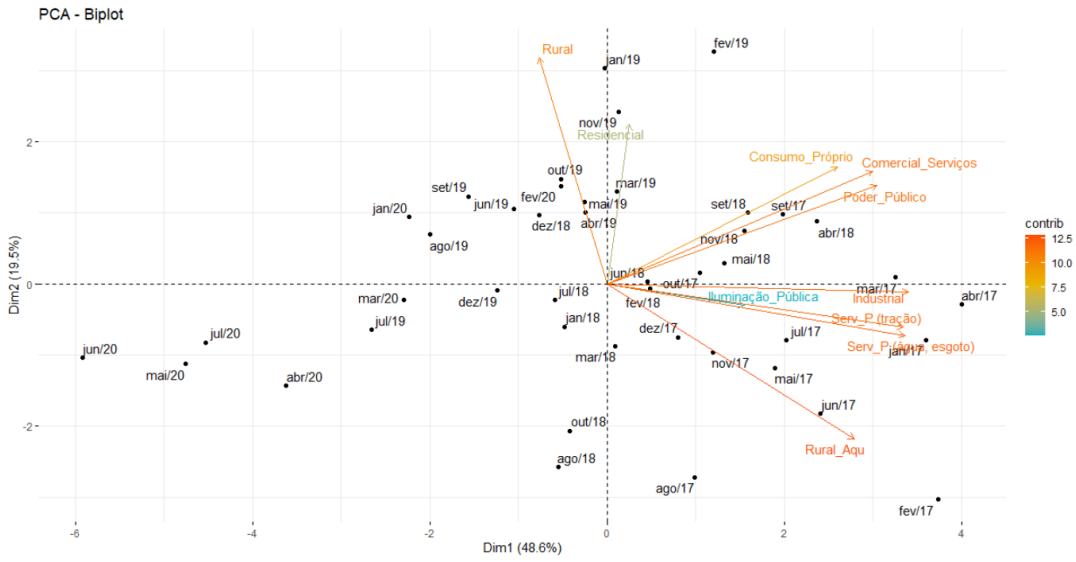


Gráfico 23 – ACP para ELETROPAULO antes da simulação

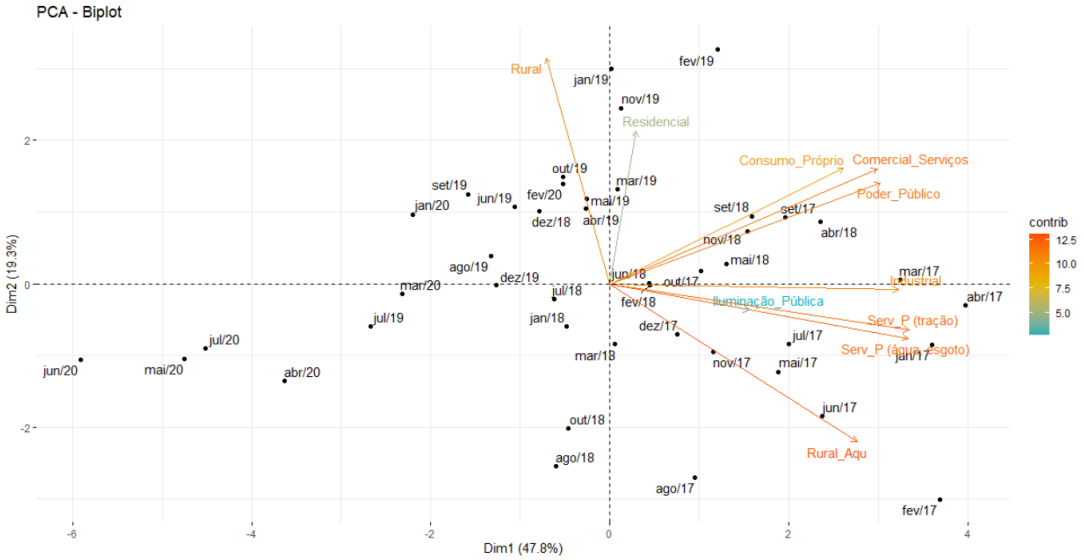


Gráfico 24 – ACP para ELETROPAULO depois da simulação

Dessa forma, observou-se que a ACP não identificou exceções na base de dados manipulada e, a alteração na **Eletropaulo**, não foi capaz de acarretar diferenças relevantes nas componentes.

CONCLUSÕES

A atividade de auditoria tem o dever de buscar incessantemente novas ferramentas, técnicas e formas para que atinja seus objetivos. Este trabalho apresentou essa necessidade de adaptação e renovação, tendo em vista que todos os processos das empresas estão passando por mudança disruptivas, principalmente durante a pandemia, e continuar com procedimentos da maneira antiga, pode não mais ser adequado. Dessa forma, este estudo abordou técnicas estatísticas que podem ser utilizadas por auditores, a fim de verificar a capacidade de identificação de exceções em quaisquer bases de dados.

O *dataset* utilizado é disponibilizado pela publicamente pela agencia reguladora ANEEL e contempla dados de consumo, receita, tarifa de energia e quantidade de Unidades Consumidoras. Neste estudo, o escopo foram esses dados das 3 maiores concessionárias, em termos de consumo de energia: Cemig, Coelba e Eletropaulo, do período de janeiro de 2017 a julho de 2020.

Das técnicas utilizadas, a **análise exploratória dos dados**, por meio da apuração da média, mediana, desvio padrão e moda, se mostrou eficaz na identificação de exceções. A visualização dessas medidas descritivas, por meio de histograma e *boxplot*, facilitou a identificação de valores extremos para todas empresas em alguns meses de 2020. Desconsiderando o ano de 2020, a Coelba apresentou valores extremos em períodos de festividades, como março e dezembro. A aplicação da **moda**, por sua vez, permitiu a identificação erros na extração de dados da base da ANEEL.

A aplicação da **correlação** nos dados de Consumo, Receita, Número de UC's e Tarifas possibilitou identificar padrões que necessitam uma análise dos microdados segregados por clientes. Isso porque, identificou-se que a Eletropaulo em 2018 apresentou correlação negativa entre Consumo e Receita para a classe "Comercial, Serviços e Outros". É uma situação que foge aos padrões das outras distribuidoras, mas que pode ter a ver com o aumento da tarifa no mesmo período, falha de medição ou furto de energia. Outro fator de destaque é o impacto da pandemia do COVID na classe Residencial e Rural que acarretou um aumento do consumo de energia, apesar do aumento da tarifa no período. E, por fim, para a Cemig especificamente em 2020, observou-se que a redução do consumo de energia tem relação inversa com o crescimento de unidades consumidoras. Desse modo, a análise dos microdados se faz necessária.

A aplicação da **Lei de Newcomb-benford**, ou lei do primeiro dígito, resultou em estreita conformidade quando se considera o *dataset* com todas as distribuidoras do país. Neste sentido, as maiores diferenças entre o padrão da LNB em relação aos consumos apurados para Cemig, Coelba e Eletropaulo foram investigadas em relação aos três primeiros dígitos. Verificou-se que o maior benefício da LNB é dar indícios de elementos que não necessariamente são valores extremos, podendo ser muito útil no direcionar de amostras a serem realizadas.

A identificação de valores influentes por meio da **distância de Cook** é um dos subprodutos da **regressão linear**. Nessa aplicação, os elementos mais influentes

estavam mais concentrados no ano de 2020, decorrentes do impacto da pandemia. Contudo, como o R^2 dos modelos teve baixa explicação, optou-se por analisar os resíduos de um modelo ARIMA de séries temporais e, como resultado, novos elementos foram identificados fora do período de pandemia. Como os períodos são compostos por dados consolidados, seria importante analisar os elementos que compõe esses dados para outras conclusões. Por isso, verificou-se que a análise dos resíduos é complementar, uma vez que o modelo de regressão linear usual identificou principalmente elementos previsíveis por serem do período de pandemia.

Em relação à **Análise de Componentes Principais**, ficou demonstrado que a redução da dimensionalidade dos dados permite um enorme apoio no trabalho dos auditores, uma vez que uma explicação de mais de 50% do *dataset* pode ser obtida pela análise de apenas 3 variáveis, com pequena perda de informações. Em relação a identificação de exceções, pela análise gráfica foi possível identificar elementos extremos que reduziram ou aumentaram a capacidade de explicação das componentes, porém o maior benefício foi mesmo o direcionamento da análise para as variáveis que tem maior capacidade de explicar a variabilidade da base de dados.

Por fim, foram criadas **simulações de consumo** de energia para o mês de agosto de 2019, a fim de se avaliar a capacidade das ferramentas em identificar essas manipulações. A moda, a correlação, a lei de Benford e a distância de Cook foram capazes de identificar essas variações nos dados, mas **a moda, a lei de Benford, seguida pela distância de Cook**, se destacaram dando maiores indícios de onde podem estar exceções relevantes. A análise de componentes, por outro lado, tem uma importância maior na compreensão da base de dados, por meio da redução da dimensionalidade dos dados.

Desse modo, análises exploratórias, como as de medidas de tendências central, com demonstração gráfica, em conjunto a Lei de Newcomb-benford e regressão linear, foram as ferramentas mais capazes de identificar possíveis exceções na base de dados utilizada.

Face ao exposto, o presente estudo respondeu aos objetivos geral na medida em que demonstrou que todas as técnicas estatísticas utilizadas foram capazes de identificar exceções. **Em relação ao objetivo específico a aplicação da moda, da correlação em conjunto com análise gráfica se mostraram mais adequadas.** Deve-se considerar, no entanto, que os dados utilizados são consolidados o que diminui a capacidade de identificações mais assertivas e que a integração de técnicas, aliada a análise gráfica traz resultados mais robustos do que a aplicação isolada.

Além disso, deve-se dar destaque para o impacto que a **pandemia da COVID-19** causou na análise de exceções. Isso porque a natural redução do consumo de energia em alguns meses de 2020 permitiu que dados extremos surgissem, sem que tenham ocorrido por motivos de fraudes ou erros. Todavia, deve-se ter em mente, que este período de redução do consumo pela COVID pode ter sido manipulado indevidamente forçando de uma redução maior do que a real ou uma retomada maior do que a real. Esta situação poderia ser identificada comparando-se os dados das outras distribuidoras do país, o que não foi objetivo deste trabalho.

Por tudo isso, o presente estudo não pretende esgotar o assunto de utilização de técnicas estatísticas no apoio a trabalhos de auditoria, sugerindo maiores aprofundamentos com a utilização de microdados por unidades consumidoras, aplicação da distância de *Mahalanobis*, Lei dos grandes números, regressão múltipla com variáveis dummy, predição de exceções por meio da aplicação de outros métodos de séries temporais e a análise das componentes principais em componentes com menor capacidade explicativa.

REFERÊNCIAS BIBLIOGRÁFICAS

ANEEL, Agência Nacional de Energia Elétrica. **Resolução Normativa n° 610 de 2014 – ANEXO 1**. Disponível em <http://www2.aneel.gov.br/cedoc/ren2014610.pdf>. Acessado em 04/11/2020

ANEEL, Agência Nacional de Energia Elétrica. **Relatórios de Consumo e Receita de Distribuição.** Disponível em: <https://www.aneel.gov.br/relatorios-de-consumo-e-receita>. Acessado em 04/11/2020

ARAÚJO, Inaldo da Paixão Santos. ***Introdução à auditoria: área governamental***. Salvador: Egba, 1988.

BAESENS, Bart; Vlasselaer, Véronique; Verbeke, Wouter. ***Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection***. ed. Wiley, 2015.

BEUREN, I. M., Longaray, A. A., Raupp, F. M., Sousa, M. A. B., Colauto, R. D., & Porton, R. A. B. (Orgs.). (2003). ***Como elaborar trabalhos monográficos em contabilidade: teoria e prática***. São Paulo: Atlas.

BOUVEYRON, Charles; Celeux, Gilles, Murphy, T. Brendan; Raftery, Adrian E. Model-Based ***Clustering and Classification for Data Science***. ed. Cambridge University Press, 2019.

BOYNTON, William C.; Johnson, Kell. ***Auditoria*** – tradução José Evaristo dos Santos. ed. Atlas, 2002.

CECON, Paulo; Silva, Anderson; Nascimento, Moysés; Ferreira, Adésio. ***Métodos Estatísticos***. Série didática. ed. UFV, 2012.

CEMIG, Companhia Energética de Minas Gerais. **Valores de Tarifas e Serviços**. Disponível em: < https://novoportal.cemig.com.br/atendimento/valores-de-tarifas-e-servicos/>. Acessado em 04/11/2020.

DIACONIS, Persi; Mosteller, Frederick. ***Methods for Studying Coincidences***. *Journal of the American Statistical Association*, 1989

DYSON, Freeman. ***The Scientist as Rebel***. New York Review Collection, p.275. 2006

DW, Deutsche Welle. ***Wirecard committed 'elaborate and sophisticated fraud' say auditors***. <https://www.dw.com/en/wirecard-committed-elaborate-and-sophisticated-fraud-say-auditors/a-53942273>. Acessado em 28/10/2020.

FOLHA, Jornal. **IRB vê fraude no pagamento de R$ 60 milhões em bônus a ex-executivos**. <https://www1.folha.uol.com.br/mercado/2020/06/irb-ve-fraude-no-pagamento-de-r-60-milhoes-em-bonus-a-ex-executivos.shtml>. Acessado em 28/10/2020

GOMES, Bruno C. **Uma Análise da Conformidade das Demonstrações de uma Empresa de Energia Elétrica com a Lei de Newcomb-Benford**. TCC UFMG-FACE. 2013

GOMES, Bruno C. **A arte da lavagem de dinheiro: Como criminosos tornam lícito o dinheiro fruto de crime**. ed. Amazon, 2019

IIA. **Missão auditoria Interna.** Disponível em: <https://iiabrasil.org.br//ippf/missao-da-auditoria-interna>. Acesso em 24/09/2020

____. **Definição de Auditoria Interna**. Disponível em: <https://iiabrasil.org.br//ippf/definicao-de-auditoria-interna>. Acessado em 24/09/2020

____. **Normas Internacionais para Prática Profissional de Auditoria Interna**. Tradução IIA Brasil. *The Institute of Internal Auditors*, 2017

____. **Declaração de Posicionamento do IIA: Fraude e a Auditoria Interna** – Tradução IIA Brasil. *The Institute of Internal Auditors*, 2019

GTAG 16, ***Global Technology Audit Guide - GTAG 16: Data Analysis Technologies***. *The Institute of Internal Auditors*, 2011

ISA 315. ***Internacional Standards of Auditing N. 315***. Disponível em < https://www.ifac.org/system/files/downloads/a017-2010-iaasb-handbook-isa-315.pdf>. Acessado em 05/10/2020

KASSAMBARA, Alboukadel. ***Machine Learning Essentials***. ed. STHDA, 2017.

LEMMON, Tom. ***Auditors must "stop pretending" it's not their job to catch fraud. 2020.*** Disponível em: <https://www.accountancyage.com/2020/09/07/auditors-must-stop-pretending-its-not-their-job-to-catch-fraud/>. Acessado em 25/09/2020

LESHIK, Edward; CRALLE, Jane. ***An introduction to algorithmic trading. Basic to advanced strategies.*** ed. Wiley Trading, 2011

LEVINE, David. **Estatística Teoria e Aplicações: Usando o Microsoft Excel em Português**. Tradução Teresa Cristina Padilha – Rio de Janeiro, LTC, 2014

MARGARIDO, Mário Antônio. **Modelos de Séries Temporais: Uma introdução com aplicações práticas**. Amazon, 2020

MORETTIN, Pedro; TOLOI, Clelia. **Análise de Séries Temporais**. Ed. Edgard Blucher, 2018

NIGRINI, J. Mark. ***Dusting your data for fraud's fingerprints: Six number patterns that fraudsters use***. Disponível em <https://www.fraud-magazine.com/cover-article.aspx?id=4295011516>. Acessado em 03/11/2020

_____, J. Mark. ***Forensic Analytics: Methods and Technique for Forensic Accounting Investigations***. John Wiley & Sons, 2012.

PCAOB, ***Auditing Standard No. 15.*** Disponível em: <https://pcaobus.org/Standards/Archived/PreReorgStandards/Pages/Auditing_Standard_15.aspx>. Acessado em 29/09/2020

RAMASUBRAMANIAN, Karthik; Singh, Abhishek. ***Machine Learning Using R - With Time Series and Industry-Based Use Cases in R.*** *ed. Apress, 2018*

SILVA, Anderson. **Métodos de Análise Multivariada em R**. ed. FEALQ, 2016

TOFFLER, Alvim. ***Future Shock***. ed. Bantam, 1984.

WASSERMAN, Larry. ***A Concise Course in Statistical Inference***. ed. Springer, NA.