REVISTA BRASILEIRA

DE ESTATÍSTICA

volume 81

número 248

Janeiro/dezembro 2023

IBGE
Instituto Brasileiro de Geografia e Estatística

Instituto Brasileiro de Geografia e Estatística – IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

Volume 81 - número 248 – jan/dez 2023

ISSN 2675-3243

R. Bras. Estat., Rio de Janeiro, v.81, n.248, p. 1-104, jan/dez 2023

Instituto Instituto Brasileiro de Geografia e Estatística – IBGE

Revista Brasileira de Estatística, ISSN 2675-3243

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site https://rbesojs.ibge.gov.br.

Os artigos submetidos à RBE não devem ter sido publicados em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração

José André de Moura Brito

Capa

José André de Moura Brito

Informações Adicionais

Revista Brasileira de Estatística/IBGE - v.1, n.1 (jan/mar. 1940) – Rio de Janeiro: IBGE, 1940.v. trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística. Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v.50 (1980-1989). Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN publicação online 2675-3243, a partir de 2019.

I. Estatística – Periódicos. II. IBGE. III. Associação Brasileira de Estatística.

Gerência de Biblioteca e Acervos Especiais CDU 31(05) RJ-IBGE/88-05 (ver. 2009) PERIÓDICO

Sumário

Nota do Editor

É com grande satisfação que trazemos mais um volume da Revista Brasileira de Estatística.

Os artigos aqui apresentados trazem propostas metodológicas combinadas com diferentes contextos de aplicação. Interessados em metodologias estatísticas irão encontrar material e referências, enquanto interessados no universo das aplicações de estatística terão exemplos interessantes.

Dois artigos trazem temas ligados à formação em Estatística. Metodologias de riscos competitivos em análise de sobrevivência são utilizadas por Adriane Portela e coautoras em um estudo que busca identificar fatores ligados à evasão e conclusão do curso de bacharelado em estatística a partir de microdados do censo educacional.

Já Anderson Ara e coautores apresentam o perfil dos primeiros cursos de Ciência de Dados no país. Formas de oferta, tipo de instituição, dentre outras características são avaliadas. É esperado o surgimento de mais cursos de Ciência de Dados e correlatos, bem como reposicionamento de cursos de estatística. Desta maneira, além de fornecer a fotografia do momento, o artigo será importante para referência futura, comparações e avaliação de evolução de cursos com este enfoque.

Avançando no volume, Fernando Ferraz e coautores trazem um artigo sobre modelagem de valores extremos na qual adotam uma abordagem Bayesiana proposta para tratar o problema de dados faltantes. Dados de precipitação em Fortaleza, CE e temperaturas em Teresina, PI são utilizados para ilustrar a metodologia baseada em modelos dinâmicos. Além do método proposto, o texto oferece uma interessante e objetiva introdução à área.

Completamos este volume com o trabalho de Caio Gonçalves e coautoras que trata do tema de desemprego com a estimação da dinâmica de desocupação no Brasil com análises de algumas unidades da federação.

Métodos de espaços de estados e séries temporais utilizados levam em conta o efeito do desenho amostral da PNAD na correlação entre estimativas.

Autores investigam séries de seguro-desemprego e desocupados e buscando identificar se há correlação entre componentes das duas séries. Além da descrição metodológica e discussão dos resultados e achados e o trabalho contribui para evidenciar a importância de produção regular de estatística de acompanhamento do tema.

Desejo a todos uma excelente leitura.

Paulo Justiniano Ribeiro Jr.
Editor RBE.

Revista Brasileira de Estatística, V.81, N. 248, p.7-30, 2023

EVASÃO E CONCLUSÃO NO CURSO DE BACHARELADO EM ESTATÍSTICA: UMA APLICAÇÃO DE ANÁLISE DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS

Adriane Caroline Teixeira Portela

adrianeportela57@gmail.com

Universidade de São Paulo e Universidade Federal de São Carlos – USP/UFSCar

Andrea Diniz da Silva

andrea.silva@ibge.gov.br

Escola Nacional de Ciências Estatísticas – ENCE/IBGE

Giovana Oliveira Silva

giovanaufba@gmail.com

Universidade Federal da Bahia – UFBA

Resumo: Nas últimas décadas, diversas políticas públicas foram criadas ou reformuladas com o objetivo de democratizar o acesso ao ensino superior no Brasil. Embora os impactos positivos dessas ações sejam visíveis, após o ingresso no ensino superior, os alunos ainda enfrentam diversos desafios que afetam sua permanência nesse nível educacional. É preciso identificar quais e como variáveis selecionadas estão associadas às taxas de evasão e de conclusão, pois apesar da abundante produção acadêmica sobre o tema, ainda há lacunas para o entendimento dos fatores associados a cursos específicos. A análise de sobrevivência foi utilizada para entender uma situação em que a evasão está competindo com a conclusão, em particular, foi considerada uma abordagem para riscos competitivos. Para tanto, foi utilizada a função de incidência acumulada e ajustado o modelo de riscos competitivos introduzido por Fine e Gray (1999). Foram utilizados microdados do censo educacional brasileiro para alunas/os do curso de Graduação em Estatística que ingressaram no ano de 2010 e foram acompanhados até o ano de 2017. Os resultados encontrados confirmam a influência de fatores apontados como relevantes na literatura, como sexo, idade, turno, envolvimento em atividade complementar e o tipo de organização acadêmica.

Palavras-chave: evasão e conclusão; bacharelado em estatística; análise de sobrevivência; riscos competitivos.

Abstract: In the last decades, several public policies were created or reformulated to democratize access to higher education in Brazil. Although the positive impacts of these actions are visible, after entering higher education, students still face several challenges that affect their permanence at this educational level. It is necessary to identify which and how variables are associated with dropout and completion rates because despite the abundant academic production on the subject, there are still gaps in the understanding of factors associated with specific courses, especially Statistics. Survival Analysis was used to approach a situation in which dropout is competing with completion, in particular competing risk analysis was used. For this purpose, the cumulative incidence function was used, and the competing risk model introduced by Fine and Gray (1999) was adjusted. Brazilian education census microdata were used and students admitted in 2010 were followed up until 2017. The results confirm the influence of relevant factors found in the literature such as sex, age, shift, engagement in complementary activity, and the type of academic organization.
Keywords: dropout and completion; bachelor in statistics; survival analyses, competing risk analyses

1.INTRODUÇÃO

Nas últimas décadas houve grandes incentivos voltados à democratização do acesso ao ensino superior no Brasil. Diversas políticas públicas foram criadas ou reformuladas a fim de possibilitar o surgimento de novos ingressantes neste nível educacional. Dentre as ações inclusivas surgidas, podemos citar a ampliação da oferta de vagas voltadas ao ingresso de estudantes de baixa renda familiar, provenientes de escolas públicas e de diferentes etnias, como também os diversos programas de financiamento estudantil. Segundo Pinto (2004), ao analisar a situação do acesso à educação superior no Brasil nos últimos 40 anos, percebe-se que a taxa de escolarização bruta na educação superior do país ainda é uma das mais baixas da América Latina, enquanto o grau de privatização é um dos mais altos do mundo.

O Plano Nacional de Educação (PNE), em vigor desde 2014, estabelece metas relacionadas à educação de nível superior, para os dez anos subsequentes. A meta 12 consiste em elevar a taxa bruta de matrícula na educação superior de 32,1% para 50% e a taxa líquida de 17,4% para 33%, para a população de 18 a 24 anos, idades que correspondem ao ensino superior de acordo com a organização do sistema educacional brasileiro. Além disso, a meta 12 em conjunto com a meta 13 do PNE, prevê a expansão para, pelo menos, 40% das novas matrículas no ensino superior no segmento público e assegurar a elevação da qualidade da educação superior, ampliando a proporção de mestres e doutores do corpo docente em efetivo exercício (BRASIL, 2014).

No Brasil, 63,2% dos jovens, de 18 a 24 anos, pertencentes ao quinto da população com os maiores rendimentos (20% mais ricos) frequentavam o ensino superior em 2018, enquanto esse percentual era de apenas 7,4% para jovens na mesma faixa etária e que pertenciam ao quinto da população com menores rendimentos (20% mais pobres). Além disso, a taxa ajustada de frequência

escolar líquida, que representa a proporção de pessoas que frequentam ou já concluíram o nível de ensino adequado à sua faixa etária, da população de 18 a 24 anos era de 25,2%. A maioria (63,8%) dos jovens nessa faixa etária não frequentava escola e não possuía ensino superior completo e 11% frequentavam a escola fora da etapa educacional adequada (IBGE,2019).

Apesar dos programas reformulados e implementados terem acarretado o aumento da população brasileira no ensino superior, estima-se que somente 17,4% da população com 25 anos ou mais de idade possuíam graduação completa em 2019 (PNAD, 2019).

Embora sejam visíveis os impactos positivos dos programas inclusivos, há que se entender que após o ingresso no ensino superior, os estudantes ainda enfrentam diversos desafios, que se estendem no decorrer da graduação, os quais afetam, diretamente ou indiretamente, a sua permanência ou não nesse nível educacional.

Os fatores associados à evasão e à conclusão podem ser relacionados com atributos específicos das instituições de ensino e atributos individuais dos alunos e são classificados como internos ou externos às instituições, respectivamente. Os internos, relacionados a cada instituição de ensino, podem ser: infraestrutura, capacitação do corpo técnico e docente, políticas de assistência estudantil, oferta de atividades extracurriculares, dentre outros. Os principais fatores externos, ligados aos alunos, são: motivação, prestígio social da carreira e condições socioeconômicas (COSTA, 2018).

A evasão e a conclusão de alunos no ensino superior tem sido tema de grandes debates nos últimos anos, por apresentar grande interesse para os setores público e privado. Se por um lado as instituições públicas esperam retornos dos investimentos de recursos públicos destinados à formação dos discentes, por outro, as instituições privadas almejam reduzir a perda de receitas com os problemas desencadeados pela prolongação ou não diplomação destes alunos, já que estes usualmente permanecem matriculados em um número reduzido de disciplinas, impedindo que a instituição oferte a vaga a alunos que poderiam cursar e pagar por uma quantidade superior de disciplinas.

Lima Junior et al., (2012) afirma que cada instituição possui suas particularidades e que devem elaborar seus próprios indicativos, isto é, caracterizar quais dos fatores apontados pela literatura são mais relevantes em seu contexto, para que assim sejam elaboradas políticas institucionais eficazes de combate à evasão. Portanto, apesar da farta produção acadêmica sobre o tema, ainda há lacunas para a compreensão de fatores associados a cursos

específicos, como o caso do curso de Estatística. Sendo assim, para contribuir com o debate acerca do tema, buscamos identificar tais fatores.

Portanto, o principal objetivo deste artigo é analisar a evasão e a conclusão no curso de Bacharelado em Estatística no Brasil, identificando variáveis internas e externas que influenciam tais fenômenos. Além disso, buscamos traçar o perfil dos alunos que estão mais propensos a experimentar os eventos de interesse; identificar e quantificar as variáveis que aceleram ou retardam o tempo até a evasão ou a conclusão e predizer as funções de incidência acumulada para perfis selecionados de indivíduos, isto é, a probabilidade acumulada de um indivíduo estar livre de qualquer um dos eventos de interesse e sofrer um desses no tempo fixado.

Este artigo está dividido em cinco seções, dentre elas, a atual que introduz o tema, lista os objetivos e detalha a estrutura do trabalho. A segunda seção aborda estudos anteriores acerca do tema em questão, indicando alguns métodos que têm sido utilizados e principais conclusões dos autores. A terceira seção reúne os métodos quantitativos utilizados. A quarta seção trata do estudo empírico desenvolvido e é composta por informações acerca dos dados utilizados, da modelagem estatística e da análise dos resultados. Por fim, na última seção são apresentadas as principais conclusões encontradas no estudo.

2.ESTUDOS SOBRE CONCLUSÃO E EVASÃO NO ENSINO SUPERIOR NO BRASIL

O fluxo escolar no ensino superior no Brasil vem sendo amplamente estudado. Pesquisas empíricas associadas com o uso de métodos estatísticos têm apresentado variados resultados que ajudam a entender os fatores que afetam a permanência, a retenção, a evasão, a conclusão, entre outras situações de vínculo do aluno junto à instituição de ensino.

Para entender o comportamento da evasão nas universidades federais brasileiras, Costa (2018) traçou o perfil do aluno que evade nestas universidades e mensurou a influência de diversas variáveis sobre o risco de evadir. Foram utilizados os dados do Censo da Educação Superior dos anos de 2011 a 2016 para acompanhar estudantes de cursos presenciais com duração de 4, 5 e 6 anos. Para alcançar seus objetivos, foram utilizadas técnicas de análise de sobrevivência não paramétrica (Kaplan-Meier) e semi-paramétrica (Regressão de Cox).

Segundo Costa (2018), o Estimador de Kaplan-Meier permitiu identificar os grupos com maiores taxas de evasão nas universidades federais brasileiras, que são: homens, pretos, moradores no interior, cursando licenciatura, alunos

de curso noturno, não beneficiários de apoio social, não participantes de pesquisa ou extensão. Baseado nos coeficientes do modelo de Cox, concluiu que: ser mulher tende a reduzir em 19% o risco de evadir; o adicional de um ano na idade do indivíduo na admissão ao curso aumenta em 1,2% o risco de evadir; os pretos apresentaram um aumento no risco de evasão de 16,6% quando comparado aos brancos; os estudantes que receberam algum tipo de auxílio financeiro tiveram o risco de evadir reduzido em cerca de 20% e por fim, os alunos que participaram de atividades de pesquisa ou extensão por um ano têm um risco 99% menor de evadir quando comparado a estudantes que não realizaram essas atividades.

Com o objetivo de analisar a evasão e retenção nos cursos de graduação, Lima Junior et al., (2012) utilizou métodos de análise de sobrevivência para investigar o fluxo escolar nos cursos de graduação em física na Universidade Federal de Rio Grande do Sul. Por meio dos dados fornecidos pelo Departamento de Controle e Registro Acadêmico (DECORDI/ UFRGS), o autor concluiu que homens e mulheres são igualmente propensos a evadir, no entanto, as mulheres demoram mais a decidir pela evasão, uma vez que o tempo de permanência até evadir é mais longo entre as mulheres. Além disso, no curso de Física, a decisão do abandono não é tomada de maneira precoce, ocorrendo após anos de retenção. O autor também concluiu que os estudantes de licenciaturas tendem a ter um tempo de retenção superior em relação aos de bacharelado. Adicionalmente, concluiu que os indivíduos que tiveram maiores pontuações no vestibular, ficam retidos por menos tempo até a conclusão em relação aos indivíduos com menores pontuações.

Com base em um estudo de caso para a Universidade de Brasília – UnB, os autores concluíram que houve aumento da evasão e redução da retenção dos estudantes que ingressaram entre 2002 e 2008 e que tiveram sua situação na UnB avaliada no ano de 2016. Além disso, enfatizam que o método proposto no estudo pode ser empregado para especificar o perfil socioeconômico do aluno que abandona assim como o perfil curricular dos cursos em que a evasão é maior.

Saccaro et al. (2019) analisaram fatores associados à evasão de estudantes que ingressaram no ano de 2009 e foram acompanhados até o ano de 2014 nos cursos das áreas de Ciências, Matemática e Computação e Engenharia, Produção e Construção do ensino superior brasileiro. Foi utilizado um método de Análise de Sobrevivência que estima os parâmetros de um modelo de aceleração, mais precisamente o modelo Accelerated Failure Time (AFT) com a distribuição lognormal.

A partir dessa aplicação utilizando bases de dados do Censo da Educação Superior, os autores concluíram que a evasão é maior nas instituições privadas; que ser homem e ter mais idade diminui o tempo de permanência do indivíduo no ensino superior, precisamente que o tempo de permanência de estudante do sexo feminino é 1,06 vezes maior do que o dos homens; e que alunos contemplados com apoio financeiro apresentam uma maior retenção, ou seja, os alunos contemplados com a Bolsa Permanência têm o seu tempo de permanência elevado para 1,7 vezes, quando comparado com aqueles que não são beneficiários. Além disso, concluíram que os alunos que têm uma atividade remunerada apresentam um tempo de permanência 2,6 vezes maior que aqueles que não possuem (SACCARO et al., 2019).

Em um estudo sobre a evasão no Ensino Superior, Silva et al. (2018) buscou identificar o perfil dos alunos evadidos do curso de Estatística da Universidade Federal da Paraíba, e analisar o tempo até a evasão destes alunos. Foram utilizadas técnicas descritivas para análise do perfil, e modelos de regressão em análise de sobrevivência para análise do tempo até a evasão dos alunos, mais precisamente o modelo Log-normal, que foi o mais adequado para os dados.

Verificou-se que para a variável sexo, mais de 50% das evasões são do sexo masculino, quanto à forma de ingresso, 72,9% desses alunos ingressaram a partir processo seletivo seriado, além disso, não houve diferenças significativas nos percentuais de evasão para alunos que cursaram o ensino básico em escola pública ou privada. Para verificar fatores relacionados ao tempo até a evasão, foi possível concluir através da razão dos tempos medianos de sobrevivência que alunos naturais de João Pessoa e que se declaram como não brancos possuem um tempo mediano de sobrevivência maior que aqueles que são provenientes de outras localidades e que se declararam como brancos (SILVA et al., 2018).

Os trabalhos citados não cobrem toda a produção acadêmica, mas ajudam a ilustrar a diversidade de resultados encontrados no estudo do tema, assim como a variedade de métodos quantitativos aplicáveis.

3.MÉTODOS EM ANÁLISE DE SOBREVIVÊNCIA PARA RISCOS COMPETITIVOS

A análise de sobrevivência, amplamente conhecida também como análise de sobrevida, consiste em um conjunto de métodos e técnicas estatísticas que são utilizados quando pretende-se analisar dados referentes ao tempo até a ocorrência de um determinado evento de interesse, chamado de

falha ou desfecho, com tempo inicial e final predefinidos e intervalo entre tais tempos conhecido como tempo de estudo (COLOSIMO; GIOLO, 2006). Os principais métodos estatísticos utilizados na análise de sobrevivência são fundamentados nos estudos realizados por Edward L. Kaplan e Paul Meier (1958); David Roxbee Cox e de Wayne Nelson (1972) e Odd Olai Aalen (1978).

Uma característica particular inerente aos dados de sobrevivência é a presença de censura. Entende-se por censura a não ocorrência do evento de interesse durante o tempo de estudo. Existem diferentes tipos de censura, dentre elas destaca-se a censura à direita, a mais frequente e encontrada neste estudo, quando o evento de interesse não ocorreu durante o tempo do estudo (KLEIN; MOESCHBERGER, 2003; COLLET, 2003;COLOSIMO; GIOLO, 2006).

Usualmente a análise de sobrevivência é utilizada em situações em que há apenas um evento de interesse e que ocorre uma única vez e são denominados de eventos únicos. O estimador não paramétrico e o modelo de regressão mais usuais para estimar as principais métricas, na presença de eventos únicos de interesse, são o estimador de Kaplan-Meier (KM) (KAPLAN; MEIER, 1958) e o modelo de Cox (COX, 1972).

Há situações em que é possível experimentar o evento mais de uma vez ou diferentes eventos decorrentes de um mesmo fator de risco, os chamados eventos múltiplos, como por exemplo gestações ou diversos efeitos de um tratamento. Um caso particular de eventos múltiplos é quando a ocorrência de um dos eventos de interesse impede a observação dos demais, chamado de riscos competitivos (CARVALHO et al., 2011). Este é o caso deste estudo, ao observar o abandono do curso de Estatística para um indivíduo, impede que seja observada a diplomação do mesmo, portanto, a causa do desfecho registrada para cada indivíduo será, então, aquela que ocorrer primeiro.

As estimativas de Kaplan-Meier e o modelo de regressão de Cox se tornam enviesados na presença de riscos competitivos porque superestimam as probabilidades de desfecho. Esse viés surge pois o estimador de KM assume que os eventos são independentes e, portanto, censura os demais eventos complementares ao evento de interesse. Ao considerar todos os riscos concorrentes como censura não-informativa, assume-se que o evento de interesse pode vir a acontecer, o que não é verdade neste caso. Isto ocorre porque se um dos eventos competitivos ocorre para um indivíduo, isto impossibilita a ocorrência de outro evento concorrente para o mesmo indivíduo (PINTILIE, 2006).

Portanto, dado que os métodos usuais não são adequados na presença de riscos competitivos, foram propostas abordagens alternativas, como a função

de incidência acumulada e modelo de riscos competitivos de Fine e Gray (1999), as quais serão apresentadas a seguir.

Conforme dito anteriormente, há situações em que o evento pode ocorrer por uma dentre k causas (k = 1, ..., m), sendo que ao observar uma das k causas impede de observar qualquer outra, portanto, é considerado o tempo até a primeira causa k observada. Com isso, há o interesse em métodos que considerem a presença dessas k causas, uma vez que se tem o objetivo de estimar a probabilidade de desfecho por cada causa. Na literatura, dados com essa característica são referenciados como riscos competitivos e possuem métodos estatísticos específicos.

As funções de grande relevância na análise de sobrevivência na presença de riscos competitivos são a função de incidência acumulada (FIA) associada à causa k, k = 1, …, m ($F_k(t)$) e a função risco causa-específica ($\lambda_k(t)$). A primeira, fornece em cada tempo t a probabilidade acumulada de desfecho devido à k-ésima causa, considerando a presença de todas as demais causas. A segunda, representa a taxa instantânea da ocorrência do k-ésimo desfecho no tempo t, condicionada a sobrevivência até o tempo t, na presença de todas as outras causas de desfecho possíveis.

3.1 FUNÇÃO DE INCIDÊNCIA ACUMULADA

Um estimador não paramétrico que estima diretamente a função de incidência acumulada causa-específica é definido como a soma em todos os tempos $t_j \leq t$ das probabilidades de se observar o evento k no tempo t_j entre indivíduos que não sofreram qualquer dos possíveis eventos imediatamente antes desse instante, isto é, a FIA é a probabilidade acumulada de um indivíduo estar livre de qualquer evento imediatamente antes de t_j e sofrer o evento k no tempo t_j (KLEIN; MOESCHBERGER, 2003, CARVALHO et al., 2011). Portanto, o estimador da FIA para cada causa k, depende do número de indivíduos que experimentaram o evento de interesse e também do número de indivíduos que não tenham experimentado qualquer um dos eventos até aquele momento.

Pode-se ter o interesse em incorporar características dos indivíduos em estudo. É possível analisar o comportamento das FIA's entre as s categorias ($s \geq 2$) de uma dada covariável. Utiliza-se então o teste de Gray, que para a causa k, a hipótese nula associada a esse teste é dada pela igualdade das FIA's nas s categorias. Para s grupos a estatística de teste associada segue distribuição qui-quadrado com $s - 1$ graus de liberdade, denotada por χ^2_{s-1} (GRAY, 1988).

3.2 MODELO DE FINE E GRAY

Fine e Gray (1999) introduziram uma forma de estimar o efeito das covariáveis modelando diretamente as FIA's de uma particular causa de desfecho k em um ambiente de riscos competitivos, em que não é preciso supor independência entre os tempos dos eventos competitivos. Além disso, na presença de um vetor de covariáveis $\boldsymbol{x}$, o modelo assume riscos proporcionais, ou seja, efeito constante das covariáveis ao longo do tempo

A função de verossimilhança considerada neste contexto é dita parcial pois na definição do grupo de risco $R(t_{j(i)})$ considera-se os indivíduos que sofreram eventos competitivos, não somente aqueles livres de qualquer evento. Mas vale ressaltar que estamos em uma abordagem diferente, o que nos leva a repensar acerca da contribuição de cada indivíduo no cálculo. Portanto, a cada indivíduo será atribuído um peso que muda a cada tempo t_j no qual ocorre o evento de interesse, afinal, o indivíduo que já experimentou um evento competitivo não pode contribuir da mesma maneira que um indivíduo que não experimentou qualquer um dos eventos. Para mais detalhes sobre a função de verossimilhança e os pesos ver Pintilie (2006) e Carvalho et al., (2011).

Acerca dos coeficientes estimados na modelagem, para testar a significância e os diferentes modelos, as covariáveis foram selecionadas considerando os critérios de seleção de variáveis vistos em Carvalho et al. (2011), os quais são: teste de Wald, que testa a hipótese nula de que o parâmetro de regressão é igual a zero; e teste da razão de verossimilhanças, o qual compara modelos aninhados avaliando se a inclusão de uma ou mais covariáveis no modelo aumenta de modo significativo a verossimilhança de um modelo em relação a outro modelo mais parcimonioso.

Para a seleção de variáveis, considera-se a presença de uma covariável por vez, e a partir deste passo inclui-se no modelo todas as variáveis significativas ao nível de 10%, retornando e retirando as covariáveis que não foram significativas na presença das demais. O conjunto de covariáveis que se manteve no modelo final foi avaliado a um nível de significância mais rigoroso (5%) e, além disso, foram investigadas as interações entre as covariáveis.

Acerca da interpretação, foi utilizada a razão de taxas de falha dada por $\exp(\beta_p)$, que representa a razão entre as taxas de dois indivíduos, i e j, que têm os mesmos valores para as covariáveis com exceção da p-ésima, portanto, mantendo as demais covariáveis fixas.

Para verificar a suposição de riscos proporcionais e a bondade do ajuste do modelo, foram utilizados os resíduos de Schoenfeld, que investiga a

proporcionalidade para cada covariável do vetor $\boldsymbol{x}$, definidos para cada indivíduo i. Além da análise gráfica dos resíduos pode-se testar a presença de correlação linear entre o tempo e o resíduo. Logo, se a hipótese nula (H_0: $\rho=0$) não for rejeitada, temos que a premissa de proporcionalidade dos riscos não é rejeitada (CARVALHO et al., 2011).

Uma medida global de ajuste, é a estimativa da probabilidade de concordância (pc), que é usada para avaliar o poder discriminatório e a acurácia preditiva do modelo, sendo que há diferentes interpretações na literatura a depender do valor da pc, aqui adotou-se as interpretações dadas em Carvalho et al., 2011.

Foi utilizado o estimador de máxima verossimilhança para estimar os efeitos das covariáveis sobre cada evento de interesse, modelando diretamente as funções de incidência acumulada (FIA), para tal foi utilizada a função *crprep* da biblioteca *mstate* no software R. Essa função calcula os pesos que cada indivíduo assume à medida que ocorrem eventos competitivos, é repetido o mesmo procedimento para cada um dos eventos de interesse, assumindo os demais como competitivos.

Após o procedimento realizado para o cálculo dos pesos, utiliza-se as mesmas técnicas adotadas no ajuste do modelo de Cox, por meio da biblioteca *survival*, a função *survifit* estima a incidência acumulada para os eventos de interesse, utilizando o argumento *etype* e, sendo de interesse, pode-se incorporar covariáveis. Além disso, com a função *coxph* estima-se os efeitos das covariáveis, indicando o evento de interesse no argumento *failcode*. Para melhor compreensão e aplicação das funções mencionadas para análise de riscos competitivos no *software* R, ver Carvalho et al. (2011). Além disso, os códigos utilizados neste estudo estão disponíveis em um projeto no *RPubs* da autora principal, ver Portela (2023).

4.APLICAÇÃO: O CURSO DE ESTATÍSTICA

O estudo refere-se a uma coorte de indivíduos que ingressaram no curso de bacharelado em Estatística no Brasil no ano de 2010 e foram acompanhados até o ano de 2017. Os resultados mostrados nessa seção incluem a probabilidade de desfecho considerando como eventos de interesse: conclusão e evasão, estimadas por meio de função de incidência acumulada (FIA), estimação dos efeitos das covariáveis sobre o risco de experimentar um dos eventos de interesse, tendo o outro evento como concorrente e predições das curvas das FIA's de acordo com perfil de indivíduos selecionados.

4.1. FONTE, DADOS E POPULAÇÃO EM ESTUDO

O Censo da Educação Superior (Censup) que, de acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), é o instrumento de pesquisa mais completo do Brasil acerca da educação superior desde 1995. O Censup é realizado anualmente pelo Inep em parceria com todas as instituições de ensino superior (IES), as quais são responsáveis por fornecer ao instituto informações detalhadas sobre quatro unidades de investigação: IES, curso, discente e docente. É possível acessar os dados do Censup no sítio do Inep e as informações podem ser obtidas via download, em formato ASCII.

No presente estudo, foram considerados os microdados referente à coorte de ingressantes no curso de Bacharelado em Estatística no ano de 2010, pois somente a partir de 2009 foram disponibilizados dados desagregados para o nível aluno e, por ser um ano de transição e familiarização, optou-se por não utilizar este ano como o início do período do estudo. Para acompanhar a situação destes alunos na IES, foram considerados os anos seguintes até 2017, pois a partir de 2018 houve recodificação da variável código do aluno, atribuída para cada discente vinculado a uma IES e, consequentemente, não foi possível manter o acompanhamento individual para os anos seguintes. Sendo assim, o recorte considerado no estudo foi o maior tempo possível dentro das limitações apresentadas.

Na base de dados referente ao ano de 2010 há 86 variáveis correspondentes à unidade de investigação, a saber, aluno. Dentre as variáveis disponíveis, algumas foram desconsideradas por não possuírem significado para o estudo ou não serem de interesse, tais como as que designam informações para uso administrativo, as que possuem o mesmo significado que outra ou as que possuem elevadas porcentagens de dados faltantes, neste estudo o critério para exclusão de covariáveis foi de 35% ou mais de dados faltantes.

As variáveis código do aluno e código da situação do aluno foram utilizadas para conectar as bases de dados do Censup do ano de 2010 com os anos seguintes, com o intuito de traçar a trajetória destes alunos no curso ao longo dos anos. Isso foi feito utilizando a variável que indica o código da situação do aluno, que nesse estudo foi recategorizada. Considerou-se o aluno ativo, aquele que estava cursando ou com a matrícula trancada; evadido, aquele que havia se desvinculado da IES ou transferido para outro curso na mesma IES e, por fim, concluso, aquele aluno que havia se formado (Figura 1). Quando o aluno passa da situação de ativo para uma situação dita de interesse, evadido ou concluso, a nova situação do aluno permanece a mesma até o final do estudo.

Figura 1 – Recategorização da situação do aluno

Fonte – Elaboração própria com base nos dados do Censup - INEP.

Em 2010, um total de 1.612 alunos iniciaram o Bacharelado em Estatística em uma das 32 IES do Brasil. Entre eles, 991 (61,5%) são do sexo masculino e 621 (38,5%) do sexo feminino (Tabela 1). A média da idade ao ingressar no curso foi de aproximadamente 22,8 anos (desvio padrão de 6,7 anos), com idade mínima de 16 anos e máxima de 65 anos. O número médio de alunos por IES foi de 50,4 alunos, sendo que a região Sudeste possuía o maior número de IES e um total de 771 (47,8%) ingressantes.

Tabela 1 - Distribuição, absoluta e percentual, dos alunos, segundo características selecionadas, Brasil, 2010

Características selecionadas	Número de alunos de acordo com as características selecionadas	
	Número de alunos	Distribuição percentual (%)
Total	**1.612**	**100**
Sexo	**1.612**	**100**
Feminino	621	38,5
Masculino	991	61,5
Turno (noturno)	**1.612**	**100**
Sim	694	43,1
Não	918	56,9
Organização acadêmica (universidade)	**1.612**	**100**
Sim	1.484	92,1
Não	128	7,9
Reserva de vaga	**1.612**	**100**
Sim	124	8,3
Não	1.488	91,7
Apoio social	**1.612**	**100**
Sim	157	9,7
Não	1.455	90,3
Ingresso por vestibular	**1.612**	**100**
Sim	1.182	73,3
Não	430	26,7
Atividade complementar	**1.612**	**100**
Sim	59	3,7
Não	1.553	96,3

Fonte - INEP. Censo da Educação Superior - Censup, 2010

Acerca do turno que frequentavam o bacharelado, o período noturno conta com 43% dos alunos, enquanto 33% frequentam o período integral, 22% o período matutino e 2% o vespertino. No total, 59 alunos participaram de algum tipo de atividade complementar de formação, em que 32 (55%) receberam ao menos algum tipo de auxílio financeiro.

Temos que dos 1612 alunos, 73,3% ingressaram por meio de vestibular próprio da IES, sendo que 1484 (92,1%) ingressaram em uma organização acadêmica do tipo universidade e os demais 128 (7,9%) ingressaram em organizações do tipo faculdade ou centro universitário.

Acerca dos ingressantes por algum tipo de reserva de vaga (deficiência, cunho étnico, ensino público, renda familiar e outros) oferecida pela IES, temos um total de 124 (8,3%). Vale ressaltar que no ano de 2010 ainda não tinha sido sancionada a Lei de Cotas no Brasil, portanto, cada IES decidia ter ou não um quantitativo de vagas reservadas para esta finalidade. Além disso, do total de ingressantes, temos que 157 (9,7%) recebiam algum tipo de apoio social (alimentação, bolsa permanência, bolsa trabalho, material didático, moradia e transporte) pela IES e somente 59 (3,7%) estavam engajados em alguma atividade de formação complementar (pesquisa, projeto de extensão, monitoria e/ou estágio).

4.2. ANÁLISE DOS RESULTADOS

Dentre outras medidas resumo do tempo em anos até a ocorrência dos eventos de interesse, destacam-se o tempo médio até a evasão (3,3 anos) e o fato de que 75% dos alunos evadiram em até 5 anos de curso. A conclusão teve tempo médio de 5,4 anos, mas dentre os alunos que formaram, apenas 25% deles concluíram o curso em até 4 anos (Tabela 2).

Tabela 2 - Medidas resumo do tempo em anos, segundo o evento de interesse, Brasil, 2010-2017

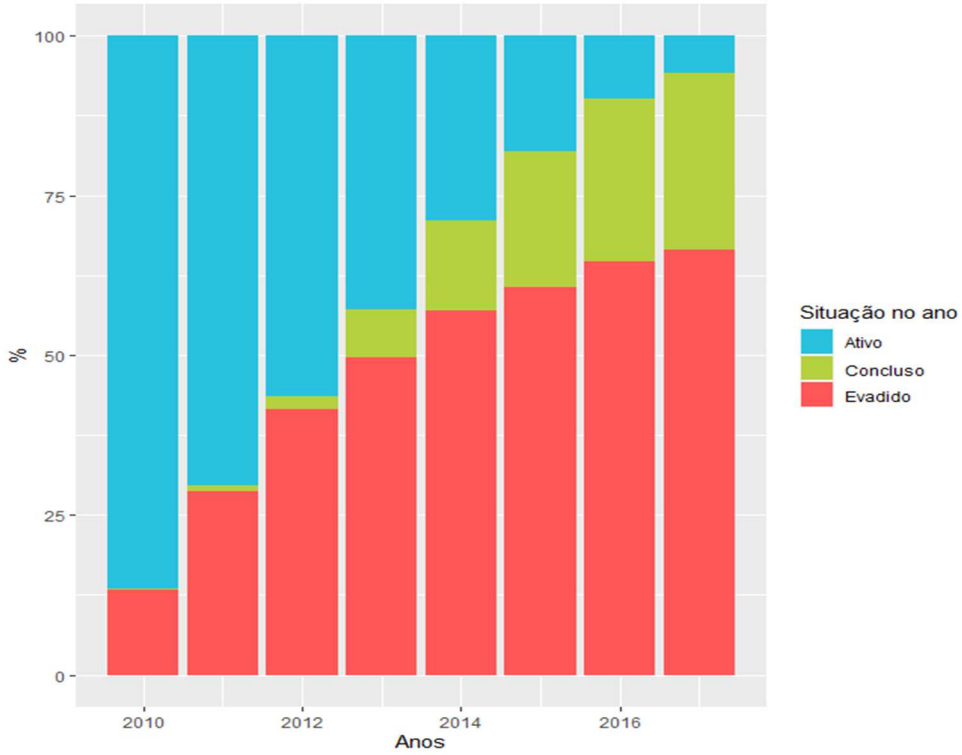
Evento de interesse	Média	Desvio	Mínimo	Q1*	Mediana	Q3**	Máximo
Evasão	3,25	1,92	1	2	3	5	8
Conclusão	5,42	1,44	1	4	5	6	8

*primeiro quartil **terceiro quartil.

Fonte - Elaboração própria com base nos dados do Censup - INEP

Ao final dos 8 anos estudados, 95 dos 1.612 alunos (5,9%) estavam ativos. Os demais alunos, 1.517 (94,1%), experimentaram um dos dois eventos de interesse: 445 (27,6%) concluíram e 1.072 (66,5%) evadiram (Tabela 2; Figura 2).

Figura 2 - Percentual acumulado da situação dos alunos ao longo dos anos, Brasil, 2010-2017



Fonte – Elaboração própria com base nos dados do Censup – INEP.

Os maiores percentuais de alunos que evadiram são vistos nos três primeiros anos. A conclusão apresenta percentuais pequenos nos primeiros anos, mostrando aumento mais significativo a partir do quarto ano, o que é esperado dado que se faz necessário cursar disciplinas antes de estar apto a concluir o curso, exceto para os alunos reingressantes ou transferidos com parte das disciplinas já cursadas.

Além disso, é possível verificar que 324 (21,4%) alunos ficaram retidos até a conclusão, que equivale a soma dos alunos que experimentaram a conclusão em um tempo maior ou igual a 5 anos (Tabela 3). Em contrapartida, o total de alunos que ficaram retidos e evadiram do curso foi 270 (17,8%) . No final do estudo restaram 95 alunos ativos, ou seja, censurados por não experimentarem um dos eventos de interesse até o fim do estudo. O tempo de permanência para estes alunos foi computado como sendo 8 anos, que equivale ao tempo do estudo.

Tabela 3 - Distribuição, absoluta e percentual, do número de alunos, segundo o tempo até experimentar um dos eventos de interesse, Brasil, 2010-2017

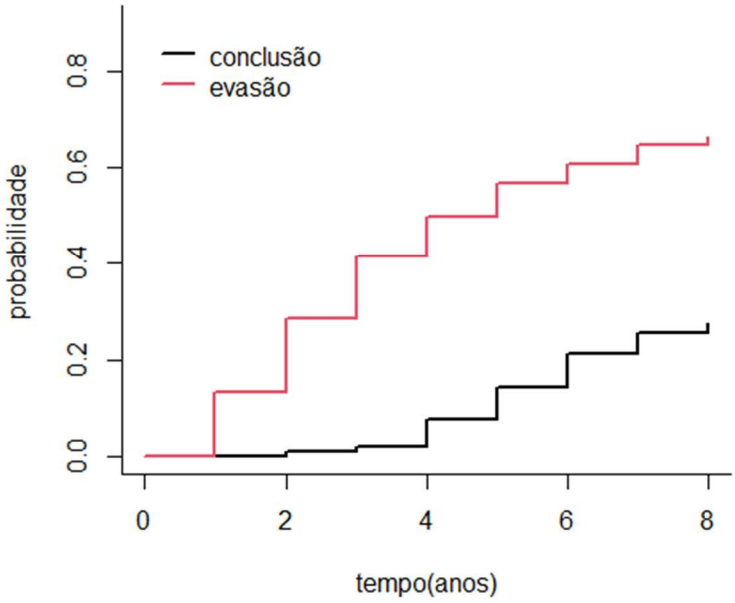
Tempo (em anos) até evento de interesse	Alunos		
	Quantidade (absoluto e %)	Conclusão (absoluto e %)	Evasão (absoluto e %)
Totais	**1517 (100)**	**445 (29,30)**	**1072 (70,70)**
1	216 (14,24)	1 (0,07)	215 (14,17)
2	260 (17,14)	13 (0,86)	247 (16,28)
3	224 (14,77)	18 (1,19)	206 (13,58)
4	223 (14,70)	89 (5,90)	134 (8,80)
5	224 (14,77)	108 (7,12)	116 (7,65)
6	172 (11,34)	112 (7,38)	60 (3,96)
7	133 (8,77)	67 (4,42)	66 (4,35)
8	65 (4,28)	37 (2,44)	28 (1,84)

Fonte - Elaboração própria com base nos dados do Censup – INEP

Preliminarmente, para verificar como as características dos alunos estão associadas ao evento de interesse, foi utilizada a função de incidência acumulada (FIA) para estimar a probabilidade de desfecho considerando como eventos de interesse: conclusão e evasão.

As curvas da FIA referentes aos dois eventos de interesse indicam que um aluno está mais propenso a vir a experimentar primeiro a evasão do que a conclusão (Figura 3). Nos primeiros anos é esperado que a FIA para a conclusão seja próxima de zero, dado que o curso de Bacharelado em Estatística tem duração típica de 4 anos. Ao final de 8 anos, a probabilidade de os alunos terem concluído foi de aproximadamente 0,27, enquanto para a evasão, esta probabilidade já era superior nos primeiros anos (Figura 3).

Figura 3 - Função de incidência acumulada para os eventos de interesse definidos, Brasil, 2010-2017



Fonte – Elaboração própria com base nos dados do Censup – INEP

Foi utilizado o teste de Gray para verificar se duas categorias das variáveis selecionadas eram significativamente diferentes para cada eventos de interesse ao nível de significância de 5%. A hipótese alternativa é de que pelo menos uma categoria difere das demais. Por exemplo, para a variável sexo e para o evento conclusão as hipóteses (nula e alternativa) são:

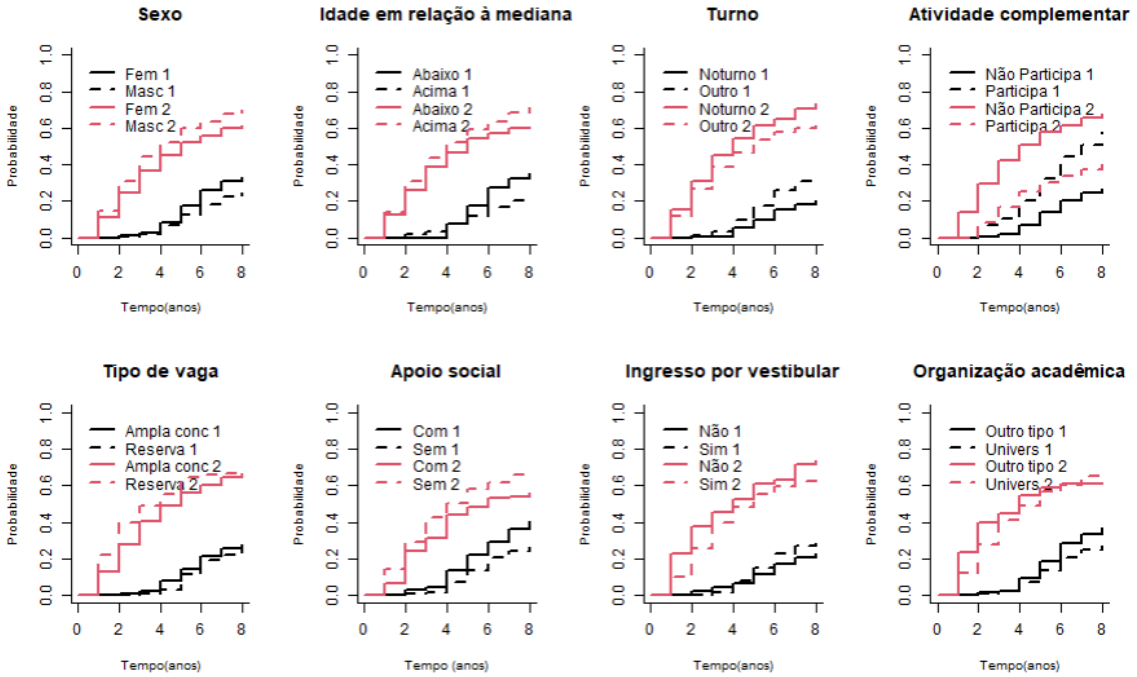
H_0: $F_{\text{conclusão\_feminino}} = F_{\text{conclusão\_masculino}}$ e H_1: $F_{\text{conclusão\_feminino}} \neq F_{\text{conclusão\_masculino}}$.

As variáveis que possuíam mais de 2 categorias, foram dicotomizadas para serem analisadas, de acordo com uma categoria de interesse, como por exemplo, idade (maior e menor que a mediana) e turno que frequenta (noturno e não noturno).

Foram analisadas 8 covariáveis: sexo, idade, turno, participação em atividades complementares, apoio social, forma de ingresso, uso de reserva de vaga e tipo de organização acadêmica, a fim de traçar o perfil dos alunos com maior probabilidade de evasão e conclusão. As variáveis uso de reserva de vaga e organização acadêmica não foram significativas para o evento de interesse evasão, sendo que para o evento conclusão apenas a última foi significativa. Para as demais variáveis, as curvas diferem entre si, ao nível de significância de 5%, (valor-p < 0,5), o que indica que é possível traçar o perfil dos alunos com

maior probabilidade de conclusão ou evasão, com base em tais características (Figura 4).

Figura 4 - Probabilidade estimada a partir da FIA para conclusão e evasão por covariável de interesse



Fonte – Elaboração própria com base nos dados do Censup - INEP

(*) 1=conclusão e 2=evasão

Os resultados indicam que estão mais propensas a concluir o Bacharelado em Estatística (linhas pretas) as mulheres, com idade abaixo da mediana, em curso não noturno, que recebem algum tipo de apoio social, fazem atividade complementar, que ingressam por vestibular próprio da IES e são matriculadas em IES do tipo faculdade ou centro universitário. Do contrário, homens, com idades acima da mediana, matriculados em IES do tipo universidade, em curso noturno, sem apoio social, que não realizam atividade complementar e não entraram por vestibular próprio estão mais propensos à evasão (linhas vermelhas).

4.3. MODELO PROPOSTO E PREDIÇÕES

Foram consideradas 8 covariáveis para ajuste do modelo: sexo, idade, turno, participação em atividades complementares, apoio social, forma de ingresso, uso de reserva de vaga e tipo de organização acadêmica. Após a seleção de variáveis permaneceram no modelo final denominado Modelo Conclusão, que considera o evento de interesse conclusão e a evasão como evento competitivo, as covariáveis: sexo, idade, turno, tipo de organização acadêmica e participação em atividades complementares. No modelo final denominado Modelo Evasão, que considera o inverso do anterior para o evento de interesse e o evento competitivo, as covariáveis que se mantiveram foram: sexo, idade, turno e participação em atividade complementar. Em relação às interações, não foram observadas interações significativas entre as covariáveis para qualquer dos dois modelos.

Para verificar visualmente se ao longo do tempo há uma possível mudança dos efeitos das covariáveis que permaneceram no modelo final, foram analisadas as distribuições dos resíduos de Schoenfeld para cada covariável. Além disto, foi testada a correlação linear dos resíduos com o tempo para cada covariável. Tanto a distribuição dos resíduos quanto os testes não rejeitaram a hipótese nula de inexistência de correlação, isto é, não se rejeita, com nível de significância de 5%, a proporcionalidade dos riscos em todas as covariáveis do estudo sobre cada um dos eventos de interesse considerados.

A probabilidade de concordância para o Modelo Conclusão foi de 0,648, e para o Modelo Evasão foi de 0,575. Portanto, ambos os modelos mostraram um resultado comum para dados de sobrevivência, segundo a regra geral discutida em Carvalho et al. (2011).

O Modelo Conclusão indica que a taxa de falha da conclusão do curso de estatística para as mulheres é de 1,45 vezes a taxa de conclusão dos homens. Além disso, a cada aumento de um ano na idade, a taxa de conclusão diminui em 6%. Os alunos matriculados em uma IES do tipo universidade têm a taxa de conclusão reduzida em 45% quando comparado com o grupo dos alunos matriculados em IES do tipo faculdade ou centro universitário. Para os alunos que frequentam o curso noturno a taxa de conclusão é 37% menor em relação àqueles que frequentam os demais turnos. Por fim, o aluno estar envolvido em atividades complementares no primeiro ano da graduação aumenta a taxa de conclusão em 2,78 vezes quando comparada aos alunos que não participam de tais atividades (Tabela 4).

O Modelo Evasão mostra que a taxa de falha da evasão do curso de Estatística para as mulheres é 20% menor que a taxa de evasão para os homens. A cada aumento de um ano na idade, a taxa de evasão aumenta em 1,8%. Para os alunos que frequentam o curso noturno a taxa de evasão é 1,26 vezes a taxa de evasão para os alunos que não frequentam o curso noturno e, por fim, os alunos que fazem atividade completar no primeiro ano da graduação possui a taxa de evasão 55% menor que a taxa dos alunos que não fazem atividade complementar (Tabela 4).

Tabela 4 - Estimativas dos efeitos das covariáveis para o modelo Fine-Gray considerando os eventos de interesse conclusão e evasão

Modelo	Covariáveis	$\hat{\beta}$	exp($\hat{\beta}$)	Z	Valor p
Conclusão	Sexo Fem.	0.369	1.446	3.845	< 0.01
	Idade	-0.059	0.943	-5.493	< 0.01
	Acadêmica Univ.*	-0.594	0.552	-3.796	< 0.01
	Turno Nort.	-0.470	0.625	-4.523	< 0.01
	Atividade Sim**	1.023	2.782	5.662	< 0.01
Evasão	Sexo Fem.	-0.219	0.803	-3.423	< 0.01
	Idade	0.019	1.018	4.588	< 0.01
	Turno Nort.	0.232	1.261	3.747	< 0.01
	Atividade Sim**	-0.799	0.449	-3.865	< 0.01

*tipo de organização acadêmica **participação em atividades complementares

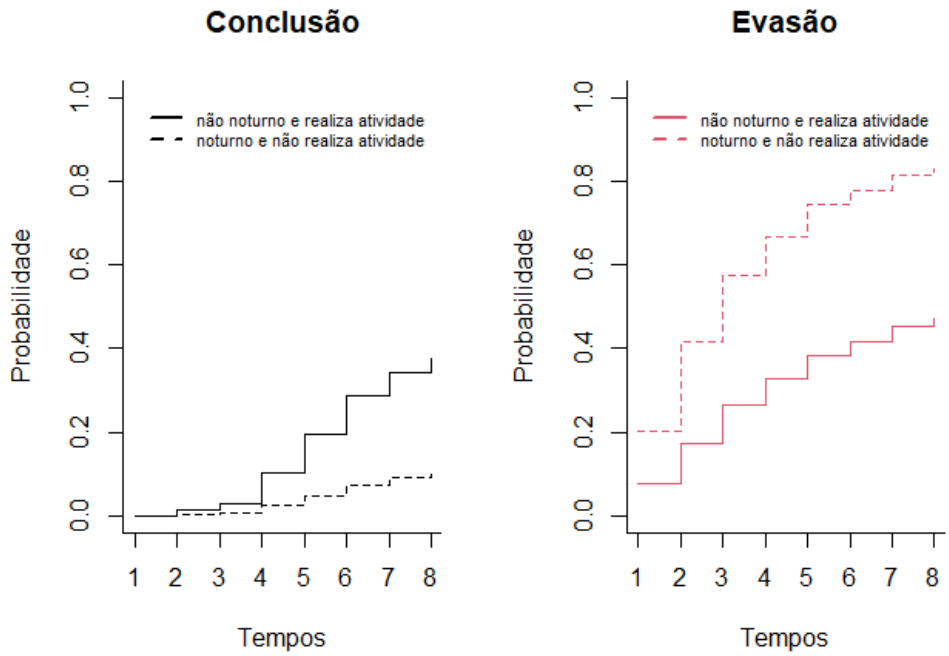
Fonte - Elaboração própria com base nos dados do Censup – INEP.

A seguir são apresentadas algumas predições com base nas funções de incidência acumulada (FIA) estimadas a partir dos dois modelos ajustados para perfis selecionados de indivíduos.

Como ilustração, para analisar as taxas de falha de conclusão e de evasão de alunas que estudam em diferentes turnos e com diferentes engajamentos em atividades complementares, foram consideradas duas mulheres de 25 anos de idade que estudam em uma organização acadêmica de tipo não universidade. A primeira aluna está matriculada em curso não noturno e realiza atividade complementar; a segunda aluna está matriculada em curso noturno e não realiza atividade complementar. Vemos que o fato de frequentar curso não noturno e fazer atividade complementar acelerou a taxa de falha da conclusão (Figura 5). O contrário foi verificado em relação à evasão, ou seja, o

fato de estar matriculado no curso noturno e não fazer atividade complementar aumentou significativamente a taxa de evasão (Figura 5).

Figura 5 - Predições conforme estimativas do modelo Fine e Gray para o evento de interesse conclusão e evasão, considerando perfis selecionados



Fonte – Elaboração própria com base nos dados do Censup – INEP

5.CONCLUSÕES

O estudo apontou que o sexo, a idade do aluno ao ingressar no curso, o tipo de organização acadêmica da IES, o turno do curso frequentado e a prática de atividade complementar possuem efeitos significativos quando o evento de interesse é a **conclusão**. Vimos que ser do sexo masculino, o acréscimo de um ano na idade, ser aluno de uma IES do tipo universidade e frequentar o curso noturno, diminuem a taxa de conclusão, mas que o oposto ocorre quando o aluno participa de atividades complementares. Para o evento de interesse **evasão**, verificou-se que o sexo, a idade, o turno e a atividade complementar foram as covariáveis significativas. Foi possível observar que ser do sexo feminino e fazer atividade complementar diminuem a taxa de evasão, no entanto, frequentar o curso noturno e o acréscimo de um ano na idade, tornam os alunos mais propensos a experimentar a evasão.

Nas predições das FIA's para comparar indivíduos com dois perfis diferentes, vimos que o fato de frequentar curso noturno e não estar inserida em

atividades complementares aumenta a taxa de evasão e diminui a taxa de conclusão, do contrário, aumenta a taxa de conclusão e diminui a taxa de evasão.

As variáveis significativas para o curso de Estatística apontadas pelo estudo acompanham os resultados encontrados na literatura para outros cursos, apesar das diferentes técnicas utilizadas. A utilização de métodos da análise de sobrevivência possibilitou um estudo longitudinal da evasão e da conclusão, permitindo uma melhor compreensão da variação do fenômeno ao longo do tempo. Observou-se que 66,5% dos alunos que ingressaram em 2010 haviam evadido do curso até 2017, enquanto somente 27,7% haviam concluído, sendo que nos primeiros anos foram observados os maiores percentuais de evasão.

Uma importante conclusão deste estudo foi a forte influência da participação em atividades complementares sobre a taxa de conclusão. Os resultados aqui discutidos mostram que participar das atividades de pesquisa, extensão ou monitoria eleva drasticamente a taxa de conclusão e reduz a taxa de evasão. Vimos que realizar atividades complementares no primeiro ano da graduação leva a taxa de conclusão em 2,78 vezes quando comparada aos alunos que não realizam tais tipos de atividades. Apesar de ser um resultado já apontado em outros estudos, sua corroboração no presente estudo, reforça a ideia de que investir na oferta de atividades complementares pode ser vista como uma alternativa para o combate à evasão a ser empregada nos primeiros anos do curso de Estatística.

De um modo geral, os modelos considerados demonstraram bons ajustes para o principal objetivo do estudo, que foi a identificação das variáveis que aceleram ou retardam as taxas de conclusão e de evasão no curso de Estatística.

Vale ressaltar que além da limitação do tempo de acompanhamento devido à mudança do código do aluno, como mencionado anteriormente, um outro fator limitante do estudo é acerca das conclusões descritas anteriormente. Os dados do estudo são secundários e tem a etapa da coleta de forma descentralizada com regime de colaboração e declaratório, englobando todos os estabelecimentos públicos e privados de educação superior, isto é, apesar do Inep disponibilizar treinamento EAD na plataforma Capacita Servidores com intuito de capacitar todos os envolvidos no preenchimento dentro da IES, o representante institucional, indivíduo investido de poderes para prestar informações em nome da instituição, pode compartilhar tarefas de inserção de dados. Portanto, a exatidão e fidedignidade das informações prestadas ao Censo Escolar depende da rigorosidade de cada recenseador institucional e de

seus auxiliares e, consequentemente, devemos ter cautela ao concluir os resultados encontrados sem desprezar esse fator limitante.

Por fim, recomenda-se que pesquisadores não desprezem a estrutura de riscos competitivos em seus estudos e, caso estejam interessados em predições, é preciso investigar a suposição de riscos proporcionais, fazendo um estudo minucioso acerca de possíveis variáveis que tenham efeitos que oscilam ao longo do tempo, para que assim sejam ajustados modelos mais adequados que incorporem estas características.

5.REFERÊNCIAS BIBLIOGRÁFICAS

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. The

Annals of Statistics, Philadelphia, v. 6, n. 4, p. 701-727.

BRASIL. Lei n.13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação

– PNE e dá outras providências. Diário Oficial da União, Brasília, DF., 26 jun 2014. Disponível em: <http://www. planalto.gov.br/ccivil_03/_ato2011-

2014/2014/lei/l13005.htm>. Acesso em: Ago. 2020.

Carvalho, M.S, Andreozzi, V.L., Codeço, C.T., Campos, D.P., Barbora, M.T.S. & Shimakura, S.E. (2011). Análise de sobrevivência: teoria e aplicações em saúde. 2 ed. Rio de Janeiro. Editora Fiocruz, 2011. 432p.

Collett, D. (2003). Modelling Survival Data in Medical Research. 2nd ed. London. Chapman and Hall/CRC, 410p.

Colosimo, E. A. & Giolo, S. R. (2006). Análise de sobrevivência aplicada. São Paulo: Editora Blucher, 370p.

Costa, F.P. (2018). Acesso e permanência no ensino superior: uma análise para as universidades federais brasileiras. 2018. 84f. Dissertações (Mestrado em Políticas Públicas) – Universidade Federal de Pernambuco, Recife - PE.

Cox, D.R. (1972) Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), London, v. 34, n. 2, p. 187-220.

Fine, J. P. & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association, New York, v. 94, n. 446, p. 496-509.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. Annals of Statistics, San Francisco, v. 16, p. 1141-1154.

IBGE. Síntese de indicadores sociais: uma análise das condições de vida da população brasileira - 2019. Coordenação de População e Indicadores Sociais. Nº 40 Rio de Janeiro: IBGE, 2019. Disponível em Síntese de Indicadores Sociais - 2019. Acesso em 29 de setembro de 2020.

Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, New York, v. 53, n. 282, p. 457-481.

Klein, J. P. & Moeschberger, M. L. (2003). Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer, 538p.

Lima Junior, P., Silveira, F. L. & Ostermann, F. (2012). Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. Revista Brasileira de Ensino de Física, v. 34, n. 1, pp.1-10, 1403.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. Technometrics, zichmond, v. 14, n. 4, p. 945-966.

Pesquisa Nacional por Amostra de Domicílios (PNAD) 2019. Disponível em: < https://www.ibge.gov.br/estatisticas/sociais/trabalho/17270-pnadcontinua.html?edicao=28203&t=publicacoes>. Acesso em: Set. 2020.

Plintilie, M. (2006). Competing Risks: a practical perspective. Chichester: John Wiley & Sons, Ltd, 224 p.

Pinto, J. M. de R. (2004). O acesso à educação superior no Brasil. Educ. Soc., Campinas, v. 25, n. 88, p. 727-756, Out. 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-73302004000300005&lng=en&nrm=iso>. Acesso em: Set. 2020.

Portela, A. C. T.(2023). Aplicação do TCC: Evasão e conclusão no curso de bacharelado em estatística: uma aplicação da análise de sobrevivência na presença de riscos competitivos. Rio de Janeiro-RJ, Brasil: Posit/Rpubs, 2023. Disponível em https://rpubs.com/AdrianePortela/1016548.

Saccaro, A., Franca, M. T. A.& Jacinto, P. A. Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. Estud. Econ., São Paulo, v.49, n. 2, p. 337-373, Apr. 2019.

Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-41612019000200337&lng=en&nrm=iso>. Acesso em: Dez. 2020.

Silva, A., Gondim Filho, A., Da Silva, C., Leite, D., Da Silva, L. & Freitas, W. (2018). Modelos de sobrevivência aplicados à evasão dos alunos de Estatística da UFPB. Revista InterScientia, v. 6, n. 2, p. 134-145, 7 dez. 2018.

Revista Brasileira de Estatística, V.81, N. 248, p.31-61, 2023

CIÊNCIA DE DADOS: UMA DESCRIÇÃO DOS PRIMEIROS CURSOS DE GRADUAÇÃO EM UNIVERSIDADES BRASILEIRAS

Anderson Ara

ara@ufpr.br

Universidade Federal do Paraná

Geisyane Karina Gonzaga Kath

geisyane.karina@ufpr.br

Universidade Federal do Paraná

Cristian Pessatti dos Anjos

cristian.anjos@ufpr.br

Universidade Federal do Paraná

Cibele Maria Russo

cibele@icmc.usp.br

Universidade de São Paulo

Wagner Bonat

wbonat@ufpr.br

Universidade Federal do Paraná

Resumo: Devido ao aumento de volume de dados, a urgência na busca de cientistas de dados devidamente qualificados tem crescido. Desta forma, as Instituições de Ensino Superior (IES) brasileiras têm buscado suprir tal demanda. Neste enredo, o objetivo deste artigo é realizar uma caracterização dos cursos de graduação em Ciência de Dados. Assim, buscou-se responder questionamentos como: os cursos têm sido ofertados em grande maioria pelas universidades públicas ou privadas? Quando começaram a ser ofertados? Costumam ser EAD (ensino à distância) ou presenciais? São do tipo tecnológico ou bacharelado? Quais grupos de disciplinas mais compõem a grade? Em quais regiões do país se concentram? Como é a oferta de vagas e qual é o perfil de ingressos em cursos do tipo bacharelado e tecnológico? Para isso, utilizou-se a junção das bases do e-MEC e do Censo da Educação Superior de 2021 e optou-se por fazer a exploração e visualização de dados considerando a técnica ACM. Entre os resultados, observa-se que há um certo equilíbrio entre as modalidades presencial e EAD, além de que em grande parte os cursos são do tipo tecnológico e costumam ser ofertados por IES privadas. Acerca das regiões, nota-se uma grande concentração de cursos presenciais na região Sudeste do Brasil.
Palavras-chave: Ciência de Dados; universidades brasileiras; graduação; ACM

Abstract: Due to the increasing volume of data, the urgency to look for suitably qualified data scientists has grown. Thus, Brazilian Higher Education Institutions (HEIs) have tried to answer this demand. In this scenario, the objective of this paper is to perform a characterization of undergraduate courses in Data Science. Thus, we aim to answer questions such as: have the courses been offered in the vast majority by public or private universities? When did they start being offered? Are they usually ODL (Online Distance Learning or in-person? Are they the technological type or baccalaureate? What groups of disciplines most make up the curriculum? In which regions of the country are they concentrated? How is the offer of vacancies and what is the profile of admissions in bachelor and technological courses? For this, the e-MEC databases and the 2021 Higher Education Census were combined, and it was decided to explore and visualize data using the MCA technique. Among the results, it is observed that there is a certain balance between the in-person and online learning modalities, in addition to the fact that most of the courses are of the technological type and are usually offered by private HEIs. Regarding the regions, a significant number of in-person undergraduate courses are concentrated in the Southeast region of Brazil.
Keywords: Data Science; brazilian universities; undergraduate; MCA

1.INTRODUÇÃO

Atualmente, têm-se disponível uma quantidade muito grande de dados nas mais diversas áreas de aplicação, auxiliando na tomada de decisões e impulsionando a sociedade em muitos aspectos e serviços, como varejo, serviços financeiros, manufatura, serviços móveis, entre outros (Agrawal et al., 2011). Assim, nota-se que não apenas as organizações dependem diretamente de dados, mas também o avanço científico, sendo que a Ciência de Dados tem se institucionalizado recentemente como educação formal (Curty & Serafim, 2016).

Neste cenário, a profissão de cientista de dados tem se tornado cada vez mais relevante e atrativa. Embora o termo ainda seja relativamente recente, segundo Monteiro-Krebs et al. (2021), foi formalmente utilizado pela primeira vez em 2001 por William S. Cleveland, no artigo *Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics* (Cleveland, 2001), e cunhado em 2008 por DJ Patil e Jeff Hammerbacher, que passaram a utilizar o termo cientista de dados como profissão no *LinkedIn* e no *Facebook* na procura de profissionais cujo trabalho principal era o de lidar com um grande volume de dados (Davenport & Patil, 2012). Contudo, foi somente no dia 20 de março de 2023 que a ocupação cientista de dados foi incluída na Classificação Brasileira de Ocupações (CBO), fazendo parte do código 2112, no qual estão inseridos os profissionais de estatística e afins (Classificação Brasileira de Ocupações, 2023).

Tendo em vista a urgência do mercado e, por conseguinte, a grande demanda de profissionais devidamente qualificados, pode-se visualizar uma rápida expansão de cursos de Ciência de Dados em universidades no Brasil. Considerando o Cadastro Nacional de Cursos e Instituições de Educação Superior (Cadastro e-MEC: https://emec.mec.gov.br/) , em novembro de 2020 registravam-se ao menos 36 cursos de graduação em Ciência de Dados no país, ao passo que

em janeiro de 2023 já constavam mais de 70 cursos, incluindo os que ainda não haviam sido iniciados.

Não obstante, há uma série de competências necessárias para compor o perfil do cientista de dados e diretrizes esperadas para a grade curricular dos cursos, para garantir ao profissional uma formação minimamente adequada para atuar no mercado de trabalho. De acordo com De Veaux et al. (2017), essas competências principais são: pensamento estatístico e computacional; fundamentos matemáticos; saber construir e avaliar modelos; algoritmos e base do software; curadoria de dados e saber comunicar o conhecimento com responsabilidade.

Devido à recente abertura desses cursos de graduação em universidades no Brasil, ainda há um grande déficit de estudos descritivos na literatura. Desse modo, é de extrema importância a caracterização de tais cursos no sentido de apoiar possíveis decisões quanto a melhorias de ofertas e servir como base para estudos posteriores. Para isso, optou-se por fazer o uso da estatística descritiva univariada, sumarizando as variáveis individualmente e assim possibilitando a obtenção de uma melhor compreensão no que tange às características dos cursos de Ciência de Dados no país. Além disso, optou-se pelo uso da estatística descritiva multivariada via análise de correspondência, que induz a exploração da relação entre as variáveis, evidenciando o estudo de associações que não são visíveis na análise univariada, mas que são importantes e que levam a uma melhor compreensão do cenário de tais cursos nas universidades brasileiras.

Neste sentido, buscou-se responder os seguintes questionamentos: os cursos têm sido ofertados em grande maioria pelas universidades públicas ou privadas? Quando o curso passou a ser ofertado? Costumam ser de modalidade de ensino EAD ou presencial? São do tipo tecnológico ou bacharelado? Quantas e quais os grupos de disciplinas mais compõem a grade curricular? Em quais regiões do país tais cursos se concentram em maioria? Como é a oferta de vagas e qual é o perfil de ingressos em cursos do tipo bacharelado e tecnológico? Como o número de disciplinas e a carga horária estão relacionadas com os cursos presenciais e EAD?

Este artigo está organizado da forma como segue. Na Seção 2 encontra-se a metodologia, com os detalhes de como os dados foram obtidos, quais os cursos incluídos e a descrição das variáveis estudadas. Na Seção 3 encontram-se os resultados, com as frequências acumuladas da criação e do início dos cursos ao longo dos anos, a sumarização das variáveis individuais para os cursos EAD, presenciais e ambos conjuntamente, mapas com o percentual de cursos presenciais em cada região do país, visualizações da análise de correspondência

e boxplots acerca das disciplinas dos cursos. Por fim, a Seção 4 exibe os comentários finais, com o resumo dos resultados apresentados.

2.METODOLOGIA

2.1 COLETA DE DADOS

Para este estudo foi realizada uma pesquisa *on-line* para verificar quais são as universidades brasileiras que ofertam os cursos de graduação em Ciência de Dados. Essa consulta foi feita pelo site que fornece a base de dados oficial dos cursos e Instituições de Educação Superior - IES (https://emec.mec.gov.br/), no dia 27 de outubro de 2022, utilizando-se os termos "Ciência de Dados" e "*Data Science*" nos buscadores, totalizando em 93 observações, das quais foram retirados os cursos que ainda não haviam sido iniciados, restando 49 cursos. Fazem parte das variáveis disponíveis no e-MEC informações como as datas de início e de criação do curso, a modalidade, o tipo de graduação e se a instituição é pública ou privada.

A partir da listagem de instituições, a *home page* de cada curso foi visitada, com o intuito de se obter a duração do curso, a grade curricular e, no caso de o curso ser ofertado na modalidade de ensino presencial, a localização geográfica. A listagem de cursos presenciais e EAD com seus respectivos *sites* está disponível no Apêndice 1 (Tabela 1.1 e Tabela 1.2, respectivamente).

Tabela 1 - Frequência Absoluta dos Cursos de Ciência de Dados no Brasil

Nome do Curso	Frequência
Ciência de Dados	34
Ciência de Dados e Inteligência Artificial	11
Data Science	4
Ciência de Dados e Inteligência Analítica	2
Ciência de Dados e Machine Learning	2
Ciência de Dados para Negócios	2
Ciências de Dados e Análise de Comportamento	1
Estatística e Ciência de Dados	1
Marketing Digital e Data Science	1

Fonte - Autores (2023)

Além disso, foi obtida uma base de dados do Censo da Educação Superior referente ao ano de 2021, sendo esta a base mais recente disponível. Embora os dados não fossem apenas de cursos de Ciência de Dados, foram filtrados os que continham tal termo no nome (como Ciência De Dados e Inteligência Artificial, Data Science, Estatística e Ciência De Dados, entre outros), resultando em 2031 observações e 200 variáveis. Contudo, devido ao fato de inúmeras universidades ofertarem cursos na modalidade EAD em vários polos diferentes, há muitas repetições dos cursos. No entanto, se destaca o fato de que nesses casos, os cursos ofertados eram do tipo tecnológico, com os mesmos códigos de IES e código de curso, mas apenas uma das observações possuía o valor da variável quantidade de vagas sendo diferente de zero. Tais repetições foram agrupadas pelas principais variáveis qualitativas (como grau acadêmico, rede da IES, modalidade de ensino e código do curso), e, em seguida, as variáveis quantitativas foram somadas, eliminando as repetições da base e restando apenas 44 cursos. Além disso, também foram excluídas as variáveis que não faziam parte do escopo do trabalho ou que estavam extremamente incompletas a ponto de ser inviável utilizá-las.

Com essas duas bases de dados disponíveis, foi realizada a junção através do código do curso. Uma vez que a junção dessas duas bases resultou em algumas variáveis repetidas, a nova base de dados foi exportada para efetuar a redução dessas variáveis. Isso porque no Censo da Educação Superior de 2021 estavam presentes alguns cursos que haviam sido excluídos da base proveniente do e-MEC, por constarem como não iniciados, mas que foram mantidos, exclusivamente nessa situação de terem feito parte do Censo. Assim, resultou-se em uma base de dados com 58 cursos, conforme a Tabela 1. As variáveis que foram utilizadas estão disponíveis na Tabela 2.

Tabela 2 - Nome, Descrição e Codificação das Variáveis que Foram Estudadas

Descrição	Variável	Codificação
Tipo de modalidade de ensino do curso	MOD_ENSINO	{PRES.; EAD}
Tipo de grau acadêmico conferido ao aluno após o término do curso	GRAU_ACAD	{TECN.;BACH.}
Rede de ensino	REDE	{PÚBL.; PRIV.}
Quantidade total de vagas oferecidas	VG	{0; 1; 2; ...}
Quantidade total de vagas oferecidas no turno diurno	VG_DIURNO	{0; 1; 2; ...}
Quantidade total de vagas oferecidas no turno noturno	VG_NOTURNO	{0; 1; 2; ...}
Quantidade total de vagas oferecidas à distância	VG_EAD	{0; 1; 2; ...}
Quantidade de ingressantes	ING	{0; 1; 2; ...}
Quantidade de ingressantes do sexo masculino	ING_MASC	{0; 1; 2; ...}
Quantidade de ingressantes do sexo feminino	ING_FEM	{0; 1; 2; ...}
Quantidade de ingressantes no turno diurno em cursos presenciais	ING_DIU_PRES	{0; 1; 2; ...}
Quantidade de ingressantes no turno noturno em cursos presenciais	ING_NOT_PRES	{0; 1; 2; ...}
Quantidade de ingressantes que concluíram o Ensino Médio em escolas públicas	ING_PROCPUB	{0; 1; 2; ...}
Quantidade de ingressantes que concluíram o Ensino Médio em escolas privadas	ING_PROCPRI	{0; 1; 2; ...}
Quantidade de concluintes	CONC	{0; 1; 2; ...}
Quantidade de concluintes do sexo feminino	CONC_FEM	{0; 1; 2; ...}
Quantidade de concluintes do sexo masculino	CONC_MASC	{0; 1; 2; ...}
Quantidade de concluintes que terminaram o Ensino Médio em escolas públicas	CONC_PROCPUB	{0; 1; 2; ...}
Quantidade de concluintes que terminaram o Ensino Médio em escolas privadas	CONC_PROCPRI	{0; 1; 2; ...}
Quantidade de ingressantes com menos de 25 anos	ING_0_24	{0; 1; 2; ...}
Quantidade de ingressantes com 25 anos ou mais	ING_25	{0; 1; 2; ...}
Quantidade de concluintes com menos de 25 anos	CONC_0_24	{0; 1; 2; ...}
Quantidade de concluintes com 25 anos ou mais	CONC_25	{0; 1; 2; ...}
Duração anual do curso	DURACAO(ANO)	{1,5; 2; 2,5; ...; 4; S. Inf.}
Carga horária do curso	CARGA_HOR	{0; 1; 2; …}
Número de disciplinas	NUM_DISC	{0; 1; 2; ...}
Ano do início de funcionamento do curso	INICIO_FUNC	{2018; ... ; 2022; S. Inf.}
Ano de criação do curso	ATO_CRIACAO	{2017; ... ; 2022 ; S. Inf.}
Região em que o curso está localizado, caso presencial	REGIÃO	{S, SD, CO, N, ND}
Se o nome do curso é apenas Ciência de Dados/ *Data Science*	NOME_CD	{CD_SIM, CD_NÃO}

Fonte - Autores (2023)

As variáveis ING_DIU_PRES e ING_NOT_PRES são complementares e estão preenchidas por valores maiores do que zero apenas nos cursos cuja modalidade de ensino é presencial, e, além disso, não há no Censo de Educação Superior de 2021 uma variável referente à quantidade de ingressantes em cursos de modalidade EAD, motivo pelo qual tal variável não consta neste trabalho. Na variável INICIO_FUNC, há a situação específica do curso de Estatística e Ciência de Dados, ofertado pela USP - São Carlos, que, embora tenha sido iniciado em 2009 com o nome de Estatística, obteve o nome atual somente a partir de 2020, sendo este o ano considerado de seu início. Ainda, é importante salientar que, embora todos os cursos que constam no Censo também constarem no site do e-MEC, nem todos os cursos presentes no e-MEC fizeram parte do Censo, pois neste trabalho foram ainda considerados os cursos iniciados em 2022 disponibilizados pelo e-MEC a fim de que o conjunto de dados fosse mais representativo. Além disso, havia alguns cursos que, de acordo com o e-MEC, foram iniciados em 2021, mas que não constavam no Censo. Possivelmente isso se deve a razão de que os dados disponibilizados na base de dados do Censo são obtidos por meio de um questionário eletrônico, em que as próprias IES são responsáveis por fazer o preenchimento acerca dos cursos. E, apesar de o INEP analisar a base de dados a fim de verificar a consistência das informações (PROPLAN, 2023), ainda existem inconsistências na mesma.

2.2 ANÁLISE DE CORRESPONDÊNCIA

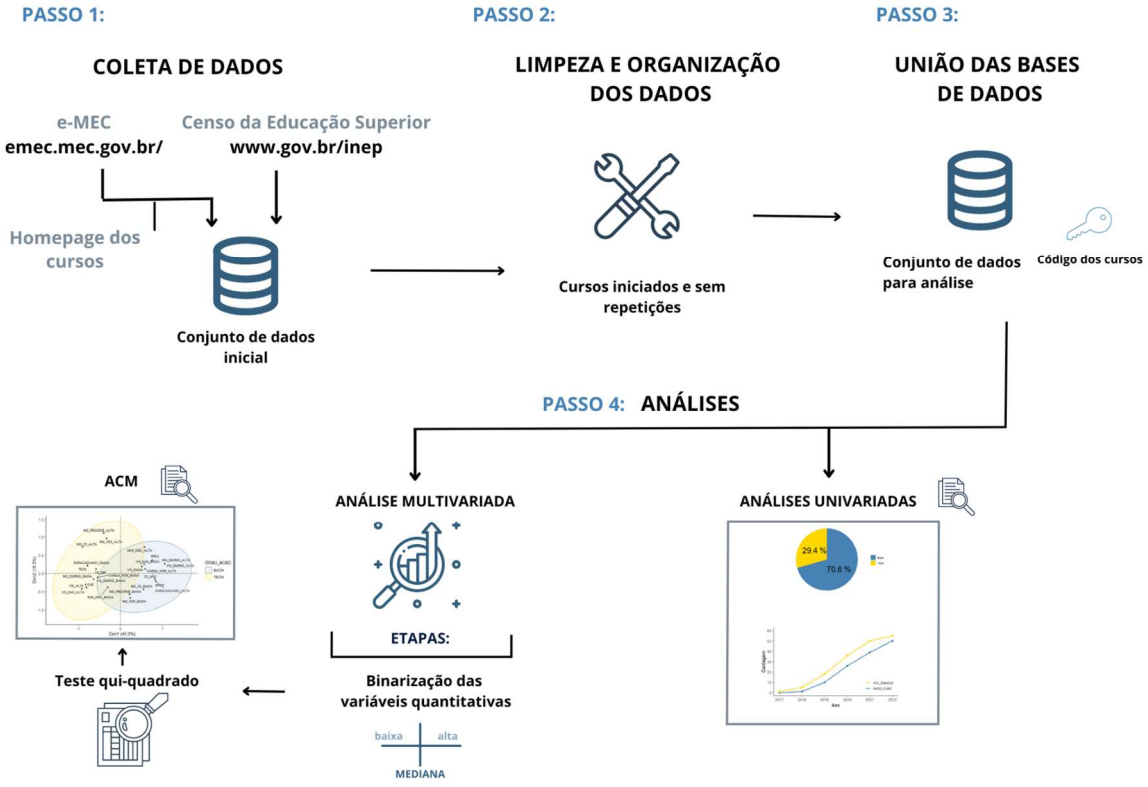
A análise de correspondência é uma técnica multivariada elaborada por um grupo de estatísticos franceses em 1960. Seu uso e metodologias são análogas a métodos como a análise de componentes principais e a análise fatorial, diferindo, sobretudo, no foco da análise de correspondência, que se direciona ao estudo de variáveis categóricas (Carvalho & Struchiner, 1992), sendo que a análise de correspondência múltipla (ACM) permite, para um conjunto de mais de duas variáveis, avaliar relações entre as variáveis e suas categorias de forma gráfica (Prado, 2012). Hair et al. (2009) pontuaram que se trata de uma técnica de interdependência que tem se popularizado no que tange à redução dimensional e ao mapeamento perceptual. Além disso, representa associações em tabelas de frequências (Johnson & Wichern, 2002) e, segundo Nascimento et al. (2017), a AC exige apenas que todos os dados sejam positivos e que estejam em uma tabela retangular, podendo a ACM ser realizada por intermédio dos métodos de matriz indicadora Z e da matriz de Burt, sendo estes comuns na literatura especializada. De uma forma geral, na matriz indicadora, nas linhas estão os indivíduos e nas colunas as categorias de cada variável (Costa, 2016).

Portanto, cabe ressaltar que a exploração e visualização de dados neste artigo também foi realizada de forma multivariada, considerando a técnica de redução de dimensionalidade ACM, tendo em vista que a base de dados que foi utilizada possuía variáveis categóricas e que mesmo as variáveis quantitativas puderam ser categorizadas de forma dicotômica. Essa abordagem foi útil para uma melhor exploração e visualização de dados, bem como para uma melhor interpretabilidade.

2.2 FLUXOGRAMA DA METODOLOGIA DE COLETA E ANÁLISE DOS DADOS

Na Figura 1 é apresentado um fluxograma que contém de forma resumida todos os passos fundamentais da metodologia deste trabalho, desde a coleta dos dados até as análises desenvolvidas.

Figura 1 - Fluxograma da metodologia de coleta e análise dos dados para os cursos



Fonte – Autores (2023)

3.RESULTADOS E DISCUSSÃO

3.1 DESCRIÇÃO GERAL DOS CURSOS

Ao todo, foram contabilizados 58 cursos na área de Ciência de Dados no Brasil, e, conforme a Tabela 3, pode-se verificar que 82,8% dos cursos são ofertados por IES de rede privada e que 60,3% são do tipo tecnológico. No que concerne à quantidade total de vagas disponíveis, tem-se que a mesma é em média 12 vezes maior que a de ingressos, aproximadamente, além de ter um valor máximo

bem destoante. Com relação ao sexo, a quantidade média de ingressos masculinos é quase 3 vezes maior, em números absolutos, que a de ingressantes femininos. Também, em números absolutos, além do fato de a média de ingressantes com 25 anos ou mais ser 3 vezes a de ingressantes com menos de 25 anos, tem-se que em média a quantidade de ingressantes que concluíram o Ensino Médio em escolas públicas é cerca de 2 vezes a de ingressantes que concluíram o Ensino Médio em escolas privadas.

Tabela 3 - Análise Descritiva Univariada dos Cursos Presenciais e EAD de Ciência de Dados no Brasil

Variável	Brasil (n=58)
REDE (% PUBL / % PRIV)	17,2 / 82,8
GRAU_ACAD (% BACH / % TEC)	39,7 / 60,3
VG (M; DP[MIN,MAX])	2599,0; 5533,5 [20, 17695]
ING (M; DP[MIN,MAX])	215,0; 548,5 [1, 3442]
ING_MASC (M; DP[MIN,MAX])	158,4; 399,1 [0, 2496]
ING_FEM (M; DP[MIN,MAX])	56,6; 149,8 [0, 946]
ING_DIU_PRES (M; DP[MIN,MAX])	9,0; 18,5 [0, 71]
ING_NOT_PRES (M; DP[MIN,MAX])	8,2; 22,2 [0, 88]
ING_PROCPUB (M; DP[MIN,MAX])	151,2; 413,0 [0, 2658]
ING_PROCPRI (M; DP[MIN,MAX])	63,8; 148,1 [0, 784]
ING_0_24 (M; DP[MIN,MAX])	49,3; 98,3 [0, 607]
ING_25 (M; DP[MIN,MAX])	165,7; 451,6 [1, 2835]
CONC (M; DP[MIN,MAX])	5,5; 17,5 [0, 96]
CONC_MASC (M; DP[MIN,MAX])	4,3; 14,0 [0, 76]
CONC_FEM (M; DP[MIN,MAX])	1,2; 3,6 [0, 20]
CONC_PROCPUB (M; DP[MIN,MAX])	3,8; 12,5 [0, 65]
CONC_PROCPRI (M; DP[MIN,MAX])	1,8; 5,7 [0, 31]
CONC_0_24 (M; DP[MIN,MAX])	0,5; 1,5 [0, 7]
CONC_25 (M; DP[MIN,MAX])	5,0; 16,4 [0, 90]
NUM_DISC (M; DP[MIN,MAX])	39,2; 13,1 [22, 90]
CARGA_HOR (M; DP[MIN,MAX])	2616,2; 615,5 [1790, 3840]
DURACAO(ANO) 1,5 2,0 2,5 3,0 3,5 4,0 S. Inf.	% 8,6 18,9 19,0 12,1 5,2 32,8 3,4

M = média; DP = desvio padrão

Fonte - Autores (2023)

Acerca das variáveis relacionadas à quantidade total de concluintes, nota-se que os valores máximos são bem menores em comparação com as variáveis relacionadas à quantidade total de ingressos. Uma possível explicação é a situação de que muitos cursos foram tanto criados quanto iniciados a partir de 2019 (Tabela 4 e Figura 2), e, com base nos dados do Censo da Educação Superior de 2021, é provável de que em alguns cursos de Ciência de Dados nenhuma turma ainda havia se formado, especialmente as de bacharelado. Apesar disso, é possível verificar (ainda na Tabela 3), em números absolutos, que a média da quantidade total de concluintes com 25 anos ou mais é 10 vezes maior que a de concluintes com menos de 25 anos, a média de concluintes do sexo masculino é um pouco mais do que 3 vezes maior do que a de sexo feminino e a média da quantidade de concluintes que cursaram o Ensino Médio em escolas públicas é cerca de 2 vezes a de escolas privadas.

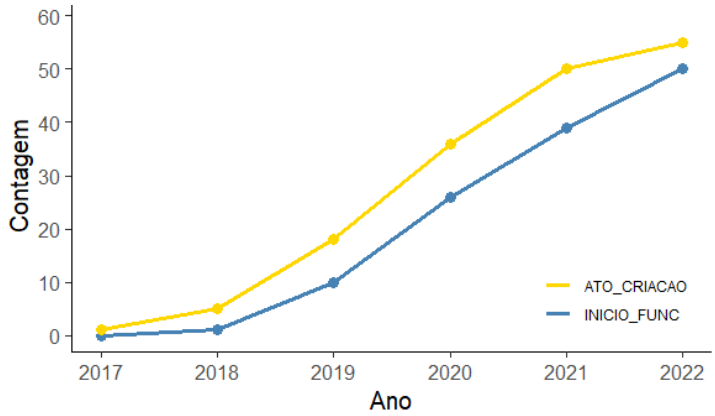
No que tange aos números médios de disciplinas e de carga horária, têm-se respectivamente os valores de 39,2 disciplinas e de 2616,2 horas. Apenas 32,8% dos cursos têm uma duração de 4 anos, o que remete ao fato de somente 39,7% dos cursos serem do tipo bacharelado.

Tabela 4 - Frequência Absoluta dos Anos de Criação e de Início dos Cursos

ATO_CRIACAO	CONTAGEM	INICIO_FUNC	CONTAGEM
2017	1	2018	1
2018	4	2019	9
2019	13	2020	16
2020	18	2021	13
2021	14	2022	11
2022	5	S. Inf.	8
S. Inf.	3		

Fonte - Autores (2023)

Figura 2 - Frequências Acumuladas dos Cursos Criados e Iniciados



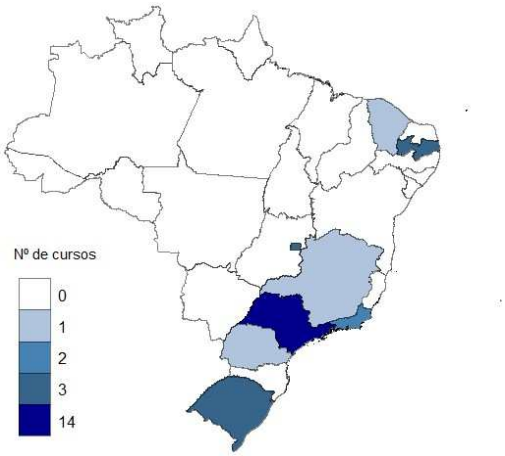
Fonte – Autores (2023)

3.2 DESCRIÇÃO DOS CURSOS PRESENCIAIS

Conforme a Tabela 5, no Brasil existem 28 cursos na área de Ciência de Dados de modalidade de ensino presencial, sendo que, considerando todo o país, 67,9% deles são do tipo bacharelado (ou seja, o maior percentual de cursos cujo grau acadêmico é bacharelado no Brasil encontra-se na modalidade presencial) e 67,9% são ofertados em IES de rede privada. Não obstante, a média da quantidade de vagas ofertadas no período diurno é ligeiramente maior do que a no período noturno, ao passo que a quantidade total de ingressos é em média cerca de metade da quantidade total de vagas disponíveis.

Nota-se que os valores máximos amostrais nas variáveis referentes aos concluintes são menores para os cursos presenciais do que para aqueles cuja modalidade de ensino é EAD (Tabela 8), o que pode ser explicado pela duração dos cursos, uma vez que há uma concentração de cursos com mais de 3 anos de duração (somando 64,2%), e, por conseguinte, pelo grau acadêmico associado a durações maiores, que é o bacharelado. Nota-se que os valores médios do número de disciplinas e de carga horária são de 44,0 disciplinas e de 2846,3 horas respectivamente. Ao contrário do que acontece na descrição geral dos cursos no Brasil (independentemente da modalidade de ensino), em números absolutos, neste caso a média da quantidade de ingressos com 25 anos ou mais é quase 2 vezes menor que a de ingressos com menos de 25 anos. Já em relação ao período, a média de ingressos no diurno é ligeiramente maior que a no noturno, bem como a média da quantidade de ingressos que cursaram o Ensino Médio em escolas públicas é cerca de 1,6 vezes maior do que em escolas privadas. Acerca dos estados, pode-se visualizar na Figura 3 que os cursos se concentram principalmente em SP, RS e PB.

Figura 3 - Mapa do Brasil com o Número de Cursos Presenciais por Estado



Fonte – Autores (2023)

Tabela 5 - Análise Descritiva Univariada dos Cursos Presenciais no Brasil

Variável	Brasil (n=28)
REDE (% PUBL / % PRIV)	32,1 / 67,9
GRAU_ACAD (% BACH / % TEC)	67,9 / 32,1
VG_DIURNO (M; DP[MIN,MAX])	52,6; 55,6 [0, 168]
VG_NOTURNO (M; DP[MIN,MAX])	46,2; 64,5 [0, 213]
ING_MASC (M; DP[MIN,MAX])	30,9; 18,9 [5, 62]
ING_FEM (M; DP[MIN,MAX])	11,2; 8,1 [1, 26]
ING_DIU_PRES (M; DP[MIN,MAX])	22,1; 23,7 [0, 71]
ING_NOT_PRES (M; DP[MIN,MAX])	20,0; 31,5 [0, 88]
ING_PROCPUB (M; DP[MIN,MAX])	26,0; 23,0 [0, 79]
ING_PROCPRI (M; DP[MIN,MAX])	16,1; 11,8 [0, 41]
ING_0_24 (M; DP[MIN,MAX])	26,4; 16,1 [4, 59]
ING_25 (M; DP[MIN,MAX])	15,7; 16,2 [1, 59]
VG (M; DP[MIN,MAX])	98,8; 79,2 [20, 282]
ING (M; DP[MIN,MAX])	42,1; 25,7 [7, 88]
CONC (M; DP[MIN,MAX])	1,1; 4,2 [0, 18]
CONC_MASC (M; DP[MIN,MAX])	0,7; 2,8 [0, 12]
CONC_FEM (M; DP[MIN,MAX])	0,4; 1,4 [0, 6]
CONC_PROCPUB (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]
CONC_PROCPRI (M; DP[MIN,MAX])	1,1; 4,2 [0, 18]
CONC_0_24 (M; DP[MIN,MAX])	0,4; 1,7 [0, 7]
CONC_25 (M; DP[MIN,MAX])	0,6; 2,6 [0, 11]
NUM_DISC (M; DP[MIN,MAX])	44,0; 14,9 [26, 90]
CARGA_HOR (M; DP[MIN,MAX])	2846,3; 501,1 [1960, 3840]
DURAÇAO(ANO)	%
1,5	0,0
2,0	3,6
2,5	3,6
3,0	25,0
3,5	10,7
4,0	53,5
S. Inf.	3,6

M = média; DP = desvio padrão

Fonte - Autores (2023)

No que concerne às regiões do país, têm-se na Tabela 6 as datas de criação e de início do primeiro curso de Ciência de Dados em cada uma delas. É possível verificar que a região Centro-Oeste é a que possui o curso mais antigo do qual se tem registro no e-MEC, enquanto a região Sul foi a última a abrir o curso de Ciência de Dados pela primeira vez.

Tabela 6 - Menores Anos de Criação e de Início dos Cursos Presenciais por Região

Região	Criação em	Região	Início em
Centro-Oeste	2017	Centro-Oeste	2018
Nordeste	2019	Nordeste	2020
Sudeste	2019	Sudeste	2020
Sul	2020	Sul	2021

Fonte - Autores (2023)

Não obstante, os cursos presenciais estão mais presentes nas regiões Sudeste, Sul e Nordeste, com 60,7% e 14,3% para as duas últimas, respectivamente (Figura 4 e Figura 5). Contudo, nota-se a grande diferença entre o Sudeste e as demais regiões, sendo que no Norte não há nenhum curso presencial que tenha participado do Censo de Educação Superior de 2021 ou que conste no e-MEC até o momento em que este trabalho foi escrito.

Na região Sul, todos os cursos dos quais se tem informação são ofertados por IES de rede privada, sendo que metade deles são do tipo bacharelado (Tabela 7), embora 100% dos cursos tenham mais do que 2,5 anos de duração. Devido ao fato de apenas 1 dos 4 cursos terem feito parte do Censo, o desvio padrão de grande parte das variáveis é zero.

Já na região Sudeste, 35,3% dos cursos são ofertados por IES da rede pública, o que representa o segundo maior percentual dentre as regiões. Apenas 29,4% são do tipo tecnológico, e, assim, tem-se o maior número médio de disciplinas do país (47,0 disciplinas) e a maior carga horária média (3059,6 horas), fazendo com que 64,7% dos cursos tenham mais de 3 anos de duração.

O Centro-Oeste tem 100% dos cursos sendo oferecidos pela rede privada, todos do tipo bacharelado. Isso faz com que 100% dos cursos tenham mais de 3 anos de duração, e que a carga horária média seja a segunda maior do país (2983,3 horas). Apenas 1 dos 3 cursos presenciais existentes fez parte do Censo de 2021, o que faz com que o desvio padrão para as variáveis provenientes do mesmo seja zero. Porém, percebe-se que é a única região do país em que a média de ingressos que concluíram o Ensino Médio em escolas privadas é maior do que a média dos ingressos provenientes de escolas públicas.

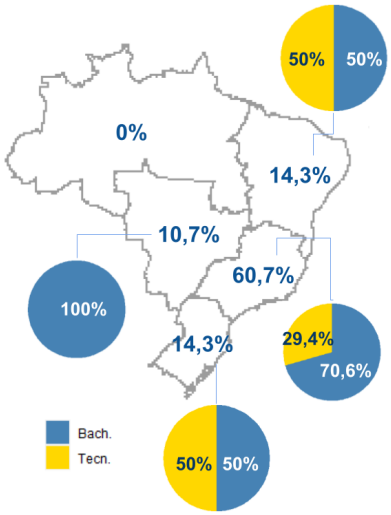
Tabela 7 - Análise Descritiva Univariada dos Cursos de Ciência de Dados Presenciais por Região Geográfica

Variável	Sul (n=4)	Sudeste (n=17)	Centro-Oeste (n=3)	Nordeste (n=4)
REDE (% PUBL / % PRIV)	0,0 / 100,0	35,3 / 64,7	0,0 / 100,0	75,0 / 25,0
GRAU_ACAD (% BACH / % TEC)	50,0 / 50,0	70,6 / 29,4	100,0 / 0,0	50,0 / 50,0
VG_DIURNO (M; DP[MIN,MAX])	168,0; 0,0 [168, 168]	44,6; 51,8 [0, 140]	124,0; 0,0 [124, 124]	30,0; 24,5 [0, 60]
VG_NOTURNO (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	41,1; 50,2 [0,142]	126,0; 0,0 [126, 126]	53,2;106,5 [0, 213]
ING_MASC (M; DP[MIN,MAX])	55,0; 0,0 [55, 55]	30,8; 20,4 [5, 62]	29,0; 0,0 [29, 29]	25,5; 15,9 [9, 46]
ING_FEM (M; DP[MIN,MAX])	16,0; 0,0 [16, 16]	11,7; 9,4 [1, 26]	10,0; 0,0 [10, 10]	8,8; 5,4 [1, 13]
ING_DIU_PRES (M; DP[MIN,MAX])	71,0; 0,0 [71, 71]	16,0; 20,9 [0, 62]	27,0; 0,0 [27, 27]	26,8; 24,4 [0, 59]
ING_NOT_PRES (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	26,5; 36,4 [0, 88]	12,0; 0,0 [12, 12]	7,5; 15,0 [0, 30]
ING_PROCPUB (M; DP[MIN,MAX])	30,0; 0,0 [30, 30]	28,8; 27,6 [0, 79]	11,0; 0,0 [11, 11]	20,2; 7,3 [15, 31]
ING_PROCPRI (M; DP[MIN,MAX])	41,0; 0,0 [41, 41]	13,7; 10,4 [0, 34]	28,0; 0,0 [28, 28]	14,0; 10,4 [3, 28]
ING_0_24 (M; DP[MIN,MAX])	51,0; 0,0 [51, 51]	26,2; 16,8 [4, 59]	23,0; 0,0 [23, 23]	21,8; 14,5 [11, 43]
ING_25 (M; DP[MIN,MAX])	20,0; 0,0 [20, 20]	16,3; 19,6 [1, 59]	16,0; 0,0 [16, 16]	12,5; 7,0 [3, 19]
VG (M; DP[MIN,MAX])	168,0; 0,0 [168, 168]	85,7; 69,4 [20, 282]	250,0; 0,0 [250, 250]	83,2;87,6 [30, 213]
ING (M; DP[MIN,MAX])	71,0; 0,0 [71, 71]	42,5; 29,0 [7, 88]	39,0; 0,0 [39, 39]	34,2; 16,9 [21, 59]
CONC (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	1,5; 5,2 [0, 18]	1,0; 0,0 [1, 1]	0,0; 0,0 [0, 0]
CONC_MASC (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	1,0; 3,5 [0, 12]	0,0; 0,0 [0, 0]	0,0; 0,0 [0, 0]
CONC_FEM (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	0,5; 1,7 [0, 6]	1,0; 0,0 [1, 1]	0,0; 0,0 [0, 0]
CONC_PROCPUB (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	0,0; 0,0 [0, 0]	0,0; 0,0 [0, 0]	0,0; 0,0 [0, 0]
CONC_PROCPRI (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	1,5; 6,0 [0, 18]	1,0; 0,0 [1, 1]	0,0; 0,0 [0, 0]
CONC_0_24 (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	0,6; 2,0 [0, 7]	1,0; 0,0 [1, 1]	0,0; 0,0 [0, 0]
CONC_25 (M; DP[MIN,MAX])	0,0; 0,0 [0, 0]	0,9; 3,2 [0, 11]	0,0; 0,0 [0, 0]	0,0; 0,0 [0, 0]
NUM_DISC (M; DP[MIN,MAX])	39,5; 6,8 [34, 48]	47,0; 17,7 [30, 90]	43,0; 10,4 [37, 55]	39,0; 13,5 [26, 57]
CARGA_HOR (M; DP[MIN,MAX])	2581,0; 574,9 [2100, 3244]	3059,6; 429,3 [2412, 3840]	2983,3; 360,8 [2775, 3400]	2369,0; 387,5 [1960, 2850]
DURACAO(ANO)	%	%	%	%
1,5	0,0	0,0	0,0	0,0
2,0	0,0	0,0	0,0	25,0
2,5	0,0	5,9	0,0	0,0
3,0	50,0	23,5	0,0	25,0
3,5	0,0	5,9	66,7	0,0
4,0	50,0	58,8	33,3	50,0
S. Inf.	0,0	5,9	0,0	0,0

M = média; DP = desvio padrão

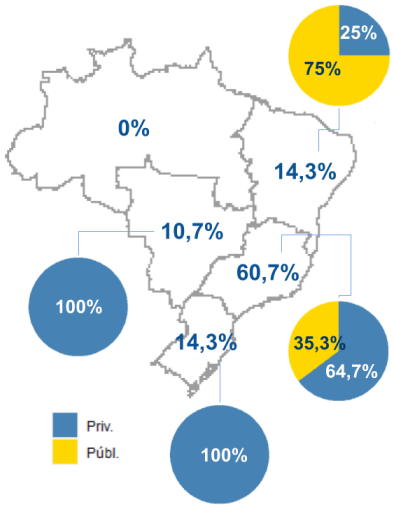
Fonte - Autores (2023)

Figura 4 – Mapa do Brasil com a Distribuição de Cursos Presenciais por Região e Grau Acadêmico



Fonte – Autores (2023)

Figura 5 – Mapa do Brasil com a Distribuição de Cursos Presenciais por Região e Rede de Ensino da IES



Fonte – Autores (2023)

Tabela 8 – Análise Descritiva dos Cursos de Ciência de Dados de Modalidade de Ensino EAD no Brasil

Variável	EAD (n=30)
REDE (% PUBL / % PRIV)	3,3 / 96,7
GRAU_ACAD (% BACH / % TEC)	13,3 / 86,7
VG_EAD (M; DP[MIN,MAX])	4329,8; 6711,5 [44, 17695]
ING_MASC (M; DP[MIN,MAX])	246,7; 503,9 [0, 2496]
ING_FEM (M; DP[MIN,MAX])	88,1; 189,8 [0, 946]
ING_PROCPUB (M; DP[MIN,MAX])	238,0; 523,4 [0, 2658]
ING_PROCPRI (M; DP[MIN,MAX])	96,8; 186,7 [0, 784]
ING_0_24 (M; DP[MIN,MAX])	65,2; 125,7 [0, 607]
ING_25 (M; DP[MIN,MAX])	269,6; 568,4 [1, 2835]
VG (M; DP[MIN,MAX])	4329,8; 6711,5 [44, 17695]
ING (M; DP[MIN,MAX])	334,8; 693,2 [1, 3442]
CONC (M; DP[MIN,MAX])	8,7; 22,2 [0, 96]
CONC_MASC (M; DP[MIN,MAX])	6,9; 17,8 [0, 76]
CONC_FEM (M; DP[MIN,MAX])	1,8; 4,4 [0, 20]
CONC_PROCPUB (M; DP[MIN,MAX])	6,4; 15,8 [0, 65]
CONC_PROCPRI (M; DP[MIN,MAX])	2,2; 6,6 [0, 31]
CONC_0_24 (M; DP[MIN,MAX])	0,6; 1,2 [0, 6]
CONC_25 (M; DP[MIN,MAX])	8,1; 20,9 [0, 90]
NUM_DISC (M; DP[MIN,MAX])	34,5; 9,2 [22, 51]
CARGA_HOR (M; DP[MIN,MAX])	2304,8; 631,7 [1790, 3840]
DURACAO(ANO)	%
1,5	16,8
2,0	33,3
2,5	33,3
3,0	0,0
3,5	0,0
4,0	13,3
S. Inf.	3,3
M = média; DP = desvio padrão	

Fonte – Autores (2023)

Por fim, nota-se na Figura 4 que 50% dos cursos situados no Nordeste são do tipo bacharelado, e, na Figura 5, que 75% são ofertados por IES de rede pública (o segundo maior percentual do país). Apesar disso, na Tabela 7 é possível visualizar que é a região com a segunda menor carga horária média (2369,0 horas) e com o menor número médio amostral de disciplinas, embora 50% dos cursos tenham mais do que 2,5 anos de duração. Destaca-se por ter cerca de 1,8 vezes mais vagas disponibilizadas para o período noturno do que para o período diurno, em média, porém, a quantidade média de ingressantes no período diurno é cerca de 3,6 vezes maior do que a de ingressantes no período noturno.

3.3 DESCRIÇÃO DOS CURSOS DE MODALIDADE EAD

Acerca da descrição dos cursos de modalidade de ensino EAD nota-se, na Tabela 8, que majoritariamente são ofertados por IES de rede privada (96,7%), bem como são do tipo tecnológico (96,7%). No entanto, a carga horária média amostral, que é igual a 2304,8 horas, é menor do que a dos cursos presenciais no Brasil (2846,3 horas), e há, em média, 34,5 disciplinas, sendo que 83,4% dos cursos têm menos do que 3 anos de duração. Ainda, há a ocorrência de cursos do tipo tecnológico com apenas 1,5 anos de duração, sendo eles ofertados por exemplo pelas IES Centro Universitário Maurício De Nassau (UNINASSAU) e Universidade Da Amazônia (UNAMA), geralmente identificados como "*Data Science*". Cabe também ressaltar a ocorrência de que, mesmo tendo sido filtrado os cursos por seus diferentes códigos, e, apesar do ano de início dos cursos serem diferentes, alguns cursos possuem as mesmas informações (como duração do curso e número de disciplinas) devido ao fato de serem ofertadas pelo mesmo grupo, como é o caso do Cruzeiro do Sul Educacional, do qual fazem parte, por exemplo, as universidades subsidiárias Universidade De Franca (UNIFRAN), Universidade Cruzeiro Do Sul (UNICSUL) e Universidade Cidade De São Paulo (UNICID) (Giordan, 2017).

Em relação às demais variáveis, enquanto nos cursos presenciais do país a quantidade média amostral de vagas era 2,3 vezes maior do que a de ingressos, no caso dos cursos cuja modalidade é EAD têm-se que a quantidade média de vagas é quase 13 vezes maior que a de ingressos. Possivelmente isso ocorra devido ao fato de o valor máximo da quantidade de vagas ser igual a 17695. Em números absolutos, a média de ingressos de sexo masculino é quase 3 vezes maior do que a de sexo feminino, bem como a média de concluintes de sexo masculino é quase 4 vezes maior do que a de sexo feminino. Tem-se ainda que a média de ingressos com 25 anos ou mais é cerca de 4 vezes maior do que a de menos de 25 anos, enquanto a média de concluintes com 25 anos ou mais é 13,5 vezes maior

do que a com menos de 25 anos. Não obstante, a quantidade média de ingressantes que concluíram o Ensino Médio em escolas públicas é mais de 2 vezes maior do que a de escolas privadas. Em relação aos concluintes, a média dos que concluíram o Ensino Médio em escolas públicas é 2,9 vezes maior do que em escolas privadas, em números absolutos.

3.4 ANÁLISE MULTIVARIADA

Para que a técnica de ACM pudesse ser empregada, primeiramente foi realizada uma categorização das variáveis quantitativas, conforme o Apêndice 2, Tabela 2.1, de modo que a categoria "ALTA" é aquela em que estão inseridos os valores maiores que a mediana da variável e na "BAIXA" os valores menores ou iguais a ela. Nota-se que nos casos em que as variáveis provenientes do Censo estavam majoritariamente preenchidas por 0, a mediana também foi zero, sendo "ALTA" qualquer dado maior que tal valor.

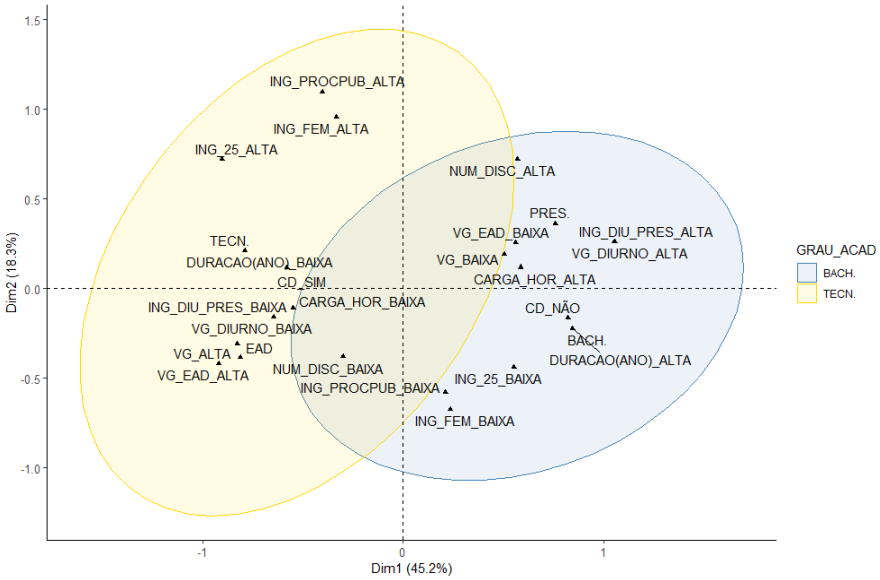
Optou-se por excluir os dados ausentes principalmente em cursos que foram iniciados em 2022 ou que foram iniciados em 2021, mas que não participaram do Censo. Para a realização da ACM foram utilizados os pacotes *{FactoMineR}* e *{factoextra}* no *software* estatístico R, versão 4.2.2, que como *default* utiliza o método da matriz indicadora para a realização do ACM.

Para uma pré-seleção de variáveis, que permitisse uma melhor visualização dos gráficos, foi realizado um teste qui-quadrado de independência ao nível de significância de 5%. Assim, foram consideradas em cada ACM apenas as variáveis cujo teste rejeitou a hipótese nula de independência em relação as 4 principais variáveis qualitativas, concernentes a modalidade de ensino, a rede da IES em que o curso foi ofertado, o nome do curso (se é apenas constituído por Ciência de Dados/ Data Science ou não) e o grau acadêmico. De acordo com Pó (2020), nos casos em que a quantidade de dados disponíveis é menor do que 40 e nos casos em que há pelo menos uma ocorrência de classe com a frequência esperada menor do que 5, utiliza-se a correção de continuidade de Yates e utilizada aqui quando necessária.

Para a variável referente ao grau acadêmico, obtiveram-se 12 variáveis significativas ao nível de 5%, ao passo que para as variáveis REDE, NOME_CD e MOD_ENSINO houve 4, 7 e 12 variáveis significativas respectivamente. Ou seja, as variáveis GRAU_ACAD e MOD_ENSINO foram as que mais apresentaram dependência com as demais, conforme o Apêndice 2, Tabela 2.2.

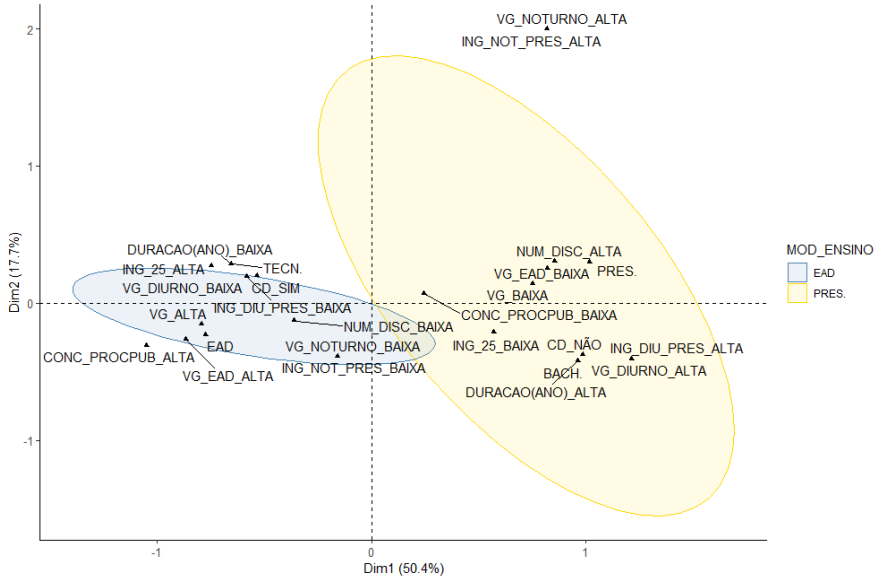
No que tange ao grau acadêmico, espera-se, em valores medianos, que sejam do tipo bacharelado os cursos com mais de 42 disciplinas, com carga horária maior do que 2692,5 horas, que ofertem mais do que 0 vagas diurno, com mais de 0 ingressos diurno e cuja duração seja maior do que 3 anos, que possuam 266 vagas ou menos, até 198,5 vagas EAD, com 30 ou menos ingressos que tenham terminado o Ensino Médio em escolas públicas, com até 12,5 ingressos do sexo feminino e até 24,5 ingressos com 25 anos ou mais e que sejam presenciais, cujo nome não é constituído apenas de Ciência de Dados (ou *Data Science*).

Figura 6 - ACM para Todos os Cursos em Relação ao Grau Acadêmico



Fonte – Autores (2023)

Figura 7 - ACM para Todos os Cursos em Relação a Modalidade de Ensino



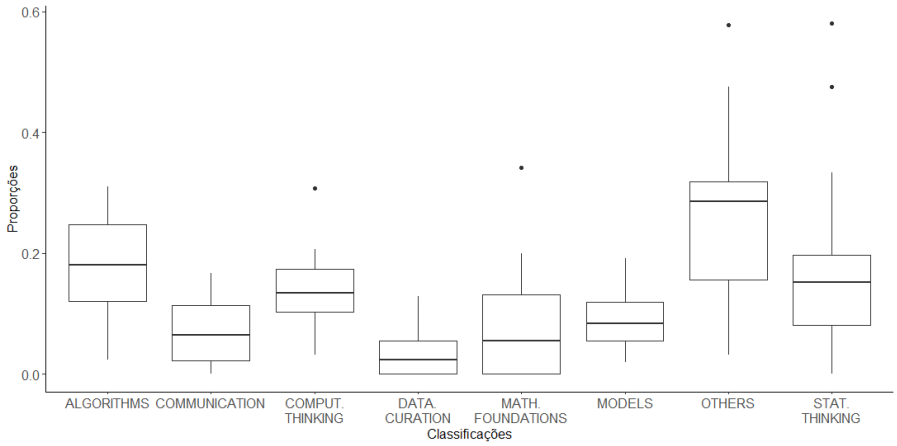
Fonte – Autores (2023)

Por fim, acerca da variável modalidade de ensino, espera-se, ainda em valores medianos, que sejam presenciais os cursos que ofertem mais do que 0 vagas no período noturno e que tenham mais de 0 ingressos neste mesmo período, que tenham mais do que 42 disciplinas, que tenham menos do que 198,5 vagas na modalidade EAD, que a quantidade de vagas seja de até 266, que tenha tido 0 concluintes que terminaram o Ensino Médio em escolas públicas, que possuam até 24,5 ingressantes com 25 anos ou mais, cujo nome não seja constituído apenas de Ciência de Dados (ou *Data Science*), que o número de ingressos diurno seja maior do que 0, bem como a quantidade de vagas no período diurno, que sejam do tipo bacharelado e que tenham uma duração maior do que de 3 anos (Figura 7). As ACM referentes a rede da IES e ao nome do curso podem ser visualizadas no Apêndice 3, Figura 3.1 e Figura 3.2.

3.5 ANÁLISE DAS DISCIPLINAS DOS CURSOS

Para realizar a análise das disciplinas, as mesmas, que haviam sido obtidas previamente nas *home pages* dos cursos por intermédio de *web scraping*, foram categorizadas em oito diretrizes que representam as áreas fundamentais de conhecimento para a formação em Ciência de Dados de acordo com De Veaux et al. (2017): *Computational thinking, Statistical thinking, Mathematical foundations, Model building and assessment, Algorithms and software foundation, Data curation, Communication and responsibility* e *Others*. O uso das diretrizes permite analisar quais áreas do conhecimento são enfatizadas em cada programa, bem como possibilita a comparação entre diferentes graus acadêmicos e redes de ensino. Na Figura 8, pode-se observar a distribuição das disciplinas em cada uma das diretrizes nos cursos. Nota-se que as diretrizes que predominam nos currículos são *Others*, *Algorithms and Software* e *Statistical Thinking*.

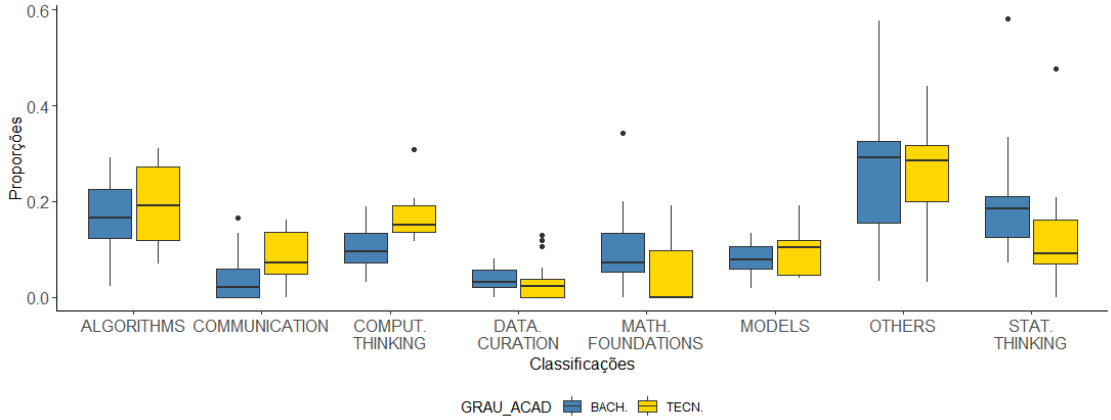
Figura 8 - Boxplot das Categorias das Disciplinas



Fonte – Autores (2023)

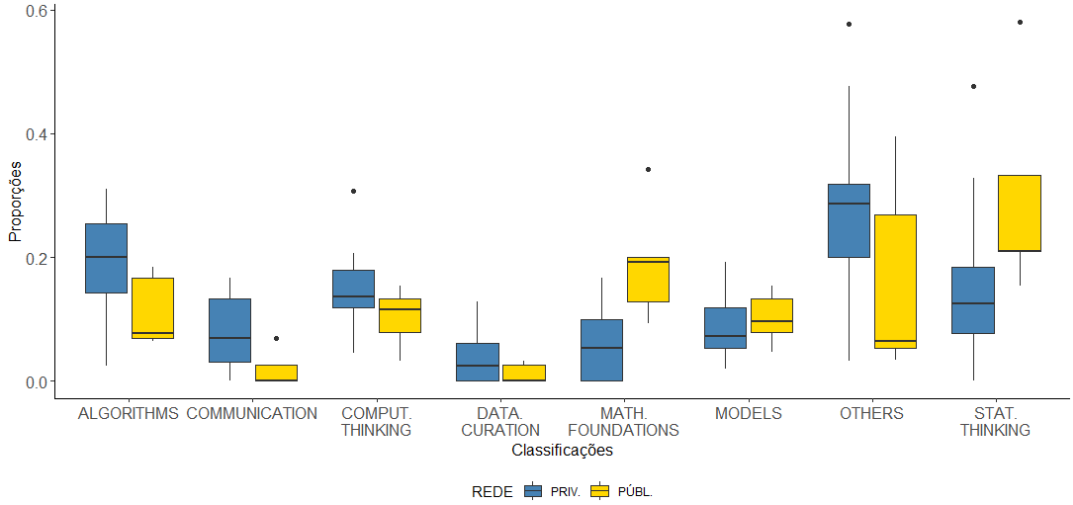
Na Figura 9, ao separar os cursos de graduação em Ciência de Dados em bacharelado e tecnólogo, percebe-se que os cursos de bacharelado tendem a enfatizar mais disciplinas de *Statistical Thinking* e *Math Foundations*. Por outro lado, os cursos de tecnólogo apresentam uma proporção sugestivamente maior de disciplinas em *Communication* e *Computational Thinking*. Na Figura 10, que divide os cursos de graduação em Ciência de Dados por instituições de ensino em rede pública e privada, podemos identificar nos cursos de rede pública uma maior presença de disciplinas em *Statistical Thinking* e *Math Foundations*, mesmo havendo apenas 10 cursos de graduação pertencentes a tal grupo, ao passo que em cursos de rede privada, observa-se uma proporção mais alta de disciplinas em *Algorithms and Software*, *Communication* e *Computational Thinking*, sendo que os cursos presentes na base de dados são majoritariamente ofertados pela rede privada (48 cursos).

Figura 9 - Boxplot das Categorias das Disciplinas por Grau Acadêmico



Fonte – Autores (2023)

Figura 10 - Boxplot das Categorias das Disciplinas por Rede da IES



Fonte – Autores (2023)

4.CONSIDERAÇÕES FINAIS

De uma forma geral, os primeiros cursos de graduação em Ciência de Dados no Brasil têm sido ofertados por IES da rede privada, sendo majoritariamente do tipo tecnológico, com um certo equilíbrio entre as modalidades de ensino presencial (48,3%) e EAD (51,7%). Grande parte dos cursos passaram a ser ofertados a partir de 2019, o que dificulta a obtenção de dados relativos aos seus egressos. Contudo, em relação aos ingressos, têm-se que, em números absolutos, a média dos ingressos com 25 anos ou mais é cerca de 3 vezes maior do que a média dos ingressos com menos de 25 anos, além de serem majoritariamente do sexo masculino e terem principalmente concluído o Ensino Médio em escolas públicas. No que concerne às regiões do país, em números absolutos, nota-se uma alta concentração de cursos presenciais de Ciência de Dados no Sudeste (60,7%), que por sua vez possui o maior número médio de disciplinas do país (47,0 disciplinas) e a maior carga horária média (3059,6 horas). Tanto a região Sul quanto a Centro-Oeste possuem todos os cursos presenciais dos quais se tem informação sendo ofertados por IES de rede privada. No Nordeste, 50% dos cursos presenciais de Ciência de Dados ofertados são do tipo bacharelado. Além disso, obteve-se que nos cursos do tipo bacharelado é esperado que a quantidade de vagas ofertadas seja menor do que no tipo tecnológico. Bem como, em números absolutos, com menos ingressantes do sexo feminino, menos ingressantes com 25 anos ou mais e menos ingressantes que tenham concluído o Ensino Médio em escolas públicas. Também, em cursos da modalidade de ensino EAD espera-se uma duração, carga horária e número de disciplinas menores do que nos cursos presenciais. Acerca das disciplinas do curso, as mesmas enfatizam o raciocínio estatístico, os algoritmos e outras áreas variadas do conhecimento. Os cursos de bacharelado geralmente apresentaram maior ênfase nas bases matemáticas e estatísticas, enquanto os cursos de tecnólogo tipicamente direcionaram-se mais às disciplinas computacionais. Com relação às redes de ensino superior, observou-se uma tendência nas instituições públicas em apresentar um currículo mais voltado para o raciocínio estatístico e matemático, enquanto as instituições privadas geralmente enfatizaram disciplinas relacionadas à área computacional e de comunicação. De uma forma geral e especificamente para as disciplinas relacionadas ao raciocínio estatístico, elas apresentam menor proporção do que as disciplinas de algoritmos e fundamentos de *softwares,* bem como também apresentam maior proporção para os cursos do tipo bacharelado e para os cursos ofertados por instituições públicas.

REFERÊNCIAS BIBLIOGRÁFICAS

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H. V., Labrinidis, A., Madden, S., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S., & Widom, J. (2011). Challenges and opportunities with big data 2011-1. Acesso em 29 de Jun de 2023. Disponível em: <http://docs.lib.purdue.edu/cctech/1>.

Carvalho, M. S. & Struchiner, C. J. (1992). Análise de correspondência: uma aplicação do método à avaliação de serviços de vacinação. Cad. Saúde Públ, 8(3), 287-301.

Classificação Brasileira de Ocupações - CBO. (2023). Acesso em 24 de Jul de 2023. Disponível em: <https://cbo.mte.gov.br/cbosite/pages/pesquisas/BuscaPorTituloResultado.jsf>.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. International Statistical Review/ Revue Internationale de Statistique, 69(1), 21-25.

Costa, A. L. A. (2016). Análise de Correspondência Simples com Novos Escores e o Uso da Análise de Correspondência Múltipla em Dados Composicionais de Granulometria de Grãos de Café. (Dissertação de mestrado). Universidade Federal Lavras (UFLA), Lavras.

Curty, R. G. & Serafim, J. S. (2016). A formação em ciência de dados: uma análise preliminar do panorama estadunidense. UEL, 21(2), 307-328.

Davenport, D. J. & Patil, T. H. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, Brighton, MA. Acesso em 22 de Ago de 2022. Disponível em: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., Tiruviluamala, N., Uhlig, P., Washington, T. M., Wesley, C. L., White, D., & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

Giordan, I. (2017). 7 curiosidades sobre a Universidade Cruzeiro do Sul. Acesso em 04 de Jul de 2023. Disponível em: <https://querobolsa.com.br/revista/ 7-curiosidades-sobre-a-universidade-cruzeiro-do-sul-unicsul>.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Análise multivariada de dados (6a ed). Porto Alegre: Bookman.

Johnson, R. A. & Wichern, D. W. (2002). Applied multivariate statistical analysis (6a ed). New Jersey: Prentice hall Upper Saddle River.

Monteiro-Krebs, L., Cappra R. & de Lima M. C. (2021). O perfil do cientista de dados no Brasil: competências e níveis de senioridade (páginas 250-272). SP: Pimenta Cultural.

Nascimento, M. M., Cavalcanti, C. & Ostermann, F. (2017). Análise de correspondência aplicada à pesquisa em ensino de ciências. In Anais do X Congresso Internacional Sobre Investigación en Didáctica de las Ciencias (pp. 1319-1324). Sevilla.

Pó, M. V. (2020). Testes não paramétricos. 19 slides. Acesso em 28 de Jun de 2023. Disponível em: <https://perguntasaopo.files.wordpress.com/2020/05/mqcs20_08_nc3a3o-paramc3a9tricos.pdf>.

Prado, M. V. B. (2012). Métodos de Análise de Correspondência Múltipla: Estudo de Caso Aplicado à Avaliação da Qualidade do Café. (Dissertação de mestrado). Universidade Federal Lavras (UFLA), Lavras.

PROPLAN - Universidade Federal do Pará (2023). Sistemas. Acesso em 26 de Jul de 2023. Disponível em: <https://proplan.ufpa.br/index.php/sistemas>.

APÊNDICE 1

Tabela 1.1 - Principais Informações Acerca dos Cursos de Ciência de Dados Presenciais que Fizeram Parte da Base de Dados

NOME DO CURSO	NOME DA IES	SIGLA DA IES	MUNICÍPIO/UF	SITE DO CURSO
Ciência de Dados e Inteligência Artificial	Centro Universitário do Instituto de Educação Superior de Brasília	IESB	Brasília/DF	iesb.br/cursos/ciencia-de-dados-e-inteligencia-artificial/
Ciência de Dados e Inteligência Artificial	Escola de Matemática Aplicada	EMAp-FGV	Rio de Janeiro/RJ	emap.fgv.br/curso/ciencia-de-dados-e-inteligencia-artificial
Ciência de Dados e Inteligência Artificial	Universidade Federal da Paraíba	UFPB	João Pessoa/PB	sigaa.ufpb.br/sigaa/public/curso/portal.jsf?id=14289031\&lc
Ciência de Dados e Inteligência Artificial	Pontifícia Universidade Católica de São Paulo	PUCSP	São Paulo/SP	pucsp.br/graduacao/ciencia-de-dados-e-inteligencia-artificial
Ciência de Dados	Universidade Anhembi Morumbi	UAM	São Paulo/SP	portal.anhembi.br/graduacao/ciencia-de-dados/
Ciência de Dados	Faculdade de Tecnologia de Adamantina		Adamantina/SP	fatec.edu.br/adamantina/ciencia-de-dados/
Ciência de Dados	Universidade de São Paulo	USP	São Carlos/SP	icmc.usp.br/graduacao/ciencia-de-dados-bacharelado
Ciência de Dados	Faculdade Tecnológica de Ourinhos	FATEC	Ourinhos/SP	fatecourinhos.edu.br/cursos/ciencia/
Ciência de Dados e Inteligência Artificial	Pontifícia Universidade Católica do Rio Grande do Sul	PUCRS	Porto Alegre/RS	pucrs.br/estudenapucrs/cursos/ciencia-de-dados-e-inteligencia-artificial
Ciência de Dados	Faculdade de Tecnologia de Santana de Parnaíba	FATEC-SPB	Santana de Parnaíba/SP	fatecsdp.edu.br/cursos/tecnologia-em-ciencia-de-dados
Ciência de Dados	Centro Universitário de João Pessoa	UNIPÊ	João Pessoa/PB	unipe.edu.br/graduacao/ciencia-de-dados/
Ciência de Dados e Inteligência Artificial	Pontifícia Universidade Católica de Campinas	PUC-CAMPINAS	Campinas/SP	puc-campinas.edu.br/graduacao/ciencia-de-dados-e-inteligencia-artificial/
Ciência de Dados para Negócios	Universidade Federal da Paraíba	UFPB	João Pessoa/PB	sigaa.ufpb.br/sigaa/public/curso/portal.jsf?id=19420831\&lc
Ciência de Dados e Inteligência Artificial	Universidade de Sorocaba	UNISO	Sorocaba/SP	uniso.br/graduacao/curso/ciencia-de-dados-e-inteligencia-artificial
Ciência de Dados	Universidade Federal do Ceará	UFC	Itapajé/CE	prograd.ufc.br/pt/cursos-de-graduacao/ciencia-de-dados-itapaje/
Ciência de Dados e Inteligência Artificial	Centro Universitário Ibmec	IBMEC	Rio de Janeiro/RJ	portal.ibmec.br/selecao

Ciência de Dados	Faculdade de Tecnologia Rubens Lara	FATEC-BS	Santos/SP	fatecrl.edu.br/cursos/ciencia-de-dados
Estatística e Ciência de Dados	Universidade de São Paulo	USP	São Carlos/SP	icmc.usp.br/graduacao/estatistica-bacharelado
Ciência de Dados	Pontifícia Universidade Católica de Minas Gerais	PUC MINAS	Belo Horizonte/MG	pucminas.br/unidade/praca-da-liberdade/ensino/graduacao/ Paginas/Ciencias-de-Dados.aspx?moda=2\&curso=
Ciência de Dados	Universidade Anhembi Morumbi	UAM	São Paulo/SP	portal.anhembi.br/graduacao/ciencia-de-dados/
Ciência de Dados	Centro Universitário das Américas	CAM	São Paulo/SP	vemprafam.com.br/cursos/ ciencia-de-dados-data-science/
Ciência de Dados	Faculdade Capital Federal	FECAF	Taboão da Serra/SP	fecaf.com.br/cursos/ciencia-de-dados
Ciência de Dados e Inteligência Analítica	Faculdade Senac Porto Alegre - FSPOA	SENAC/RS	Porto Alegre/RS	senacrs.com.br/cursos/curso-superior-de-tecnologia-em-ciencia-de-dados-e-inteligencia-analitica_WyIxNTg2lixudWxsLG51bGwsbnVsbF0
Ciência de Dados e Inteligência Analítica	Faculdade de Tecnologia Senac Pelotas	FATEC SENAC PELOTAS	Pelotas/RS	senacrs.com.br/cursos/curso-superior-de-tecnologia-em-ciencia-de-dados-e-inteligencia-analitica_WyIxNTg4lixudWxsLG51bGwsbnVsbF0
Ciência de Dados e Inteligência Artificial	Centro Universitário Fundação Santo André	CUFSA	Santo André/SP	fsa.br/graduacao/ciencia-de-dados-inteligencia-artificial/
Ciência de Dados e Machine Learning	Centro Universitário de Brasília	UNICEUB	Brasília/DF	uniceub.br/pdp/graduacao/ti/ciencia-de-dados-e-machine-learning-245
Ciência de Dados e Machine Learning	Centro Universitário de Brasília	UNICEUB	Brasília/DF	uniceub.br/pdp/graduacao/ti/ciencia-de-dados-e-machine-learning-245
Ciência de Dados para Negócios	Fae Centro Universitário	FAE	Curitiba/PR	fae.edu/cursos/182302806/ciencia+de+dados+para+negocios.htm

Fonte - Autores (2023)

Tabela 1.2 - Principais Informações Acerca dos Cursos de Ciência de Dados EAD que Fizeram Parte da Base de Dados

NOME DO CURSO	NOME DA IES	SIGLA DA IES	SITE DO CURSO
Ciência de Dados	Universidade Cruzeiro do Sul	UNICSUL	cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados	Universidade Cidade de São Paulo	UNICID	
Ciência de Dados	Universidade de Franca	UNIFRAN	cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados e Inteligência Artificial	Centro Universitário Unidom - Bosco	UNIDOM - BOSCO	unidombosco.edu.br/cursos/ciencia-de-dados-e-inteligencia-artificial-ead/
Data Science	Centro Universitário Maurício de Nassau	UNINASSAU	graduacao.uninassau.digital/nossos-cursos/cst-em-data-science/427/60/2
Data Science	Universidade da Amazônia	UNAMA	graduacao.unama.br/nossos-cursos/cst-em-data-science/427/94/2
Ciência de Dados	Centro Universitário da Serra Gaúcha	FSG	cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados	Centro Universitário Estácio de Santa Catarina - Estácio Santa Catarina		estacio.br/cursos/graduacao/ciencia-de-dados
Ciência de Dados	Universidade do Vale do Itajaí	UNIVALI	ead.univali.br/cursos-graduacao/ciencia-de-dados-ead
Ciência de Dados	Centro Universitário Estácio de Ribeirão Preto	ESTÁCIO RIBEIRÃO PRE	portal.estacio.br/graduacao/ci\%C3\%AAncia-de-dados
Ciência de Dados	Centro Universitário Internacional	UNINTER	uninter.com/graduacao-ead/ciencia-de-dados-2/
Data Science	Centro Universitário do Norte	UNINORTE	graduacao.uninorte.com.br/nossos-cursos/cst-em-data-science/427/130/2
Ciência de Dados	Universidade Estácio de Sá	UNESA	estacio.br/cursos/graduacao/ciencia-de-dados
Ciência de Dados	Universidade Nove de Julho	UNINOVE	uninove.br/cursos/graduacao-ead/ead/tecnologia-em-ciencia-de-dados-ead
Ciência de Dados	Universidade Vila Velha	UVV	uvv.br/ead/graduacao/ciencia-de-dados/
Ciência de Dados	Fundação Universidade Virtual do Estado de São Paulo	UNIVESP	univesp.br/cursos/bacharel-em-ciencia-de-dados
Ciência de Dados	Centro Universitário Braz Cubas		cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados	Universidade Positivo	UP	cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados	Centro Universitário Ritter dos Reis	UNIRITTER	uniritter.edu.br/graduacao/ciencia-de-dados/
Ciência de Dados	Centro Universitário Joaquim Nabuco De Recife	UNINABUCO RECIFE	graduacao.uninabuco.digital/nossos-cursos/ciencia-de-dados/618/5/2
Ciência de Dados	Universidade Anhanguera	UNIDERP	anhanguera.com/curso/ciencia-de-dados-tecnologo
Data Science	Universidade Universus Veritas Guarulhos	Univeritas UNG	graduacao.ung.br/nossos-cursos/cst-em-data-science/427/32/2

Ciências de Dados e Análise de Comportamento	Universidade Cesumar	UniCesumar	inscricoes.unicesumar.edu.br/curso/ciencias-de-dados-e-analise-de-comportamento
Marketing Digital e Data Science	Centro Universitário das Américas	CAM	famonline.com.br/cursos/marketing-digital-data-science/
Ciência de Dados	Centro Universitário Anhanguera Pitágoras Ampli		ampli.com.br/graduacao/ciencia-de-dados
Ciência de Dados e Inteligência Artificial	Universidade de Sorocaba	UNISO	uniso.br/graduacao/virtual/ciencia-de-dados-ead
Ciência de Dados	Universidade Católica de Santos	UNISANTOS	ead.unisantos.br/cursos-graduacao/ciencia-de-dados-ead
Ciência de Dados	Centro Universitário de João Pessoa	UNIPÊ	cruzeirodosulvirtual.com.br/graduacao/ciencia-de-dados/
Ciência de Dados	Centro Universitário Anhanguera Pitágoras Unopar de Niterói	UNIAN-RJ	unopar.com.br/curso/ciencia-de-dados-tecnologo/
Ciência de Dados	Centro Universitário Anhanguera Pitágoras Unopar de Campo Grande		pitagoras.com.br/curso/ciencia-de-dados-tecnologo/

Fonte - Autores (2023)

APÊNDICE 2

Tabela 2.1 Codificação e Mediana das Variáveis Categorizadas para a Análise Descritiva Multivariada

VARIÁVEL	CODIFICAÇÃO
MOD_ENSINO	{PRES.; EAD}
NOME_CD	{CD_SIM; CD_NÃO}
GRAU_ACAD	{TECN.; BACH.}
REDE	{PÚBL.; PRIV.}
VG	{BAIXA: <=266,0; ALTA: >266,0}
VG_DIURNO	{BAIXA: <=0,0; ALTA: >0,0}
VG_NOTURNO	{BAIXA: <=0,0; ALTA: >0,0}
VG_EAD	{BAIXA: <=198,5; ALTA: >198,5}
ING	{BAIXA: <=40,0; ALTA: >40,0}
ING_MASC	{BAIXA: <=32,5; ALTA: >32,5}
ING_FEM	{BAIXA: <=12,5; ALTA: >12,5}
ING_DIU_PRES	{BAIXA: <=0,0; ALTA: >0,0}
ING_NOT_PRES	{BAIXA: <=0,0; ALTA: >0,0}
ING_PROCPUB	{BAIXA: <=30,0; ALTA: >30,0}
ING_PROCPRI	{BAIXA: <=14,0; ALTA: >14,0}
CONC	{BAIXA: <=0,0; ALTA: >0,0}
CONC_FEM	{BAIXA: <=0,0; ALTA: >0,0}
CONC_MASC	{BAIXA: <=0,0; ALTA: >0,0}
CONC_PROCPUB	{BAIXA: <=0,0; ALTA: >0,0}
CONC_PROCPRI	{BAIXA: <=0,0; ALTA: >0,0}
ING_0_24	{BAIXA: <=20,0; ALTA: >20,0}
ING_25	{BAIXA: <=24,5; ALTA: >24,5}
CONC_0_24	{BAIXA: <=0,0; ALTA: >0,0}
CONC_25	{BAIXA: <=0,0; ALTA: >0,0}
DURACAO(ANO)	{BAIXA: <=3,0; ALTA: >3,0}
CARGA_HOR	{BAIXA: <=2692,5; ALTA: >2692,5}
NUM_DISC	{BAIXA: <=42,0; ALTA: >42,0}

Fonte - Autores (2023)

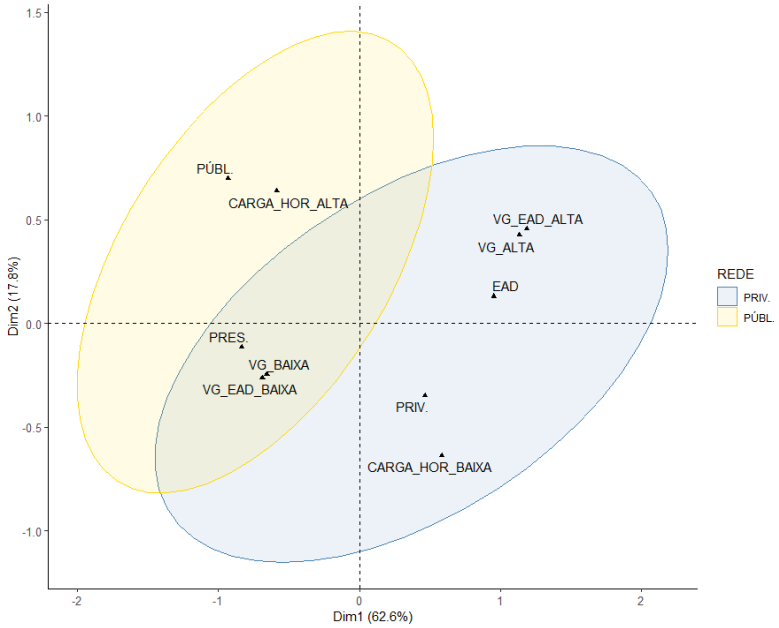
Tabela 2.2 - p-valores Obtidos no Teste Qui-quadrado de Independência para as Principais Variáveis Qualitativas

	GRAU_ACAD	REDE	NOME_CD	MOD_ENSINO
GRAU_ACAD		0,518	**<0,001**	**0,001**
MOD_ENSINO	**0,001**	**0,001**	**0,013**	
VG	**0,005**	**0,012**	0,106	**<0,001**
VG_NOTURNO	0,631	0,117	1,000	**<0,001**
VG_DIURNO	**<0,001**	0,152	**<0,001**	**<0,001**
VG_EAD	**0,005**	**0,012**	0,106	**<0,001**
ING	0,180	0,600	0,906	0,503
ING_MASC	0,117	0,719	0,746	0,358
ING_FEM	**0,028**	0,719	0,331	0,358
ING_DIU_PRES	**<0,001**	0,152	**<0,001**	**<0,001**
ING_NOT_PRES	0,631	0,117	1,000	**<0,001**
ING_PROCPUB	**0,009**	0,600	0,157	0,199
ING_PROCPRI	1,000	0,719	0,746	0,759
CONC	0,919	0,507	1,000	0,244
CONC_FEM	1,000	0,627	1,000	0,369
CONC_MASC	0,740	0,767	0,970	0,159
CONC_PROCPUB	0,252	0,219	0,380	**0,027**
CONC_PROCPRI	1,000	0,767	1,000	0,539
ING_0_24	0,754	0,719	0,746	1,000
ING_25	**<0,001**	1,000	**0,005**	**0,008**
CONC_0_24	1,000	0,767	1,000	0,539
CONC_25	0,548	0,627	0,770	0,097
DURACAO(ANO)	**<0,001**	0,483	**0,009**	**0,004**
CARGA_HOR	**0,010**	**0,053**	0,264	0,067
NUM_DISC	**0,021**	0,111	**0,039**	**0,005**
NOME_CD	**<0,001**	1,000		**0,013**

Fonte - Autores (2023)

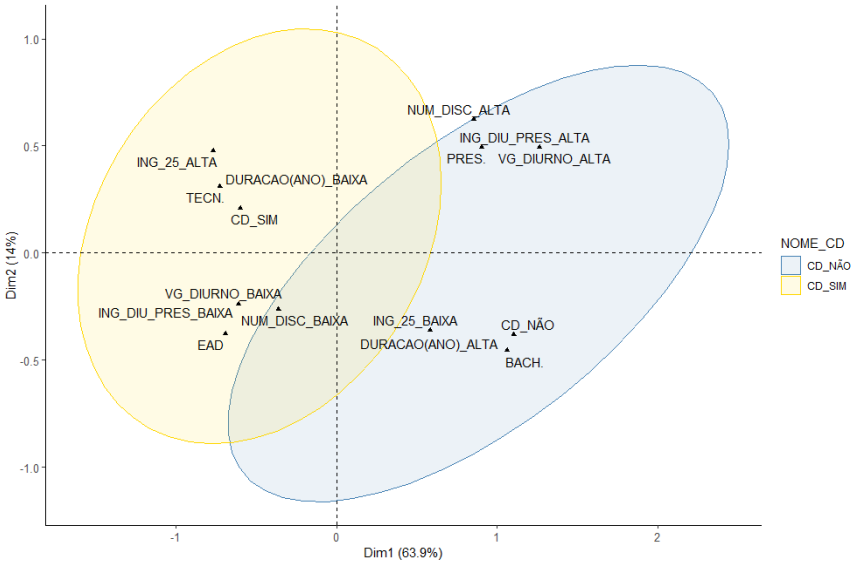
APÊNDICE 3

Figura 3.1 - ACM para Todos os Cursos em Relação a Rede da IES



Fonte – Autores (2023)

Figura 3.2 - ACM para Todos os Cursos em Relação ao Nome do Curso



Fonte – Autores (2023)

Revista Brasileira de Estatística, V.81, N. 248, p.62-75, 2023

MODELOS DE DADOS FALTANTES EM EXTREMOS

Fernando Ferraz do Nascimento

fernandofn@ufpi.edu.br

Universidade Federal do Piauí

Zeferino Gomes da Silva Neto

zeferinon@gmail.com

Universidade Federal de Viçosa

Francisco Carlos Melo Filho

melocarlosf05@hotmail.com

Universidade Federal do Piauí

Resumo: A Teoria dos Valores Extremos (TVE) tem se tornado uma área cada vez mais importante no estudo do comportamento de fenômenos naturais que provocam grandes danos à sociedade. Basicamente, esta análise consiste em estudar uma série temporal de uma variável ambiental. Em valores extremos, as duas principais abordagens utilizadas são; análise de máximos, em que se utiliza máximos mensais ou anuais e realizamos a modelagem pela distribuição de Valores Extremos Generalizados (*Generalized Extreme Value*, GEV); outra abordagem é analisar os excessos acima de um determinado limite através da Distribuição Generalizada de Pareto (*Generalized Pareto Distribution*, GPD). Em ambas as situações, pode haver período da série em que não há dados registrados, sendo estes classificados como valores faltantes. Este trabalho apresenta uma nova abordagem para valores extremos, sob o paradigma bayesiano onde além da estimação dos parâmetros variantes no tempo da distribuição, os valores faltantes no tempo também são parâmetros a serem estimados. Aplicações de níveis de precipitação e temperatura em cidades da região Nordeste do Brasil mostraram que o modelo é eficiente em estimar valores faltantes com grande precisão.
Palavras-chave: Eventos extremos; dados faltantes; Bayes; MCMC; precipitação; dados de temperatura.

Abstract: Extreme Value Theory (EVT) has become an increasingly important area in the study of behavior of natural phenomena that provoke a great damage to the society. Basically, this analysis consists of studying a time series of an environmental variable. In extreme values, the two main approaches used are; analysis of maximums, in which we took monthly or annual maxima and perform the modeling by the Generalized Extreme Values (GEV) distribution; another approach is to analyze excesses above a given threshold through the Generalized Pareto Distribution (GPD). In both situations, there may be period in the series in which there is no data recorded, these being classified as missing values. This work presents a new approach to extreme values, under the Bayesian paradigm where in addition to the estimation of the time-varying parameters of the distribution, the missing values in time are also parameters to be

estimated. Applications of rainfall and temperature levels in cities in the northeastern region of Brazil showed that the model is efficient in estimating missing values with great accuracy.
Keywords: Extreme events; Missing data; Bayesian; MCMC; precipitation; temperature data.

1.INTRODUÇÃO

Fenômenos extremos possuem grande importância em diversas áreas do conhecimento, dentre elas, a climatologia e outras áreas ambientais. Tais fenômenos também têm grande destaque na economia e ajudam a estimar danos, sendo suas estimativas probabilísticas fundamentais para o planejamento de ações futuras. Na área climatológica, as chuvas têm sido amplamente estudadas em diversas regiões do mundo devido à sua relevância no ciclo hidrológico e na preservação dos seres vivos na Terra.

As secas representam um problema grave para a sociedade e para os ecossistemas naturais (Dinpashoh, Fakheri-Fard e Moghaddam, 2004). Assim, inúmeras metodologias têm sido usadas para analisar a variabilidade da precipitação.

O aquecimento global tem sido um tema importante nos últimos anos devido à rápida mudança climática mudança que o planeta vem sofrendo nas últimas décadas. Estas mudanças implicam em algumas mudanças no número de eventos de temperatura extrema. Mudanças em extremos de temperatura são mais responsáveis pelas mudanças na natureza do que mudanças na temperatura média (Parmesan, Root and Willing, 2000).

Verões extremamente quentes ou invernos muito frios podem influenciar a agricultura, o consumo de energia, além de ocasionar maiores problemas de saúde da população. Compreender com que frequência eventos extremos ocorrem é importante para mitigar o impacto na sociedade das perdas e danos que esses eventos possam acarretar.

1.1. A DISTRIBUIÇÃO GEV

Os primeiros estudos com distribuição GEV vêm da distribuição máxima de alguns casos específicos, onde era necessário conhecer a distribuição comum F_x. Contudo, em alguns casos esta distribuição é desconhecida, sendo necessário obter resultados assintóticos. Von Mises (1954) e Jenkinson (1955) propuseram a distribuição GEV, denotada por G e com função distribuição acumulada dada por:

$$G(x|\xi,\sigma,\mu) = \begin{cases} exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right\} se\ \xi \neq 0; \\ exp\left\{-exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right\} se\ \xi = 0; \end{cases} \quad (1)$$

onde $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$.

Esta distribuição possui 3 parâmetros: um parâmetro de locação μ, de escala σ e de forma ξ. O caso em que $\xi = 0$ a distribuição G corresponde a distribuição *Gumbel*. O caso ξ<0 corresponde a distribuição de *Frechet* e o caso $\xi > 0$ corresponde a distribuição *Weibull*.

1.1. A DISTRIBUIÇÃO GPD

Suponha que nós temos $X_1, X_2, \cdots, X_N$ variáveis com função de distribuição F. Seja x_F o limite superior de F. Considere u sendo o limiar suficientemente alto. Nós denotamos como excess os valores X_i tal que $X_i > u$ e N_u é o número de valores maiores do que u, $N_u = \sum_{i=1}^{N} 1_{(x_i>U)}$ onde $1_{(X_i>U)} = 1$ se $X_i > u$, e 0 caso contrário

A distribuição de Pareto Generalizada (GPD) foi desenvolvida por Pickands (1975) e foi baseada no seguinte teorema:

Teorema 1.1. Seja X uma variável aleatória com função de distribuição F, pertencente ao domínio de atração de uma distribuição GEV. Então quando $u \to \infty$, $F(x|u) = P(X \leq u + x|X > u)$ tem distribuição GPD, com a seguinte função distribuição

$$F(x|\xi,\sigma,u) = \begin{cases} 1-\left(1+\xi\frac{(x-u)}{\sigma}\right)^{-1/\xi} & if\ \ \xi \neq 0, \\ 1-exp\left\{-\frac{x-u}{\sigma}\right\}, & if\ \ \xi = 0. \end{cases} \quad (2)$$

onde $x-u>0$ para $\xi \geq 0$ e para $\xi < 0$. O caso $\xi = 0$ é a distribuição exponencial.

1.2.1 ESCOLHENDO O LIMIAR

É necessário ter muita cautela na escolha do limiar, pois um valor muito alto de u implicará em um pequeno número de observações na cauda, podendo resultar em uma maior variabilidade dos estimadores. Por outro lado, um limiar que não seja suficientemente elevado não satisfaz os pressupostos teóricos da GPD e pode resultar em estimativas erradas.

Alguns trabalhos anteriores citam métodos gráficos de escolha do limiar. Coles (2001) utiliza o gráfico MRL, baseado na média empírica dos excessos para diferentes limiares, escolhendo o valor ótimo do limiar que apresenta

tendência linear de crescimento da média. Cunnane (1979) utiliza o Gráfico de Índice de Dispersão (DIPLOT), que se baseia na razão entre a média amostral e a variância amostral do excesso para diferentes limiares, escolhendo o valor onde essa razão se aproxima de 1.

Outros trabalhos estimam o limiar como parâmetro do modelo. Scarrott e Macdonald (2012) fazem uma ampla revisão dos vários métodos de estimativa do limiar. Alguns deles utilizam uma abordagem Bayesiana para estimar os parâmetros do modelo. Behrens, Gamerman e Lopes (2004) consideram uma distribuição Gama para dados abaixo do limite u, enquanto Nascimento, Gamerman e Lopes (2012) utilizam uma abordagem semiparamétrica utilizando mistura finita de distribuições.

1.2.1 MODELANDO EXTREMOS COM DADOS FALTANTES

Uma dificuldade frequente em qualquer pesquisa científica é a incidência de dados faltantes. Para resolver estes problemas surgiram as técnicas de imputação de dados faltantes (Nunes, Kluck e Fachel, 2009). Essas técnicas visam preencher os dados faltantes e viabilizar a análise com todas as observações. No contexto Bayesiano, podemos considerar as observações faltantes como parâmetros a serem estimados. Desta forma, o vetor paramétrico possui, além dos parâmetros do modelo, o vetor de valores faltantes. Assim, podemos fazer a inferência Bayesiana para amostrar pontos de dados faltantes considerando suas distribuições condicionais completas através de procedimentos iterativos como a técnica Markov Chain Monte Carlo (MCMC) (Gamerman e Lopes, 2006).

O objetivo deste trabalho é apresentar uma metodologia para modelagem de valores extremos na presença de dados faltantes, tanto para abordagens por máximos através da distribuição GEV quanto para os excessos através da distribuição GPD. Em ambos os casos, as observações faltantes são parâmetros a serem estimados pelo modelo. Para realizar esta modelagem e estimativa de dados faltantes capturam o comportamento específico no tempo, consideramos modelos com parâmetros variantes no tempo para as distribuições GEV estendendo o modelo de Huerta e Sansó (2007), e para a distribuição GPD considerando o modelo proposto por Nascimento, Lopes e Gamerman (2016).

2. MODELOS DINÂMICOS PARA EXTREMOS

Nas séries temporais, investiga-se como a variação dos dados pode mudar ao longo do tempo. Esse tipo de alteração é comum para dados de valores extremos. Ao avaliar dados ambientais, por exemplo, precipitação, velocidade do vento e temperatura, seus níveis podem estar correlacionados com a sazonalidade, além de tendência crescente ao longo dos anos devido às mudanças climáticas no planeta (Nascimento, Gamerman e Lopes, 2012).

No modelo de Nascimento, Lopes e Gamerman (2016), cada observação x_t, $t = 1, \dots, n$, que é menor do que o limiar u, é modelado usando uma abordagem semi-paramétrica, e uma observação x_t maior que o limiar é modelado pela distribuição GPD. Neste modelo, dois dos parâmetros da GPD mudam com o tempo, enquanto que o limiar permanece estático.

2.1 MODELAGEM DOS EXCESSOS VARIANDO NO TEMPO

No modelo de Nascimento, Lopes and Gamerman (2016), a distribuição complete da densidade antes e depois do limiar é dada por

$$H_t(x_t|\mu,\eta,p,u,\xi_t,\sigma_t) = \begin{cases} F_G(x_t|\mu,\eta,p) \;\; se\; x_t < u \\ F_G(u,\mu,\eta) + [1 - F_G(u,\mu,\eta)]F(x_t|u,\xi_t,\sigma_t)\; se\; x_t > u \end{cases}, \quad (3)$$

onde $F_G(.|\mu,\eta,p)$ é a distribuição de Mistura finite de Gamas, definida em Wiper, Rios Insua e Ruggeri (2001) e F é a distribuição GPD como em (2).

O modelo linear dinâmico (DLM), visto por exemplo em West e Harrison (1997), é usado para modelar os parâmetros da GPD, que mudam com o tempo através da seguinte estrutura

$$\begin{pmatrix} \xi_t \\ \sigma_t \end{pmatrix} = F_t'\beta_t + \nu_t \text{ e } \beta_t = G_t\beta_{t-1} + \omega_t \quad (4)$$

Nascimento, Lopes e Gamerman (2016), propondo que a GPD(ξ_t, σ_t, u) tem os parâmetros da cauda reparametrizados, com a dinâmica em $(\ell\sigma_t, \ell\xi_t)$, where $\sigma_t = \exp(\ell\sigma_t)$ e $\xi_t = \exp(\ell\xi_t) - 1$.

O modelo DLM para os parâmetros da cauda são dados por

$$\begin{aligned} \ell\xi_t &= \theta_{\xi,t} + \nu_{\xi,t} \\ \theta\xi_t &= \theta_{\xi,t-1} + \omega_{\xi,t} \\ \ell\sigma_t &= \theta_{\sigma,t} + \nu_{\sigma,t} \\ \theta\sigma_t &= \theta_{\sigma,t-1} + \omega_{\xi,t} \end{aligned} \quad (5)$$

onde, $\theta_{\xi,t}, \theta_{\sigma,t}, t = 0, \dots, n, V_\xi, W_\xi, V_\sigma, W_\sigma$ são os hiperparâmetros do modelo.

2.2 MODELAGEM DINÂMICA PARA OS MÁXIMOS DA DISTRIBUIÇÃO

Para a distribuição de máximos, a abordagem é um pouco mais simples, pois não há quebra de distribuição como no modelo da subseção anterior, mas para a distribuição GEV temos 3 parâmetros. As formas dinâmicas dos parâmetros σ_t e ξ_t é feita da mesma maneira do que dos parâmetros da GPD, mas agora a atualização de μ_t é feita da seguinte maneira

$$\mu_t = \theta_{\mu,t} + \nu_{\mu,t} \tag{6}$$

$$\theta_{\mu,t} = \theta_{\mu,t-1} + \omega_{\mu,t}$$

onde, $\theta_{\mu,t}, t = 0, \dots, n, V_\mu, W_\mu$ são os hiperparâmetros do modelo. Então, dada uma amostra de k máximos $Y_1, \dots, Y_k$, nós temos

$$X_t \sim GEV(\mu_t, \sigma_t, \xi_t), \tag{7}$$

cuja distribuição é da mesma forma como em (1).

3. INFERÊNCIA BAYESIANA PARA EXTREMOS COM DADOS FALTANTES

A grande novidade deste trabalho é realizar a estimação de dados faltantes para dados extremos, tanto para excessos quanto para máximos. Seja A o conjunto de valores faltantes. Desta forma, a função de verossimilhança dos dados pode ser decomposta em duas partes

$$L_{mispar}(\theta, X_{mis}|X_{obs}) = L(\theta, X_{obs}|X_{mis}) = f(X_{obs}, X_{mis}|\theta), \tag{8}$$

onde X_{mis} são os valores faltantes, X_{obs} são os valores observados e θ é o vetor paramétrico do modelo.

Considerando uma função para (θ, X_{mis}) com X_{obs} fixado, e estimando θ maximizando $L_{mispar}(\theta, X_{mis}|X_{obs})$ em θ e X_{mis}.

Considerando uma distribuição a priori uniforme para os dados faltantes, temos que a distribuição condicional completa de cada dado faltante X_{mis} pode ser amostrado pela distribuição como em (3) para dados de excessos, e como (7) para a distribuição dos máximos.

3.1 DISTRIBUIÇÃO *GPD*

Considerando a independência entre os parâmetros da cauda, e na cauda tomando apenas a dependência dos parâmetros do modelo dinâmico em (5), a distribuição a priori dos parâmetros do modelo como em (3) pode ser escrita da seguinte forma:

$$\pi\left(\mu, \eta, \boldsymbol{p}, l\sigma, l\xi, u, \theta_\sigma, \theta_\xi, V_\sigma, V_\xi, W_\sigma, W_\xi\right) = \pi(\mu, \eta, \boldsymbol{p})\pi(u)\pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma) \times \pi(l\xi, \theta_\xi, V_\xi, W_\xi),$$

As prioris para Mistura de Gamas da parte da não cauda são escolhidas como em Nascimento, Lopes e Gamerman (2016), dadas por

$$\pi(\Xi) \propto \prod_{i=1}^{k} \left(\eta_j^{a_j - 1} \exp(-b_j \eta_j)\, \mu_j^{-c_j - 1} \exp\left(-\frac{d_j}{\mu_j}\right) p_j^{\gamma_j - 1} \right) I_R(\underline{\mu}).$$

A priori dos parâmetros $\pi(l\xi, \theta_\xi, V_\xi, W_\xi)$ e $\pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma)$ podem ser escritas como

$$\pi(l\xi, \theta_\xi, V_\xi, W_\xi) = \prod_{t=1}^{n} (\pi(l\xi_t | \theta_{\xi,t}, V_\xi)\pi(\theta_{\xi,t} | \theta_{\xi,t-1}, W_\xi))\pi(\theta_\xi, 0)\pi(V_\xi)\pi(W_\xi)$$

$$\pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma) = \prod_{t=1}^{n} (\pi(l\sigma_t | \theta_{\sigma,t}, V_\sigma)\pi(\theta_{\sigma,t} | \theta_{\sigma,t-1}, W_\sigma))\pi(\theta_0, \sigma)\pi(V_\sigma)\pi(W_\sigma)$$

onde N é Normal e G é a distribuição Gama.

A escolha dessas estruturas de distribuição a priori, de acordo com a forma dinâmica em (5), visa facilitar os cálculos de distribuições posterioris, sendo possível encontrar distribuições condicionais completas conhecidas para a maioria dos parâmetros, e realizar estimativas pelo amostrador de Gibbs.

Então, com base na função de verossimilhança e na distribuição anterior, a densidade posterior pode ser escrita como:

$$\pi(\Theta, x) \propto \prod_{t: x_t < u, x_t \notin A} [F_G(x_t | \mu, \eta, p)] \prod_{x_t \geq u, x_t \notin A} [(1 - F_G(x_t | \mu, \eta, p)) f(x_t | \xi_t, \sigma_t, u)]$$

$$\times \prod_{t: x_t < u, x_t \in A} [F_G(x_t | \mu, \eta, p)] \prod_{x_t \geq u, x_t \in A} [(1 - F_G(x_t | \mu, \eta, p)) f(x_t | \xi_t, \sigma_t, u)]$$

$$\times V_\xi^{\frac{n}{2} + f_\xi - 1} exp\left(-\frac{V_\xi}{2} \sum_{t=1}^{n} (l\xi_t - \theta_{\xi,t})^2 - o_\xi V_\xi \right) exp\left(-\frac{1}{2C_{\xi,0}} (\theta_{\xi,0} - m_{\xi,0})^2 \right)$$

$$\times W_\xi^{\frac{n}{2} + l_\xi - 1} exp\left(-\frac{W_\xi}{2} \sum_{t=1}^{n} (\theta_{\xi,t} - \theta_{\xi,t-1})^2 - m_\xi W_\xi \right)$$

$$\times V_\sigma^{\frac{n}{2} + f_\sigma - 1} exp\left(-\frac{V_\sigma}{2} \sum_{t=1}^{n} (l\sigma_t - \theta_{\sigma,t})^2 - o_\sigma V_\sigma \right) exp\left(-\frac{1}{2C_{\sigma,0}} (\theta_{\sigma,0} - m_{\sigma,0})^2 \right)$$

$$\times W_\sigma^{\frac{n}{2} + l_\sigma - 1} exp\left(-\frac{W_\sigma}{2} \sum_{t=1}^{n} (\theta_{\sigma,t} - \theta_{\sigma,t-1})^2 - m_\sigma W_\sigma \right) \quad (9)$$

É possível encontrar uma forma conhecida para a distribuição condicional completa para a maioria dos parâmetros definidos no modelo linear dinâmico em (5). A única exceção é para $l\xi_t$ e $l\sigma_t$ quando $x_t > u$. O método de se obter esta estimação é similar ao método de Nascimento, Lopes e Gamerman (2016).

Para amostrar uma observação faltante x_t, baseado na distribuição 3.2, nós temos

$$\pi(x_t|\Theta_{-x_t}) \sim H(x_t|\mu_t, \eta_t, \xi_t, \sigma_t, u)$$

Onde H é a distribuição em (3).

3.2 DISTRIBUIÇÃO GEV

Considerando os parâmetros da GEV variando no tempo de acordo com o modelo dinâmico descrito em (5) e (6), a distribuição a priori pode ser escrita como:

$$\pi(l_\sigma, l_\xi, \mu, \theta_\sigma, \theta_\xi, \theta_\mu, V_\sigma, V_\xi, V_\mu, W_\sigma, W_\xi), W_\mu$$
$$= \pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma) \times \pi(l\xi, \theta_\xi, V_\xi, W_\xi) \times \pi(l\mu, \theta_\mu, V_\mu, W_\mu)$$

A distribuição a posteriori dos parâmetros $\pi(l\xi, \theta_\xi, V_\xi, W_\xi)$, $\pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma)$ e $\pi(l\mu, \theta_\mu, V_\mu, W_\mu)$ pode ser escrita como

$$\pi(l\xi, \theta_\xi, V_\xi, W_\xi) = \prod_{t=1}^{n} \left(\pi(l\xi_t|\theta_{\xi,t}, V_\xi)\pi(\theta_{\xi,t}|\theta_{\xi,t-1}, W_\xi)\right) \pi(\theta_{\xi,0})\pi(V_\xi)\pi(W_\xi)$$

$$\pi(l\sigma, \theta_\sigma, V_\sigma, W_\sigma) = \prod_{t=1}^{n} \left(\pi(l\xi_\sigma|\theta_{\sigma,t}, V_\sigma)\pi(\theta_{\sigma,t}|\theta_{\sigma,t-1}, W_\sigma)\right) \pi(\theta_{0,\sigma})\pi(V_\sigma)\pi(W_\sigma)$$

$$\pi(l\mu, \theta_\mu, V_\mu, W_\mu) = \prod_{t=1}^{n} \left(\pi(\mu_t|\theta_{\mu,t}, V_\mu)\pi(\theta_{\mu,t}|\theta_{\mu,t-1}, W_\mu)\right) \pi(\theta_{\mu,0})\pi(V_\mu)\pi(W_\mu)$$

Além disso, baseado na função de verossimilhança e na distribuição a priori, a densidade a posteriori tem a forma proporcional dadapor:

$$\pi(\Theta, x) \propto \prod_{x_t \notin A} [g(x_t|\xi_t, \sigma_t, \mu_t)] \times \prod_{x_t \in A} [g(x_t|\xi_t, \sigma_t, \mu_t)]$$
$$\times \pi(l_\sigma, l_\xi, \mu, \theta_\sigma, \theta_\xi, \theta_\mu, V_\sigma, V_\xi, V_\mu, W_\sigma, W_\xi, W_\mu) \quad (10)$$

É possível encontrar uma forma conhecida para a distribuição condicional complete para a maioria dos parâmetros definidos (5) da mesma

maneira como no caso da distribuição GPD. A única exceção é para $l\xi_t$, $l\sigma_t$ e μ_t que são amostrados pelo algoritmo de Metropolis-Hastings.

Para amostrar uma observação faltante x_t, baseado na distribuição a posteriori em (10), nós temos que

$$\pi(x_t|\Theta_{-x_t}) \sim g(x_t|\mu_t, \eta_t, \xi_t, \sigma_t)$$

Onde g é a densidade da GEV com em (1.1).

4. APLICAÇÕES

4.1 DADOS PLUVIOMÉTRICOS

Os dados de precipitação pluvial foram analisados em Fortaleza, cidade localizada na região Nordeste do Brasil, onde os dados foram coletados diariamente, de 1993 a 2016. A Figura 2 mostra a série de dados de precipitação de Fortaleza. Observe que existem alguns longos períodos em que não há registro de precipitação.

Para esta aplicação, analisamos os dados conforme o modelo de Nascimento, Lopes e Gamerman (2016), incluindo agora a abordagem deste trabalho para estimar os dados faltantes desta série conforme procedimento visto na Seção 3.1.

A Tabela 1 mostra a média posterior dos parâmetros de mistura Gama e GPD de cauda, sendo dois componentes para a mistura de Gamas, com intervalos de credibilidade de 95%. A tabela mostra os parâmetros de cada faixa e seu peso na mistura, sendo expresso pelo parâmetro p, percebe-se que o componente 1 obtém um peso um pouco maior, com média de 0,54, enquanto o componente 2 tem a média de 0,46. Em relação a $\theta_{0,\xi}$, a distribuição encontrada gira em torno de 0,74 com desvio padrão de aproximadamente 0,05, enquanto a $\theta_{0,\sigma}$ apresenta média de 0,101 com desvio padrão de aproximadamente 0,07. Ainda para a Tabela 1, em relação às variâncias da estrutura dinâmica, a média e a variância da distribuição são praticamente os mesmos dados da distribuição anterior, ou seja, os dados fornecem poucas informações sobre os parâmetros, esse mesmo resultado foi mostrado nas simulações realizadas por Nascimento, Lopes e Gamerman (2016).

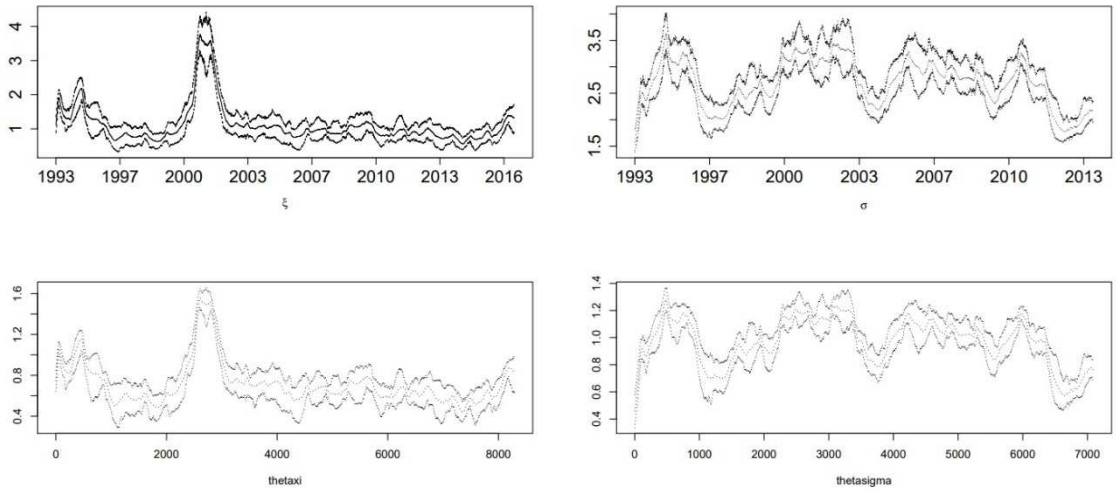
Tabela 1 – Média e interval de confiança de 95% CI para estimação dos dados de Fortaleza

Parâmetro	μ_1	μ_2	η_1	η_2
Média	0,0102	9,1493	5,6898	0,6352
Intervalo	(0,0101;0,0103)	(8,5233;9,8266)	(5,0954;5,9738)	(0,6012;0,6672)
Parâmetro	$p1$	$p2$	$\theta_{0.\xi}$	$\theta_{0.\sigma}$
Média	0,5404	0,4595	0,7381	0,1012
Intervalo	(0,5294;0,5521)	0,4478;0,4706)	(0,6305;0,8381)	(-0,0438;0,2320)
Parâmetro	V_ξ	V_σ	W_ξ	W_σ
Média	1990,6	1998,6	9968,7	1000,0
Intervalo	(1816,3;2181,1)	(1942,7;2064,3)	(9783,3;1015,3)	(9937,3;1005,9)

A Figura 1 mostra o interval de 95% para $\xi, \sigma, \theta_\xi, \theta_\sigma$. Nós Podemos notar que para os parâmetros ξ e θ_ξ há um aumento onde há uma grande parcela de dados não observados. Em relação a σ e θ_σ, a estimação se comporta com muitas oscilações, tendo um decréscimo entre as parcelas com muitos dados faltantes seguidos, e no final diminui também no final das observações. Portanto, em períodos com dados faltantes, o parâmetro ξ é maior do que em outros períodos. Um comportamento inverso é observado para σ. Em relação aos períodos sem dados faltantes, os parâmetros oscilam de maneira mais estável.

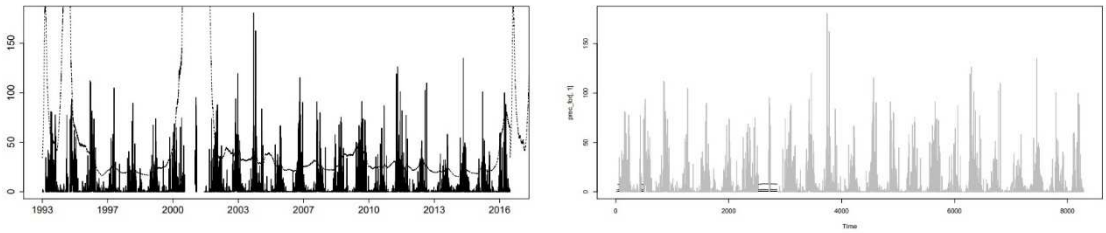
A parte a esquerda da Figura 2 mostra os valores observados e estimados do quantil 99% para a cidade de Fortaleza ao longo do tempo. Podemos notar que o quantil observado captura bem a oscilação, porém, onde existe uma grande lacuna com dados faltantes em que a curva quantílica é difícil de visualizar graficamente, assumindo valores muito elevados, a solução para gerar esses valores muito elevados é um dos objetivos na continuidade da pesquisa. A parte direita da Figura 2 mostra o gráfico pluviométrico registrado em Fortaleza-CE, após o preenchimento dos dados faltantes, realizado com o modelo de mistura de Gamas com o GPD, onde não podemos estimar valores tão elevados, principalmente na parcela onde ocorre uma grande sequência de dados faltantes, o máximo é em torno de 7mm e a mediana em torno de 2mm. Estudar o comportamento de grandes parcelas com dados faltantes é um dos objetivos na continuidade da pesquisa.

Figura 1 – Estimação dos parâmetros: Média e intervalos de 95%



Fonte – Elaboração própria

Figura 2 – Séries das observações com as médias a posteriori do quantil 99%



Fonte – Elaboração própria

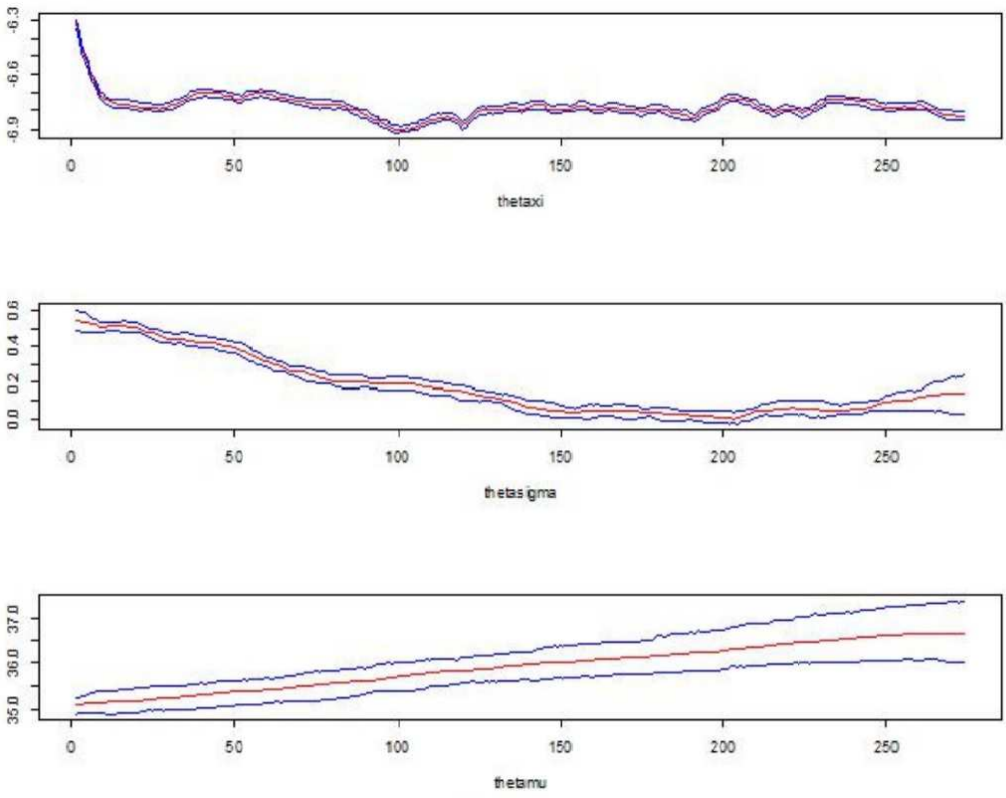
4.2 DADOS DE TEMPERATURA

A segunda aplicação consiste na análise de dados de temperatura máxima na cidade de Teresina, região Nordeste do Brasil, de 1993 a 2016. Para esta aplicação, analisamos as máximas mensais pela abordagem da distribuição GEV com dados faltantes conforme descrito na Seção 3.2. Amostramos aleatoriamente 10 pontos na série original de máximos, ocultamos esses pontos na modelagem e depois estimamos esses 10 pontos para verificar se esse método é eficiente para prever valores faltantes. A Figura 3 mostra as distribuições posteriores dos parâmetros, juntamente com o intervalo de credibilidade de 95%. Podemos observar uma tendência de crescimento do parâmetro relacionado μ no tempo, enquanto os demais parâmetros também variam, tendo períodos de altos e baixos.

A Figura 4 ilustra a série incompleta com os valores estimados em vermelho e os valores reais da série destacados em preto. É possível perceber

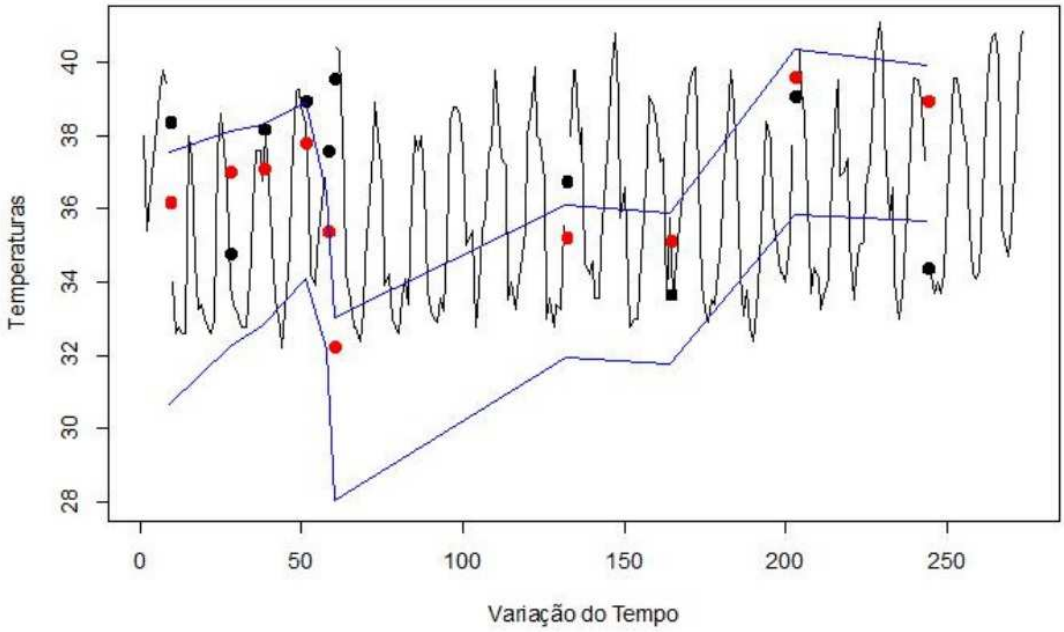
que a estimativa foi precisa em sua maioria, sendo as temperaturas estimadas muito próximas das reais.

Figura 3 – Estimação dos parâmetros dinâmicos: Média e intervalos de 95%



Fonte – Elaboração própria

Figura 4 – Série histórica máxima mensal de Teresina, com valores faltantes (preto) e valores estimados pelo modelo (vermelho). Linhas azuis: intervalo de credibilidade dos dados faltantes.



Fonte – Elaboração própria

5. CONCLUSÕES

Este trabalho apresentou uma forma de tratar dados faltantes no contexto de valores extremos. Foi feita uma abordagem integral, tanto analisando os máximos (GEV) quanto os excessos (GPD), considerando um modelo base onde os parâmetros das distribuições variam no tempo de acordo com um DLM. Os resultados das aplicações de dados ambientais mostraram que o método é eficiente na estimativa de valores faltantes, mas o método tem maior dificuldade em estimar corretamente quantis elevados quando há uma gama muito grande de dados faltantes, como foi o caso dos dados de precipitação.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Coles, S. G. (2001). An Introduction to Statistical Modelling of Extreme Values. Springer.

Gamerman, D. & Lopes, H. F. (2006). Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd edition). Baton Rouge: Chapman and Hall/CRC.

Behrens, C., Gamerman, D. & Lopes, H. F. (2004). Bayesian analysis of extreme events with threshold estimation. Statistical Modelling, 4, 227–244.

Cunnane, C. (1979). Note on the poisson assumption in partial duration series model. Water Resource Research, 15, 489–494.

Dinpashoh, Y., Fakheri-Fard, A. & Moghaddam, M. (2004). Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods. Journal of Hydrology, 297, 109–123.

Huerta, G. & Sanso, B. (2007). Time-varying models for extreme values.

Environmental and Ecological Statistics, 14(3), 285–299.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological events. Quarterly Journal of the Royal Meteorology Society, 81, 158–171.

Nascimento, F. F., Gamerman, D. & Lopes, H. F. (2012). A Semiparametric Bayesian approach to extreme value estimation. Statistics and Computing, 22, 661-675.

Nascimento, F. F., Lopes, H. F. & Gamerman, D. (2016). Time-varying extreme pattern with dynamic models, 25(1), 131-149.

Nunes, L. N., Kluck, M. M. & Fachel, J. (2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Caderno de Saúde Pública, Rio de Janeiro, 25(2), 268–278.

Parmesan, C., Root, T. L. & Willing, M. R. (2000). Impacts of Extreme Weather and Climate on Terrestrial Biota. Bulletin of the American Meteorological Society, 81, 443–450.

Pickands, J. (1975). Statistical inference using extreme order statistics. Annals of Statistics, 3, 119–131.

Scarrott, C. J. & Macdonald, A. (2012). A Review of extreme value threshold estimation and uncertainty quantification. REVSTAT 10 33–60.

von Mises, R. (1954). La distribution de la plus grande de n valeurs. American Mathematical Society, 2, 271–294.

West, M. & Harrison, J. (1997). Bayesian Forecasting and Dynamic Models. New York: Springer - Second Edition.

Wiper, M., Rios Insua, D. & Ruggeri, F. (2001). Mixtures of Gamma distributions with applications. Journal of Computational and Graphical Statistics, 10, 440–454.

Revista Brasileira de Estatística, V.81, N. 248, p.76-104, 2023

ESTIMAÇÃO DA DINÂMICA DA DESOCUPAÇÃO NO BRASIL COM MODELOS DE ESPAÇO DE ESTADOS PARA PESQUISAS AMOSTRAIS

Caio César Soares Gonçalves

ccsgonc@gmail.com

Fundação João Pinheiro e ENCE/IBGE

Luna Hidalgo

luna.hidalgo@ibge.gov.br

IBGE

Denise Britz do Nascimento Silva

denise.silva@ibge.gov.br

ENCE/IBGE

Resumo: Este trabalho examina a dinâmica da desocupação no Brasil com base em estimativas mensais, do total de desocupados do Brasil e de unidades da federação selecionadas, produzidas a partir da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC). O artigo utiliza a abordagem baseada em modelos de séries temporais univariados e bivariado sob a formulação de espaço de estados, incorporando o efeito que o desenho amostral da PNADC exerce sobre a autocorrelação das estimativas. O modelo bivariado integra a série de beneficiários do seguro-desemprego como informação auxiliar. Além disso, estimativas da primeira diferença foram obtidas para identificar se as variações mensais são estatisticamente significativas entre os anos de 2012 a 2020. Os resultados encontrados apontam que a série de total de beneficiários do seguro-desemprego não apresentou tendência comum com a série de total de desocupados. As divergências envolvem questões burocráticas, conceituais e de elegibilidade do seguro-desemprego, além de questões associadas ao nível de informalidade do mercado de trabalho brasileiro. Adicionalmente, as estimativas de primeira diferença mostraram-se significativas por longos períodos, evidenciando que as frequentes mudanças no país, sejam econômicas, políticas, ou até mesmo as ações para enfrentamento da pandemia da COVID-19, provocaram, e continuam provocando, alterações no mercado de trabalho.
Palavras-chave: Indicadores do mercado de trabalho; Seguro-desemprego; Modelos estruturais de séries temporais; Pesquisa amostral; Estimativas mensais.

Abstract: This paper examines the dynamics of unemployment in Brazil, and for selected states, based on monthly estimates of the total number of unemployed persons obtained from the Continuous National Household Sample Survey (PNADC). The paper exploits univariate and bivariate time series state-space models that take into account the effect of the survey sample design on the autocorrelation of the estimates. The bivariate model incorporates the series of

unemployment insurance benefits (claimant counts) as auxiliary information. In addition, estimates of the first difference were obtained to identify whether the monthly variations were statistically significant from 2012 to 2020. The results show that unemployment insurance benefits series did not show a common trend with the series of unemployed persons. The differences are due to bureaucratic, conceptual and unemployment insurance eligibility issues and other issues associated with the high level of informality in the Brazilian labour market. Additionally, the first difference estimates proved to be statistically significant for long periods, showing that the frequent changes in the country's, economic and political situation, or even the actions to face the COVID-19 pandemic have caused, and continue to influence the labour market dynamics.
Keywords: Labour market indicators; Claimant count; Structural time series models; Sample surveys; Monthly estimates.

1. INTRODUÇÃO

O acompanhamento contínuo de informações que retratam a situação socioeconômica de um país ou de uma região faz parte das análises de conjuntura. Quanto mais rápido um dado estiver disponível, melhor o entendimento do fenômeno naquele local e, consequentemente, ações podem ser planejadas e executadas pela administração pública e por outros agentes.

No passado recente, o Brasil enfrentou algumas crises que provocaram alterações inclusive em seus indicadores macroeconômicos, como no Produto Interno Bruto (PIB), na inflação e na taxa de desocupação.

Na última década, o PIB nacional registrou recessão e a inflação, embora tenha se mantido dentro do patamar estabelecido pela política de metas de inflação, com exceção dos anos de 2015 e 2017, apresentou oscilações e pressões diferenciadas no período. Na desocupação, um dos indicador-chave do mercado de trabalho, as alterações que vinham sendo notadas se agravaram em 2020 com a chegada da pandemia de COVID-19 no Brasil. Com isso, a demanda por informações conjunturais mensais se intensificou, em especial às relacionadas com o mercado laboral.

Desde março de 2016, a Pesquisa Nacional por Amostra de Domicílios Continua – PNADC, realizada pelo Instituto Brasileiro de Geografia e Estatística – IBGE, é a fonte oficial dos indicadores de mercado de trabalho. Mensalmente, a partir de dados acumulados de três meses consecutivos, divulga resultados nacionais, e trimestralmente divulga resultados regionais, para grandes regiões (GRs) e unidades da federação (UFs). "Assim, os indicadores da PNAD Contínua produzidos mensalmente não refletem a situação de cada mês, mas, sim, a situação do trimestre móvel que finaliza a cada mês" (IBGE, 2022a).

De forma a suprir a ausência de indicadores mensais não providos pelo IBGE, o Banco Central do Brasil (Bacen) e o Instituto de Pesquisa Econômica Aplicada (Ipea) realizaram alguns estudos específicos nessa linha. No estudo do

Bacen, foi usado um modelo de espaço de estados para construção da série temporal mensal de média móvel trimestral de alguns indicadores do mercado de trabalho da PNADC (Banco Central do Brasil, 2020). No caso do IPEA, Hecksher (2020) produziu estimativas mensais do mercado de trabalho baseadas no desenho amostral da pesquisa, sem fazer uso de modelagem estatística, utilizando um procedimento para identificar indiretamente nos microdados públicos da pesquisa o mês de realização da entrevista.

Com a substituição da Pesquisa Mensal de Emprego (PME) em 2016 pela PNADC, os indicadores conjunturais do país passaram a ser produzidos para todas as UFs[1], a partir de uma amostra maior e mais espalhada pelo território nacional. Por outro lado, as estimativas mensais passaram a ser calculadas para trimestres, pois Lila & Freitas (2007) demostraram que as estimativas de variação mensal (diferenças mês a mês) não eram estatisticamente significativas e, portanto, indicadores trimestrais seriam suficientes para o acompanhamento conjuntural do mercado de trabalho brasileiro. Entretanto, tal estudo foi elaborado com dados da PME analisando resultados de pouco mais de um ano (janeiro de 2002 a abril de 2003).

Diante desse contexto, esse trabalho analisa a dinâmica da desocupação por meio do cálculo de estimativas mensais para o total de desocupados do Brasil e de duas UFs (com diferentes tamanhos de amostra), utilizando a abordagem baseada em modelos de séries temporais univariados que consideram na sua formulação características do desenho amostral da pesquisa, como apresentado por Brakel & Krieg (2009) para a Holanda e Elliott & Zong (2019) para o Reino Unido, e modelos bivariados como empregados por Harvey & Chung (2000) incorporando informações do seguro-desemprego.

Além disso, estimativas de primeira diferença são calculadas para identificar se as variações mensais são estatisticamente diferentes de zero durante o período de análise de 2012 a 2020, evidenciando as alterações da dinâmica da desocupação no Brasil. Alguns dos estudos que abarcam as estimativas de variação dos totais de pessoas desocupadas baseadas em modelos de espaço de estados são Harvey & Chung (2000) e Boonstra & Brakel (2019).

O registro administrativo do número total de beneficiários mensal do seguro-desemprego (Ministério do Trabalho e Previdência, 2021) é utilizado para avaliar a possibilidade de aprimorar a precisão das estimativas do total de desocupados por meio de um modelo bivariado de tendências comuns. Sob essa

[1] A PME era realizada em apenas seis regiões metropolitanas (Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo e Porto Alegre) (IBGE, 2007).

ótica de integração de dados, Harvey & Chung (2000) exploraram a possibilidade de modelar séries com trajetórias comuns do total de desocupados com o número total de beneficiários do seguro fornecido pelo Reino Unido e encontrou resultados a favor do modelo bivariado. Schiavoni *et al.* (2021) testaram a mesma estratégia para o caso holandês. Gonçalves *et al.* (2022) apontaram que não foi possível encontrar vantagens em termos de precisão do uso de modelos bivariados com o total de beneficiários do seguro-desemprego, porém o estudo foi realizado considerando como variável de interesse a taxa de desocupação, o que leva ao questionamento se o mesmo resultado seria encontrado na estimação baseada em modelos para o total de desocupados.

Na seção 2 desse artigo são apresentadas a evolução do total de desocupados no Brasil e do total de parcelas pagas de seguro-desemprego mensal. Na seção 3, abordam-se os modelos de espaço de estados em suas formulações univariada e bivariada, além de apresentar os estimadores de diferenças e procedimentos de estimação. Os resultados são discutidos na seção 4 a partir do modelo com melhor precisão na comparação entre modelos univariados e bivariados e considerando os resultados da primeira diferença do componente de tendência. Por fim, a seção 5 contém as considerações finais.

2. DADOS DE DESOCUPAÇÃO NO BRASIL

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) é a maior pesquisa amostral realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), abrangendo aproximadamente 70 mil domicílios mensalmente. Criada com o objetivo de permitir o acompanhamento das estatísticas trimestrais e da evolução da força de trabalho no curto, médio e longo prazos, a PNADC também incorpora outras temáticas relevantes (IBGE, 2022b).

A PNADC faz parte do Sistema Integrado de Pesquisas Domiciliares (SIPD), que foi estabelecido pelo IBGE para garantir uma abordagem coordenada, gerencial e conceitualmente harmonizada nas pesquisas amostrais domiciliares. A amostra da PNADC é selecionada a partir de uma amostra mestra, evitando assim a ocorrência de domicílios comuns em mais de uma pesquisa do Sistema Integrado. Essa amostra mestra foi construída com base no Censo Demográfico de 2010 e leva em consideração as atualizações da Base Operacional Geográfica (BOG) e do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) (IBGE, 2022b).

A PNADC, iniciada em caráter experimental em outubro de 2011, entrou em produção definitiva em janeiro de 2012, com as primeiras publicações ocorrendo em 2014. Ela substituiu a Pesquisa Mensal de Emprego (PME) e a

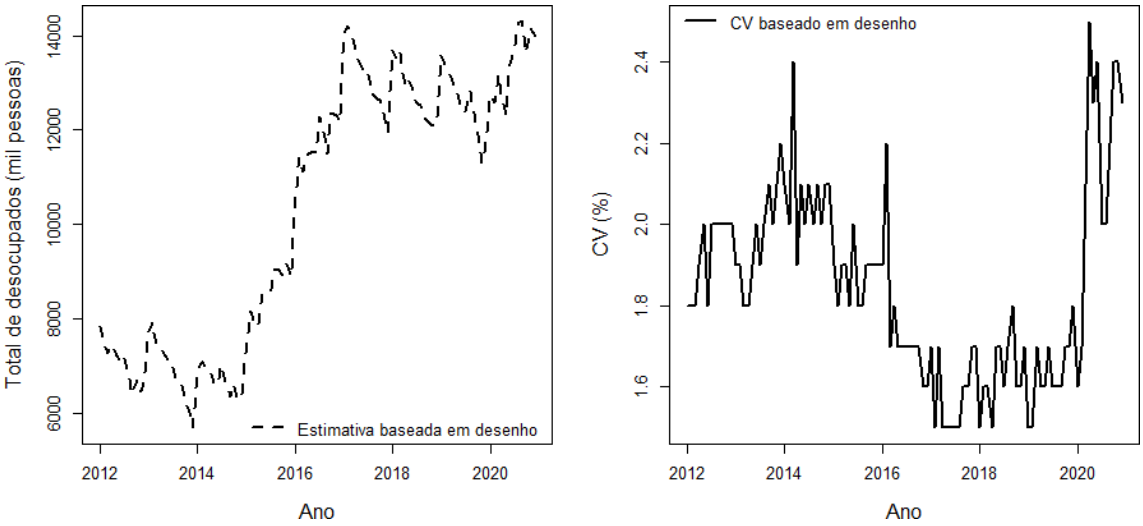
Pesquisa Nacional por Amostra de Domicílios (PNAD). A substituição da PME pela PNADC marcou a ruptura das séries temporais relacionadas ao mercado de trabalho. Completando dez anos em 2022, a PNADC já possibilita o retorno aos estudos de tendências, incluindo a aplicação de métodos de análise de séries temporais (IBGE, 2022b).

A amostra da PNADC é uma amostra probabilística de múltiplos estágios de domicílios, com um plano amostral conglomerado em dois estágios de seleção. No primeiro estágio, as unidades primárias de amostragem (UPAs), que são usualmente os setores censitários, são selecionadas com uma probabilidade proporcional ao número de domicílios em cada estrato definido. A definição dos estratos é feita no âmbito do Sistema Integrado de Pesquisas Domiciliares (SIPD). No segundo estágio, são selecionados 14 domicílios dentro de cada UPA da amostra usando uma amostragem aleatória simples, com base na lista atualizada do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE). Além disso, a PNADC utiliza um esquema de rotação da amostra conhecido como 1-2(5) (IBGE, 2022b).

Ao utilizar exclusivamente a amostra coletada para cada mês de referência, são obtidas estimativas baseadas em desenho amostral (conhecidas como estimativas *design-based* em inglês) ou, como também referenciado na literatura, estimativas diretas. Segundo IBGE (2014a, p. 34), o estimador de total divulgado pelo órgão é obtido pela soma da variável de interesse para cada registro, levando em consideração o peso amostral correspondente. Isso quer dizer que os mesmos procedimentos adotados para as estimativas trimestrais ou anuais pelo IBGE foram aplicados para calcular as estimativas mensais diretas apresentadas nesse artigo.

A Figura 1 ilustra a evolução do total de desocupados no Brasil, juntamente com o coeficiente de variação dessas estimativas.

Figura 1 – Estimativas mensais baseadas no desenho amostral do total de desocupados e respectivos coeficientes de variação - Brasil - jan. 2012 a dez. 2020



Fonte - Elaboração própria com base nos dados básicos PNADC/IBGE

Para contextualizar o quadro econômico brasileiro no início da série temporal produzida pela PNADC, pode-se considerar três indicadores: PIB, inflação e desocupação. Em meados de 2012, o Brasil encontrava-se em um status de arrefecimento da economia, tendo o PIB crescido 7,5% e 4,0% em 2010 e 2011, respectivamente, e 1,9% em 2012. No ano de 2013, o crescimento foi de 3,0%, porém nos anos seguintes a situação econômica se agravou com o indicador aumentando apenas 0,5% em 2014 e registrou recessão em 2015 e 2016 com taxas -3,5% e -3,3%, respectivamente, recuperando-se relativamente em 2017 com crescimento de 1,3% e em 2018 com 1,8% (IBGE, 2020).

Em relação ao nível de preços, a inflação medida pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA) foi 5,8% em 2012, sendo que apresentou valores dentro do intervalo das metas de inflação em todo o período de análise, com exceção de 2015 quando a inflação alcançou 10,7% naquele ano (IBGE, 2021).

De acordo com a Figura 1, o total de desocupados em dezembro de 2012 correspondia a 6,7 milhões de pessoas, permanecendo em torno desse patamar até o agravamento da crise política e econômica em 2015. Uma escalada do total de desocupados começou no mês de janeiro de 2015 e atingiu um pico no mês de fevereiro de 2017, com 14,195 milhões de desocupados. Esse novo nível alcançado pela série permaneceu até o mais recente agravamento devido aos efeitos da pandemia de COVID-19 na economia nacional.

Com medidas de restrição de circulação adotadas em diferentes UFs a partir de março de 2020, a desocupação avançou até a marca de 14,296 milhões de desocupados em setembro de 2020. Ressalta-se que existem elementos distintos nesse aumento da desocupação dado que, com a mobilidade urbana restrita, o movimento de busca por trabalho pode ter se reduzido, o que indica a relevância de analisar a situação do conjunto de pessoas fora da força de trabalho para complementar a análise do fenômeno da desocupação nesse período, o que não será abordado nesse estudo.

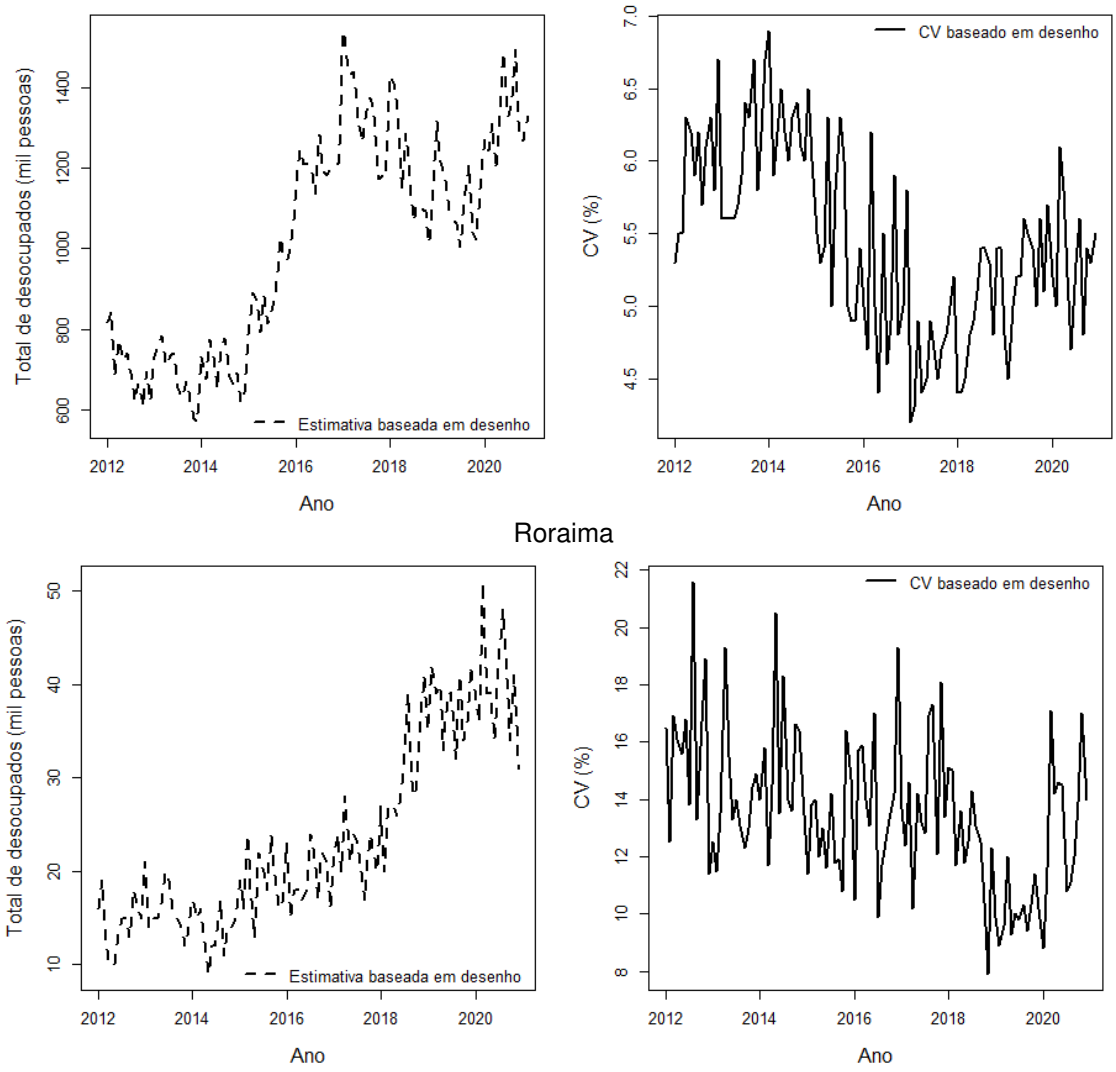
A Figura 1 sugere uma trajetória ascendente no final da série temporal. No entanto, é necessário estimar a tendência dessa série, inclusive removendo o componente sazonal para melhor compreendê-la. Esses componentes serão analisados posteriormente.

Em termos da precisão dessas estimativas mensais que não são publicadas pelo IBGE, observou-se que o nível de 1,6% de coeficiente de variação no período final da série (Figura 1) torna as estimativas adequadas para publicação, mesmo com os efeitos do aumento da imprecisão em 2020 devido à pandemia de COVID-19. Essa avaliação de adequação segue as recomendações feitas pelo instituto de que estimativas com CV de até 15% são passíveis de publicação (IBGE, 2014b). No entanto, sabe-se que, para as análises socioeconômicas do país, um olhar regionalizado é necessário devido às diferenças de comportamento das principais variáveis econômicas e sociais ao longo do território brasileiro. Assim, nesse artigo ilustra-se com os resultados de duas UFs, escolhidas por suas características distintas de tamanho de amostra: Minas Gerais (amostra grande) e Roraima (amostra pequena).

A Figura 2 apresenta as estimativas mensais baseadas no desenho amostral para o total de desocupados de Minas Gerais e Roraima, bem como o comportamento ao longo do tempo do coeficiente de variação dessas estimativas. A trajetória para o caso de Minas Gerais é bastante similar ao descrito para o caso do Brasil na Figura 1, porém com uma trajetória de queda mais acentuada após 2017 e níveis mais altos de coeficiente de variação.

No caso de Roraima, a Figura 2 mostra uma trajetória diferente da nacional para o total de desocupados nesse estado. A transição entre dois níveis visualizados no Brasil e em Minas Gerais apresenta-se mais suavizada para o caso de Roraima, além dessa série apresentar um comportamento mais errático que as anteriores.

Figura 2 – Estimativas mensais baseadas no desenho amostral do total de desocupados e respectivos coeficiente de variação - Minas Gerais e Roraima - jan. 2012 a dez. 2020



Fonte - Elaboração própria com base nos dados básicos PNADC/IBGE

Em termos da precisão, o coeficiente de variação das estimativas do total de desocupados de Roraima foi mais alto comparativamente aos obtidos para as estimativas do Brasil e Minas Gerais, e até mesmo superior a 15% em vários pontos do tempo, o que revela a necessidade de um uso mais cauteloso dessas estimativas diretas, caso fossem divulgadas. Assim, como esperado para pequenas amostras, a imprecisão se eleva. Porém, conforme mencionado anteriormente, a abordagem baseada em modelos possui o potencial de produzir estimativas mensais não só para o Brasil, mas também para as UFs, mais precisas que essas estimativas diretas baseadas no desenho amostral (apresentadas pela Figura 2), bem como permite estimar seu componente de tendência e possibilitar a avaliação de mudanças de curto prazo ocorridas nas séries.

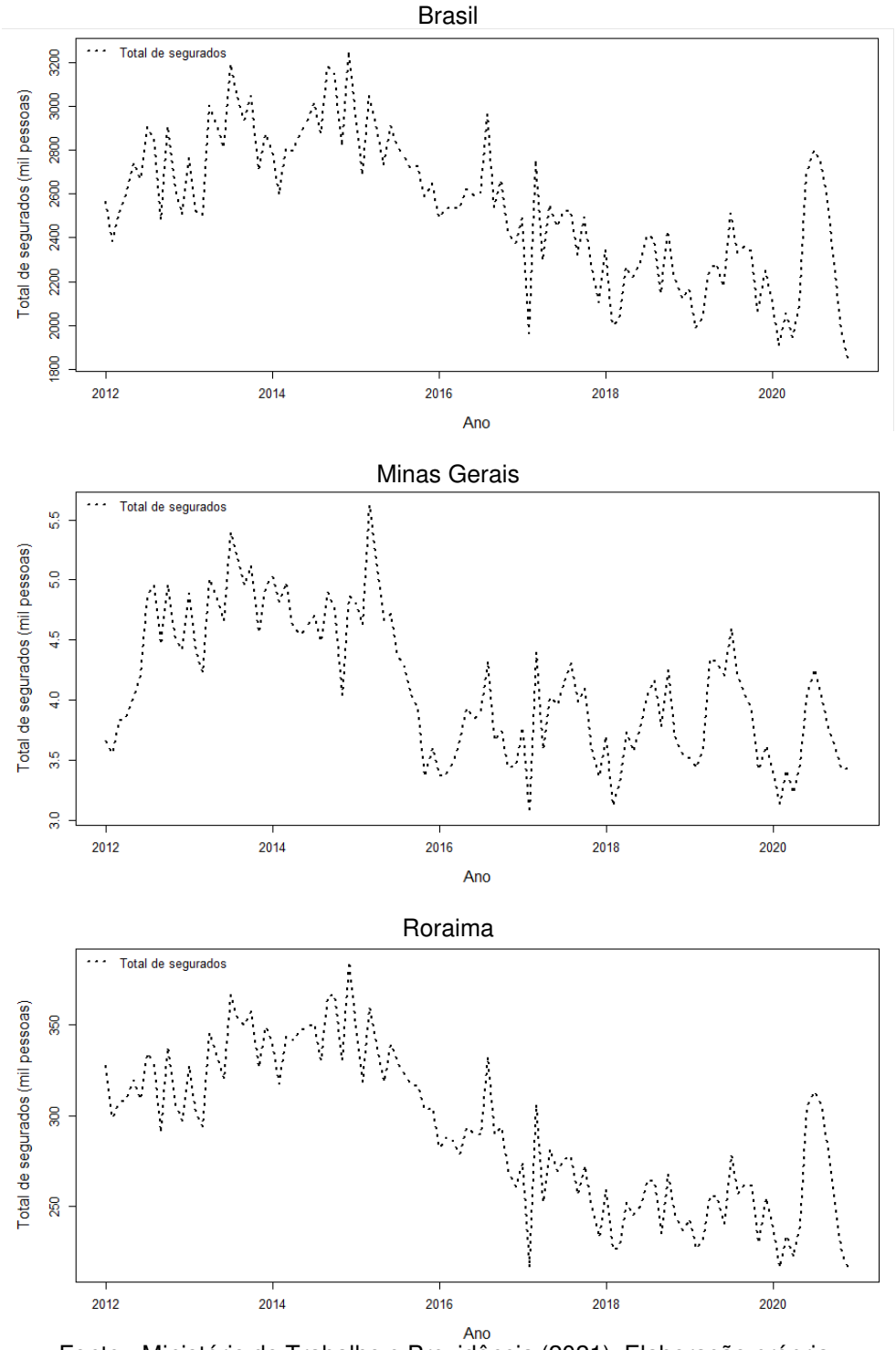
2.1 Seguro-desemprego

Uma das políticas públicas de emprego e renda existentes já consolidadas e abrangentes no país é o seguro-desemprego. Tal política fornece uma transferência monetária temporária para trabalhadores formais dispensados sem justa causa do trabalho e de acordo com pré-requisitos determinados. Conforme Ramos (2003), o seguro-desemprego se classifica como uma política de emprego passiva, no sentido de buscar amenizar a condição de desempregado.

As solicitações são realizadas pelo requerente, que tem seu benefício analisado bem como a especificação do número de parcelas a receber. Portanto, em termos comparativos com o número total de desocupados pode existir um "descompasso burocrático" com o número de beneficiários, dado que existe um intervalo de tempo entre o momento que a pessoa foi dispensada do emprego até o recebimento da primeira parcela do seguro-desemprego. Conforme as regras em vigor, o requerente pode iniciar o processo a partir do sétimo dia após a data da demissão e dentro de até 120 dias. Além disso, a pessoa não será considerada desocupada para a PNADC caso não busque um trabalho nesse período de recebimento - o que pode revelar também um "descompasso conceitual" com a série de total de desocupados, dado que nem todos aqueles que estão incluídos no total de segurados necessariamente estão desocupados. Ademais, podem estar ocupados em um trabalho informal ao mesmo tempo que recebem o benefício, e, portanto, serão contabilizados pelo IBGE como ocupados.

Os dados do seguro-desemprego são divulgados pelo Ministério do Trabalho e Previdência com periodicidade mensal. A Figura 3 apresenta o número total de parcelas (benefícios) pagas(os) por mês, o que representa o número total de pessoas seguradas em determinado mês. Observa-se que o comportamento da série do total de segurados é distinto do apresentado pela série do total de desocupados com uma relevante observação sobre a mudança do nível na direção inversa ao indicado nas Figuras 1 e 2. O comportamento esperado seria que, com o agravamento da crise econômica e política no país em 2015 a 2017, o número de requerentes e, consequentemente, o número de benefícios pagos por mês pelo seguro-desemprego aumentariam. No entanto, observou-se o comportamento contrário nesse período.

Figura 3 – Total de segurados com o seguro-desemprego - Brasil, Minas Gerais e Roraima – jan. 2012 a dez. 2020



Fonte - Ministério do Trabalho e Previdência (2021). Elaboração própria

Por se tratar de uma política pública, alterações em seu escopo podem influenciar no comportamento da série descolando do comportamento observado no total de desocupados estimado pela PNADC. A política de seguro-desemprego sofreu alterações em 2015. Em uma tentativa de cortes de gastos

como medida de política fiscal, o governo reformulou o programa do seguro-desemprego com a edição da Medida Provisória no. 665/2014, que posteriormente foi convertida na Lei no. 13.134/2015. Nesse sentido, não se apresentou como uma modificação dentro das discussões de eficiência e eficácia de uma política, mas sim uma alteração pautada em controle fiscal (Ipea, 2016). A Tabela 1 resume as alterações ocorridas ilustrando as modificações para o caso de primeira solicitação do benefício e o quão restritivas foram as novas regras. Por exemplo, retirou-se a possibilidade de acesso ao seguro-desemprego a trabalhadores com menos de um ano no trabalho formal.

Tabela 1 – Regras para acesso ao seguro-desemprego – 1ª solicitação

Legislação (Vigência)	Meses trabalhados	Número de parcelas do benefício
Lei nº 7.998 (Até Fev./2015)	6 a 11 meses	3
	12 a 23 meses	4
	24 meses ou mais	5
MP nº 665 (Março a Maio/2015)	18 a 23 meses	4
	24 meses ou mais	5
Lei nº 13.134 - (Após Jun./2015)	12 a 23 meses	4
	24 meses ou mais	5

Fonte - Adaptado do Banco Central do Brasil (2016, p. 110)

Como um grupo de pessoas passaram a não ser aptas ao seguro-desemprego, o Banco Central do Brasil (2016) apontou que ocorreu uma elevação da oferta de mão de obra devido às mudanças das regras de concessão do seguro-desemprego, porém esse efeito seria relativamente pequeno, entre 0,1 p.p. a 0,2 p.p. acrescidos na taxa de desocupação. De forma que a elevação da desocupação teria se dado por outros fatores, como a adoção das políticas de ajuste econômico que levaram à recessão, conforme ressaltou Pochmann (2015).

Com isso, existe a necessidade de modelar essa mudança de nível na série de total de segurados ocorrida devido a alteração da política, já que foi uma modificação das pessoas elegíveis. Ainda assim, vale ressaltar que existe uma diferença considerável das séries de total de desocupados e de total de segurados. Esse "descompasso de elegibilidade" é decorrente das regras dessa política, dentre as quais se destacam: assinatura da carteira e comprovações referentes ao tempo de contribuição e meses trabalhados no mercado formal, entre outros aspectos levados em consideração no cálculo do benefício. Dessa forma, nem todas as pessoas desocupadas estão aptas a receber o benefício. Enquanto no Brasil como um todo existiam 13,4 milhões de desocupados em junho de 2020, apenas 2,7 milhões de pessoas receberam o benefício nesse mês, conforme Figuras 1, 2 e 3. A magnitude da discrepância evidencia os efeitos da informalidade no país que, nesse caso específico, apresenta-se com um dos impeditivos de requerer o seguro-desemprego. As relações existentes entre seguro-desemprego e informalidade são exploradas no estudo de Mourão *et al.* (2013).

3. MODELOS ESTRUTURAIS

O modelo para a série $\hat{y}_t$ de estimativas mensais do total de desocupados baseadas no desenho amostral pode ser escrito no formato de espaço de estados com base em Harvey (1989) e Durbin & Koopman (2012):

$$\hat{y}_t = H_t \boldsymbol{\alpha}_t + I_t, \qquad I_t \sim \mathcal{N}(0, \sigma_I^2) \tag{1}$$

$$\boldsymbol{\alpha}_t = G\boldsymbol{\alpha}_{t-1} + W\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(0, Q) \tag{2}$$

A Equação 1 é referenciada na literatura como equação de observação ou de medida e a Equação 2 é a equação de transição ou de estado. A determinação do modelo ocorre pelo comportamento ao longo do tempo do vetor de estados $\boldsymbol{\alpha}_t$ de dimensão $h\ x\ 1$ na Equação 2. Porém, como esse vetor não é observado diretamente, expressa-se uma relação do vetor de estados com o vetor de observações $\hat{y}_t$ de dimensão $T\ x\ 1$ da Equação 1. Assim, T é o total de observações e h é o total de variáveis no vetor de estado. As matrizes H_t, G e W são conhecidas e a matriz Q e o escalar σ_I^2 serão estimados. Todas as matrizes podem variar com o tempo, porém, para esse caso específico apenas a matriz H_t é variante no tempo. Assume-se que os termos de erros I_t e $\boldsymbol{\eta}_t$ são serialmente independentes e não dependentes entre si (Durbin & Koopman, 2012).

3.1 Modelo para o total de desocupados

Ao empregar um modelo estrutural para uma série temporal advinda de uma pesquisa amostral repetida, é necessário considerar a estrutura de correlação das observações imposta pelo desenho da pesquisa. Nesse sentido, constrói-se um modelo de extração de sinal, conforme abordado inicialmente por Scott & Smith (1974) e Scott *et al.* (1977), em que a série observada $\hat{y}_t$ é decomposta em seu verdadeiro valor $\boldsymbol{\theta}_t$ associado a um fenômeno e um distúrbio que representa o erro amostral e_t, conforme Equação 3. O parâmetro populacional não observável $\boldsymbol{\theta}_t$ representa os componentes clássicos de séries temporais: tendência (L_t), sazonalidade (S_t) e irregular (I_t) (Equação 4).

$$\hat{y}_t = \boldsymbol{\theta}_t + e_t \tag{3}$$

$$\boldsymbol{\theta}_t = L_t + S_t + I_t, \qquad I_t \sim \mathcal{N}(0, \sigma_I^2) \tag{4}$$

L_t representa um componente de tendência-ciclo em conjunto, dado que se trata de um modelo de curto prazo. Com intuito de apresentar esse componente com menores oscilações, optou-se por um modelo que inclui apenas um distúrbio na Equação 6 de inclinação ($\eta_{R,t}$). Esse modelo é referenciado na literatura como modelo de tendência suavizada. A diferença em relação ao modelo de tendência local é que a Equação 5 não inclui um distúrbio do nível da tendência ($\eta_{L,t}$) (Durbin & Koopman, 2012). Assim, a tendência depende do seu comportamento no período anterior mais um termo que representa a inclinação (R_t), que por sua vez é modelada por um passeio aleatório.

$$L_t = L_{t-1} + R_{t-1} \tag{5}$$

$$R_{t-1} = R_{t-1} + \eta_{R,t}, \qquad \eta_{R,t} \sim \mathcal{N}(0, \sigma_R^2) \qquad cov(\eta_{R,t}, \eta_{R,t'}) = 0, \forall t \neq t' \tag{6}$$

A sazonalidade é expressa em um formato trigonométrico, sendo que o somatório de seis frequências ($S_{\iota,t}$) gera o componente sazonal (S_t) conforme as Equações 7 a 10.

$$S_t = \sum_{\iota=1}^{\frac{S}{2}=6} S_{\iota,t} \tag{7}$$

$$\begin{pmatrix} S_{\iota,t} \\ S^*_{\iota,t} \end{pmatrix} = \begin{bmatrix} cos(\pi\iota/6) & sen(\pi\iota/6) \\ -sen(\pi\iota/6) & cos(\pi\iota/6) \end{bmatrix} \begin{pmatrix} S_{\iota,t-1} \\ S^*_{\iota,t-1} \end{pmatrix} + \begin{pmatrix} \eta_{S,\iota,t} \\ \eta^*_{S,\iota,t} \end{pmatrix} \tag{8}$$

$$\begin{pmatrix} \eta_{S,\iota,t} \\ \eta^*_{S,\iota,t} \end{pmatrix} \sim \mathcal{N}\left(0, \sigma_S^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \qquad onde \quad \iota = 1, \cdots, 6 \tag{9}$$

$$cov\left(\eta_{S,\iota,t}, \eta_{S,\iota,t'}\right) = 0, \;\; cov\left(\eta^*_{S,\iota,t}, \eta^*_{S,\iota,t'}\right) = 0, \qquad \forall t \neq t' \tag{10}$$

O erro amostral é expresso de forma proporcional ao erro padrão das estimativas baseadas no desenho amostral, conforme proposto por Binder & Dick (1989), o que resulta no erro amostral escalonado $\tilde{e}_t$ na Equação 11. Esse mecanismo permite capturar os efeitos relacionados à pesquisa, como as oscilações no tamanho da amostra. Portanto, além da própria série de interesse baseada no desenho amostral, a série temporal do desvio padrão das estimativas também é incorporada no modelo.

$$e_t = \hat{c}_t \tilde{e}_t, \qquad \hat{c}_t = \sqrt{var(\hat{y}_t)}, \;\; \tilde{e}_t = \hat{\phi}\tilde{e}_{t-3} + \eta_{\tilde{e},t}, \qquad \eta_{\tilde{e},t} \sim \mathcal{N}(0, \sigma^2_{\tilde{e}}) \tag{11}$$

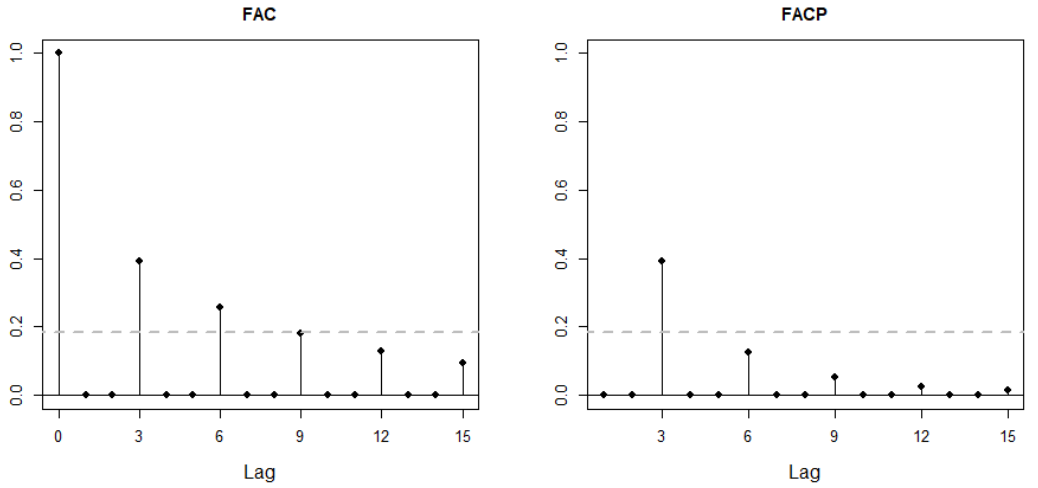
Outra informação exógena é o parâmetro do modelo que expressa a estrutura de correlação do erro amostral imposta pela sobreposição das amostras ao longo do tempo. A PNADC é uma pesquisa de painéis rotativos na qual o domicílio é entrevistado uma vez a cada trimestre, por 5 trimestres consecutivos (IBGE, 2022b). Sendo assim, não há sobreposição mensal da amostra da PNADC e, portanto, não há correlação de *lag* um e dois entre os erros amostrais, mas apenas aquela associada ao par de meses t e $t-3$. Conforme Brakel & Krieg (2009), o processo estocástico associado ao erro amostral pode ser modelado como um AR(3) com apenas um coeficiente diferente de zero. Essa identificação também é possível pelo cálculo das funções de autocorrelação e autocorrelação parcial. Para isso, as autocovariâncias dos erros amostrais foram calculadas pela sobreposição das unidades primárias de amostragem da PNADC (IBGE, 2022b), sendo que a estimativa de ϕ foi obtida pela resolução das equações de Yule-Walker. A Figura 4 ilustra as funções de autocorrelação (FAC) e autocorrelação parcial (FACP) para os erros amostrais das estimativas do total de desocupados do Brasil, considerando o período desde o início da série da PNADC em janeiro de 2012 a dezembro de 2020.

Os parâmetros estimados para a equação do erro amostral (ϕ) foram de 0,39 para o Brasil, 0,41 para o Roraima e 0,36 para Minas Gerais. Ressalta-se o comportamento de decrescimento da FAC e truncamento da FACP no *lag* três

ratificando a proposição de representar o processo do erro amostral com um modelo AR(3) em todos os casos.

Por fim, é importante ressaltar que o componente irregular e os distúrbios, nomeadamente I_t, $\eta_{R,t}$, $\eta_{S,t}$ e $\eta_{\tilde{e},t}$, são considerados mutuamente não correlacionados.

Figura 4 – Função de Autocorrelação (FAC), Função de Autocorrelação Parcial (FACP) e intervalos de confiança do processo do erro amostral das estimativas da PNADC do total de desocupados no Brasil



Fonte – Elaboração Própria

3.2 Modelo para o total de segurados

Para o caso da série de seguro-desemprego, representada por x_t (total de pessoas seguradas no mês t), não há o erro amostral e_t como na Equação 3, já que a série se refere a um registro administrativo e, consequentemente, a Equação 11 não se aplica.

$$x_t = L_{x,t} + S_{x,t} + I_{x,t}, \qquad I_{x,t} \sim \mathcal{N}\left(0, \sigma^2_{I,x}\right) \tag{12}$$

Além disso, é necessária a inclusão de uma variável de intervenção para captar a mudança de nível provocada pela alteração da política de seguro-desemprego como mencionada na seção 2.1. Conforme seção 7.6 de Harvey (1989), o efeito da intervenção pode ser incorporado como uma variável indicadora d_t do tipo pulso na equação de nível (Equação 13), em vez de inserida diretamente na equação de mensuração. Assim, d_t assume valor unitário apenas em março de 2015 e o valor zero para todos os demais meses. A inclusão dessa *dummy* de intervenção no modelo foi testada, bem como o mês para o seu valor unitário.

$$L_{x,t} = L_{x,t-1} + R_{x,t-1} + \lambda d_t \tag{13}$$

$$R_{x,t-1} = R_{x,t-1} + \eta_{R,x,t}, \qquad \eta_{R,x,t} \sim \mathcal{N}\left(0, \sigma^2_{R,x}\right). \quad cov\left(\eta_{R,x,t}, \eta_{R,x,t'}\right) = 0, \forall t \neq t' \tag{14}$$

A sazonalidade de x_t é modelada da mesma forma que a indicada pelas Equações 7 a 10, as Equações 15 a 18 representam a sazonalidade no formato trigonométrico:

$$S_{x,t} = \sum_{\iota=1}^{\frac{S}{2}=6} S_{x,\iota,t} \tag{15}$$

$$\begin{pmatrix} S_{x,\iota,t} \\ S^*_{x,\iota,t} \end{pmatrix} = \begin{bmatrix} cos(\pi\iota/6) & sen(\pi\iota/6) \\ -sen(\pi\iota/6) & cos(\pi\iota/6) \end{bmatrix} \begin{pmatrix} S_{x,\iota,t-1} \\ S^*_{x,\iota,t-1} \end{pmatrix} + \begin{pmatrix} \eta_{S,x,\iota,t} \\ \eta^*_{S,x,\iota,t} \end{pmatrix} \tag{16}$$

$$\begin{pmatrix} \eta_{S,x,\iota,t} \\ \eta^*_{S,x,\iota,t} \end{pmatrix} \sim \mathcal{N}\left(0, \sigma^2_{S,x} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \qquad onde \quad \iota = 1, \cdots, 6 \tag{17}$$

$$cov\left(\eta_{S,x,\iota,t}, \eta_{S,x,\iota,t'}\right) = 0, \quad cov\left(\eta^*_{S,x,\iota,t}, \eta^*_{S,x,\iota,t'}\right) = 0, \qquad \forall t \neq t' \tag{18}$$

3.3 Modelo bivariado

O modelo univariado pode ser generalizado para incluir várias séries temporais. No caso específico, há o interesse em estimar conjuntamente as séries de total de desocupados ($\hat{y}_t$) e o número pessoas beneficiadas com o seguro-desemprego (x_t) por mês t. Portanto, esse modelo bivariado pode ser expresso como:

$$z_t = H_t \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, V) \tag{19}$$

$$\boldsymbol{\alpha}_t = G\boldsymbol{\alpha}_{t-1} + W\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(0, Q), \quad t = 1, \cdots, T \tag{20}$$

e z_t pode ser escrito de acordo com a extração de sinal na presença de ruído por:

$$z_t = \begin{pmatrix} \hat{y}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \theta_{y,t} \\ \theta_{x,t} \end{pmatrix} + \begin{pmatrix} e_t \\ 0 \end{pmatrix} \tag{21}$$

Conforme já mencionado, x_t é um registro administrativo e, portanto, não possui erro amostral. Segundo Durbin & Koopman (2012), esse é um modelo para representar o processo de duas séries temporais aparentemente não correlacionadas em que cada série possui sua própria especificação do processo do sinal e erro amostral (não existente para x_t) independentes, porém os distúrbios dos componentes não observáveis, como da tendência, podem se correlacionar. Essa relação será incluída na matriz Q. A forma bivariada dessa matriz específica para o presente estudo, no qual se investiga a possibilidade de componentes de tendências comuns, tem a seguinte forma:

$$Q = \begin{pmatrix} \sigma^2_{S,\hat{y}} & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{S,x} & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{R,\hat{y}} & \rho^R_{x,\hat{y}}\sigma_{R,\hat{y}}\sigma_{R,x} & 0 \\ 0 & 0 & \rho^R_{x,\hat{y}}\sigma_{R,\hat{y}}\sigma_{R,x} & \sigma^2_{R,x} & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\tilde{e},\hat{y}} \end{pmatrix} \tag{22}$$

A covariância entre os distúrbios dos termos de erro das equações de inclinação das séries x_t e $\hat{y}_t$ é escrita em função da correlação linear e dos desvios.

3.4 Estimadores de diferença

Dois principais resultados do processo de modelagem são as estimativas dos componentes não observáveis, a saber: as séries da tendência e do sinal. Destaca-se que, sob essa abordagem, o sinal é obtido a partir da soma das séries de tendência e sazonalidade, após extrair os dois componentes de erro (o componente irregular e o erro amostral).

Adicionalmente, como na análise de conjuntura econômica, a estimação da variação de indicadores (entre meses ou trimestres) é também relevante, decidiu-se apresentar estimadores de diferença, especificamente da tendência, para responder sobre sua significância estatística no curto prazo. A diferença da tendência em relação ao mês anterior será expressa como $\Delta_t = L_t - L_{t-1} = R_{t-1}$ conforme Equação 5. Os intervalos de confiança para os parâmetros, com base nos estimadores Δ_t, permitem uma avaliação da das mudanças mensais diferentes de zero do total de desocupados no Brasil e UFs selecionadas.

3.5 Procedimentos de estimação

Considerando a característica das séries temporais de não possuírem observações independentes, a função de verossimilhança pode ser expressa como a função de densidade condicional, conforme mostrado a seguir:

$$L(y;\boldsymbol{\psi}) = \prod_{t=1}^{T} p(y_t|Y_{t-1}) \qquad (23)$$

onde $p(y_t|Y_{t-1})$ representa a distribuição condicional de y_t dado o conjunto de informações até o tempo $t-1$, ou seja, $Y_{t-1} = \{y_{t-1}, y_{t-1}, \dots, y_1\}$. Os estimadores de máxima verossimilhança são obtidos maximizando essa função em relação ao vetor de hiperparâmetros $\boldsymbol{\psi}$ (Harvey; 1989, p. 125).

No caso do modelo de séries temporais de pesquisas repetidas apresentado na seção 3.1, é assumido que os distúrbios (I_t e $\boldsymbol{\eta}_t$) e o vetor de estado inicial ($\boldsymbol{\alpha}_0$) são normalmente distribuídos. Utiliza-se o resultado da distribuição normal multivariada para calcular recursivamente a distribuição de $\boldsymbol{\alpha}_t$ condicionada as observações disponíveis para o tempo t. Por sua vez, essas distribuições condicionais possuem distribuição normal com média e covariâncias completamente especificadas, sendo elas computadas pelo filtro de Kalman.

O filtro de Kalman é um algoritmo de estimação recursiva que calcula o estimador ótimo do vetor de estado no tempo t com base nas informações disponíveis até o tempo t, ou seja, o conjunto de observações $\hat{y}_t$ (estimativas da abordagem baseada no desenho amostral) dado que as matrizes do sistema juntamente com $\widehat{\boldsymbol{\alpha}}_0$ e P_0 que representam a média do vetor de estado inicial $\boldsymbol{\alpha}_0$ e a sua respectiva matriz de covariância P_0 são consideradas conhecidas e não incluídas explicitamente no conjunto de informações (Harvey; 1989, p. 104). Foram produzidas estimativas para as séries filtradas nas quais as estimativas para o mês t são calculadas com as informações obtidas até o tempo t, e também as séries suavizadas que utilizam, para obtenção de estimativas para o mês t, toda a informação disponível até o final da série (tempo T). Para mais detalhes sobre o filtro de Kalman ver Harvey (1989).

Entre um dos resultados do filtro, tem-se que $\boldsymbol{\alpha}_t$ condicionado a Y_{t-1} possui distribuição normal com média $\boldsymbol{a}_{t|t-1}$ e matriz de covariância $P_{t|t-1}$. A equação de mensuração (Equação 1) pode ser reescrita como:

$$\hat{y}_t = H_t\boldsymbol{a}_{t|t-1} + H_t(\boldsymbol{\alpha}_t - \boldsymbol{a}_{t|t-1}) + I_t, \tag{24}$$

sendo a distribuição condicional de $\hat{y}_t$ normal com média $\tilde{y}_{t|t-1} = H_t\boldsymbol{a}_{t|t-1}$ e variância $F_t = H_tP_{t|t-1}H_t' + \sigma_I^2$. Consequentemente, a Equação 23 pode ser reescrita como:

$$\log L = -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|F_t| - \frac{1}{2}\sum_{t=1}^{T} v_tF_t^{-1}v_t' \tag{25}$$

onde $v_t = \hat{y}_t - \tilde{y}_{t|t-1}, t = 1, \ldots, T$ é a inovação ou também denominada erros de predição um passo à frente.

Testes de diagnósticos foram utilizados para checar as hipóteses de normalidade desses resíduos padronizados, especificamente o teste de Shapiro-Wilk (função *shapiro.test* do R base - R Core Team (2020)), que avalia a semelhança com a distribuição normal sob a hipótese nula de que os resíduos possuem distribuição normal. . Para a verificação de evidências de correlação serial sob a hipótese nula de que os resíduos são independentemente distribuídos foi empregado o teste de Ljung-Box (função *Box.test* do R base) que utiliza os valores do correlograma da série. Por fim, para averiguar possível heterocedasticidade dos resíduos, optou-se pelo teste H (construído de maneira independente no R a partir de Durbin & Koopman (2012, p. 39)) que compara a soma dos quadrados dos resíduos de dois subconjuntos exclusivos e sob a hipótese nula de homoscedasticidade. Para a comparação dos modelos, e escolha da melhor especificação, foram realizados testes da razão de

verossimilhanças e os cálculos dos critérios AIC e BIC (Durbin & Koopman, 2012).

Com as informações obtidas da abordagem baseada no desenho amostral $\hat{y}_t$, $\hat{c}_t$ e $\hat{\phi}$ e da série de registro administrativo x_t, usou-se uma rotina de otimização da função de verossimilhança do modelo para estimar os hiperparâmetros σ_I^2, σ_R^2, σ_S^2, $\sigma_{\tilde{e}}^2$ e $\rho_{x,\hat{y}}$ para os modelos univariados do total de desocupados e do total de segurados, e também para o caso bivariado. Foram aproveitadas as funções disponíveis (*dlmMLE* e *dlmLL*) no pacote dlm (Petris, 2010). O mesmo pacote possui funções para o filtro de Kalman (*dlmFilter* e *dlmSmooth*), utilizado para obtenção das estimativas dos componentes não observáveis, de suas variâncias e das covariâncias (*dlmSvd2var*) entre os termos do vetor de estados que são necessários para o cálculo da precisão das estimativas, por sua vez, expressas pelos coeficientes de variação e intervalos de confiança.

4 ANÁLISE DOS RESULTADOS E DISCUSSÃO

Foram obtidos resultados dos modelos univariados e bivariados para Brasil, Minas Gerais e Roraima. O período de análise foi de janeiro de 2012 a dezembro de 2020, compreendendo um total de 108 observações. Os modelos estimados, conforme apresentado na seção 3, possuem tendência suavizada com sazonalidade estocástica modelada na forma trigonométrica e componente irregular estocástico. No entanto, os hiperparâmetros $\sigma_{S,\hat{y}}^2$, $\sigma_{S,x}^2$, $\sigma_{I,\hat{y}}^2$ e $\sigma_{I,x}^2$ foram testados para análise da significância estatística por meio da estimação de modelos aninhados, empregando testes de razões de verossimilhanças. A Tabela 2 apresenta os hiperparâmetros estimados e os testes de diagnósticos das inovações padronizadas.

Os resultados da Tabela 2 referem-se aos modelos que apresentaram os menores valores de AIC e BIC e consideram os resultados dos testes de razões de verossimilhanças para definir a melhor especificação do modelo. Não foram encontradas evidências estatísticas de que a formulação incluindo intervenção (Equação 13) seja melhor do que modelo que não a considere. Portanto, os modelos selecionados foram sem intervenção.

Identificou-se que, para todos os casos, o componente sazonal pode ser considerado determinístico ($\sigma_{S,\hat{y}}^2 = 0$) para a série do total de desocupados. Para a série do total de beneficiários do seguro-desemprego isso também ocorre para os modelos do Brasil e Minas Gerais. Verificou-se também que não há evidência para rejeitar a hipótese de nulidade da variância do termo irregular para o caso

de Roraima, resultando em um modelo de tendência determinística ($\sigma^2_{R,\hat{y}} = \sigma^2_{I,\hat{y}} = 0$).

Tabela 2 – Hiperparâmetros estimados e testes de diagnósticos para os modelos univariados e bivariados

Hiperparâmetros	Brasil		Minas Gerais		Roraima	
	Univariado	Bivariado	Univariado	Bivariado	Univariado	Bivariado
$\hat{\sigma}^2_{R,\hat{y}}$	4.858,58	4.489,71	33,00	31,31	0,037	0,055
$\hat{\sigma}^2_{I,\hat{y}}$	4.422,83	4.301,32	1.476,34	1.434,79	-	-
$\hat{\sigma}^2_{\tilde{e},\hat{y}}$	1,08	1,11	0,57	0,60	0,987	0,967
$\hat{\sigma}^2_{S,x}$	-	-	-	-	0,001	0,001
$\hat{\sigma}^2_{R,x}$	14,54	15,21	0,42	0,35	0,003	0,003
$\hat{\sigma}^2_{I,x}$	22.272,06	22.508,81	276,14	279,79	0,045	0,045
$\hat{\rho}^R_{x,\hat{y}}$	-	-0,65	-	-0,51	-	0,73
Testes	Valor p					
$\rho^R_{x,\hat{y}} = 0$		0,26		0,24		0,05
Total de desocupados - $\hat{y}$	Valor p					
Shapiro-Wilk	0,55	0,66	0,48	0,50	0,41	0,53
Ljung-Box	0,82	0,82	0,43	0,43	0,54	0,66
H	0,04	0,04	0,12	0,11	0,35	0,29
Total de segurados - x	Valor p					
Shapiro-Wilk	0,06	0,01	0,06	0,01	0,07	0,15
Ljung-Box	0,22	0,17	0,17	0,17	0,01	0,01
H	0,25	0,25	0,21	0,2	0,97	0,98

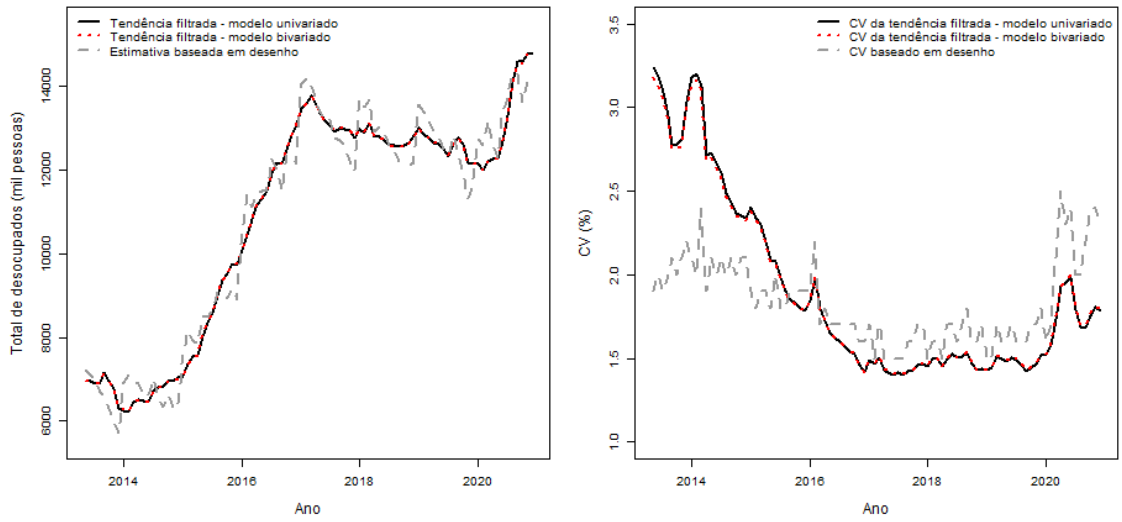
Fonte – Elaboração Própria

Os resultados dos testes dos resíduos padronizados revelaram que a normalidade, a ausência de autocorrelação e os erros homocedásticos foram comprovados ao nível de um por cento de significância, concluindo-se pela adequação do modelo. Ressalta-se que testes de sensibilidade de valores iniciais foram conduzidos para identificar a existência de outros máximos locais, sempre resultando na escolha da combinação de hiperparâmetros com o menor valor de log-verossimilhança.

É interessante ainda destacar na Tabela 2 os valores da correlação entre os distúrbios dos termos de erros das inclinações da série de total de desocupados e total de segurados. Para o Brasil e Minas Gerais, a correlação foi negativa e para Roraima, positiva. Essas correlações foram testadas com testes de razão de verossimilhanças em que o modelo escolhido foi estimado considerando a correlação $\rho^R_{x,\hat{y}} = 0$. Nesse caso, essa hipótese não foi rejeitada, ou seja, as correlações estimadas não são estatisticamente significativas, de forma que o modelo univariado apresenta a melhor especificação. Essa conclusão foi a mesma tanto para o Brasil quanto para os estados de Minas Gerais e Roraima.

Esse é um resultado relevante, confirmando o que já havia sido apontado em Gonçalves *et al.* (2022) ao estudar a integração de dados de seguro-desemprego com a taxa de desocupação. Assim, não há evidências que essas séries temporais possuam tendência comum. Em termos das precisões, os erros-padrão das estimativas de tendência suavizadas são apresentados na Figura 5 tanto para o modelo univariado quanto para o bivariado, ilustrando que ambos os modelos produzem resultados convergentes, mas o modelo bivariado não produz estimativas mais precisas. Essas mesmas conclusões foram encontradas para os casos de Minas Gerais e Roraima. Assim, o modelo final escolhido é a versão univariada cujas estimativas dos hiperparâmetros foram apresentadas na Tabela 2.

Figura 5 – Estimativas mensais baseadas no desenho amostral e em modelo do total de desocupados e coeficiente de variação – Brasil



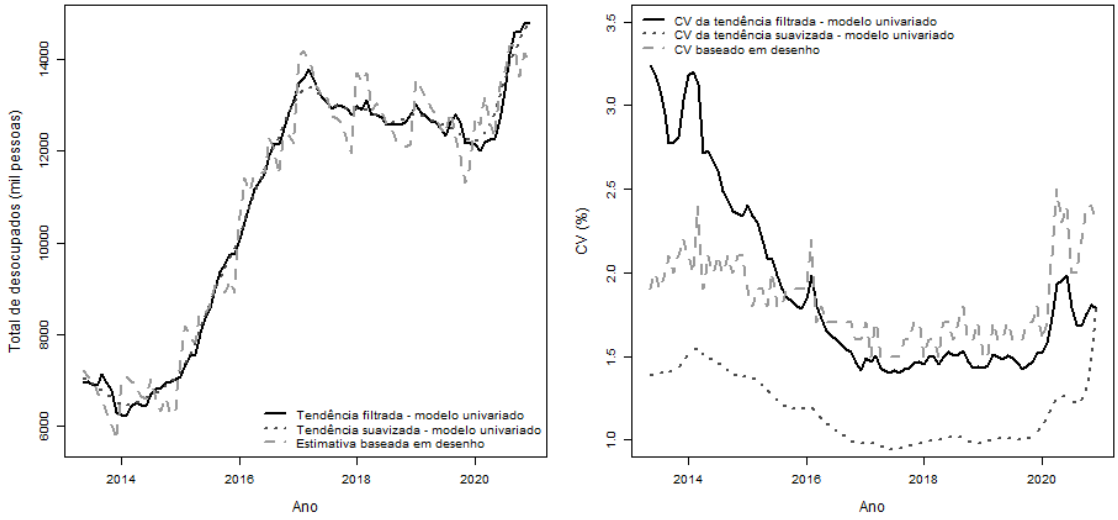
Fonte – Elaboração Própria

A Figura 6 exibe as tendências para o total de desocupados do Brasil, Minas Gerais e Roraima comparando com o valor das estimativas baseadas no desenho amostral, além de ilustrar as diferenças de precisão com os respectivos coeficientes de variação.

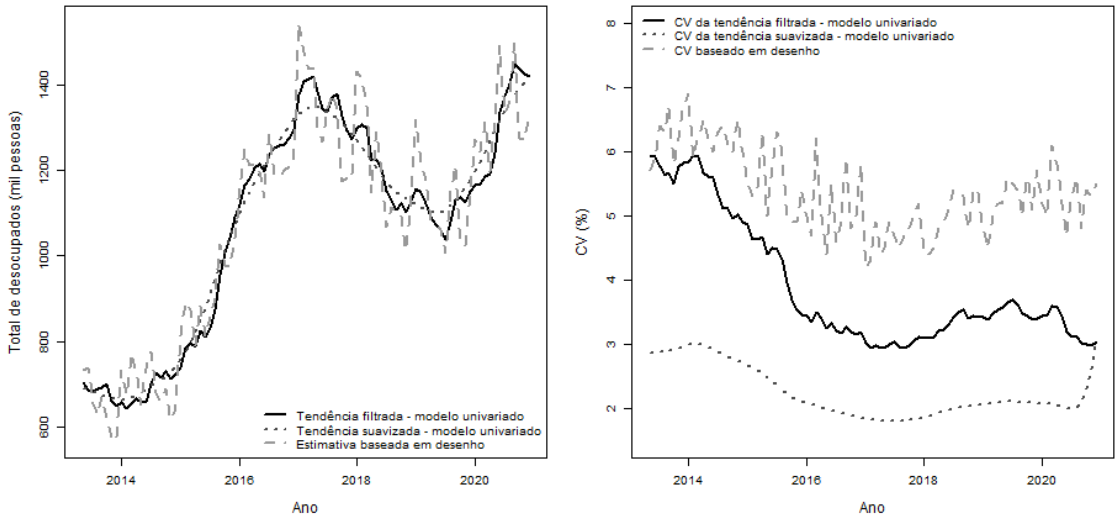
O comportamento da tendência do total de desocupados no Brasil evidencia o crescente aumento da desocupação ao longo do ano de 2020. Minas Gerais apresentou uma inversão nos últimos meses considerando a estimativa filtrada da tendência, porém a tendência suavizada ainda indicou uma trajetória ascendente até o final de 2020. No caso de Roraima, a tendência suavizada indicou uma inversão de trajetória após alcançar os níveis mais altos da série histórica nesse estado. Em relação aos picos históricos, tanto o Brasil quanto Minas Gerais superaram em 2020 os níveis de desocupação observados pela crise econômica e política de 2015-2017.

Figura 6 – Estimativas mensais baseadas no desenho amostral e em modelo do total de desocupados e coeficientes de variação - Brasil, Minas Gerais e Roraima

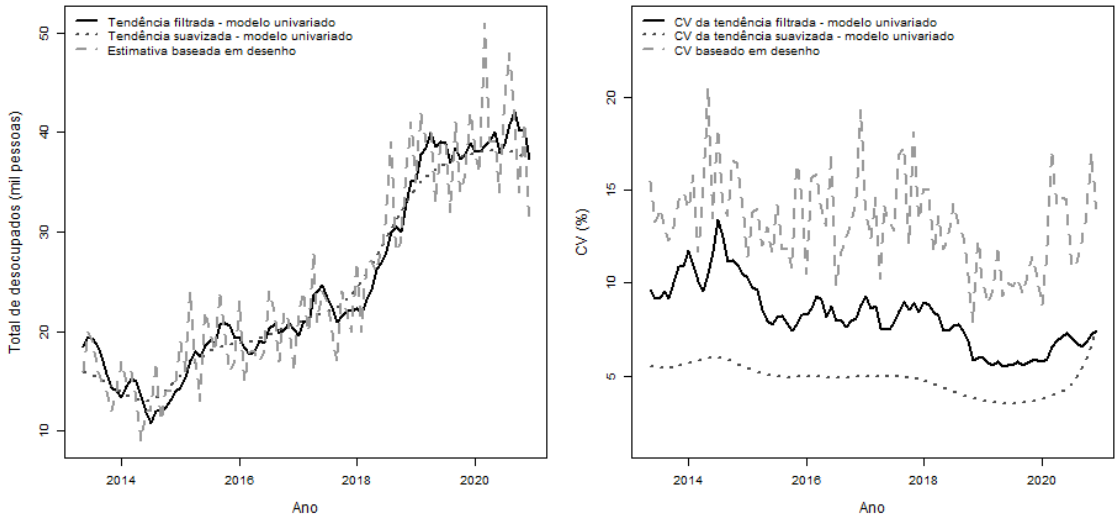
Brasil



Minas Gerais



Roraima



Fonte – Elaboração Própria

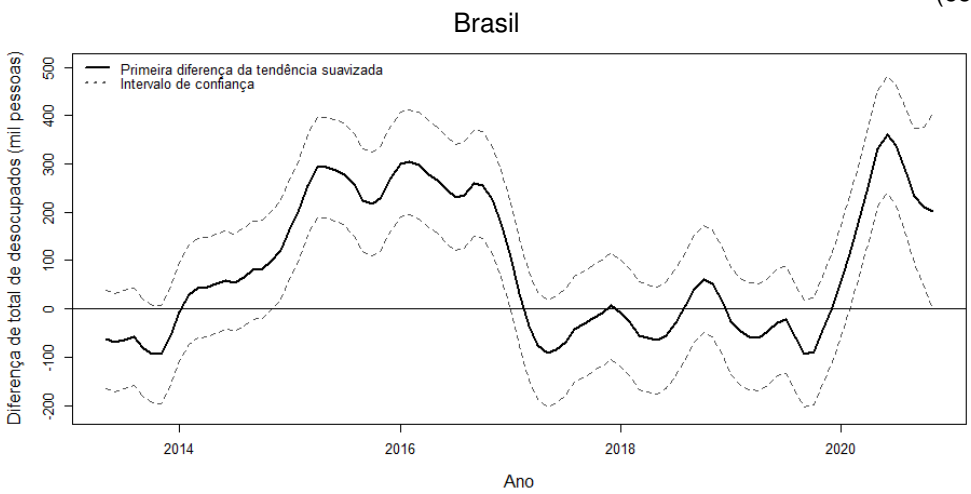
Em termos da precisão estatística das estimativas comentadas anteriormente, existem vantagens na produção das estimativas baseadas em modelo comparativamente àquelas baseadas no desenho amostral. Apesar de não serem notadas no início da série, as vantagens crescem com o tempo e são maiores para estados com amostras menores. A Figura 6 ilustra os percentuais dos coeficientes de variação menores em grande parte do tempo e para Roraima verifica-se uma redução de um patamar em torno de 15% para 8% no caso das estimativas filtradas, e para 5% no caso da tendência suavizada.

Dada a variabilidade observada em estados com amostras menores, como Roraima, a Figura 6 também destaca a vantagem de interpretar o comportamento pela tendência em detrimento de se observar apenas a série temporal da estimativa baseada no desenho amostral. Acompanhar as estimativas dos movimentos da tendência a curto prazo é uma estratégia recomendada nas análises de conjuntura.

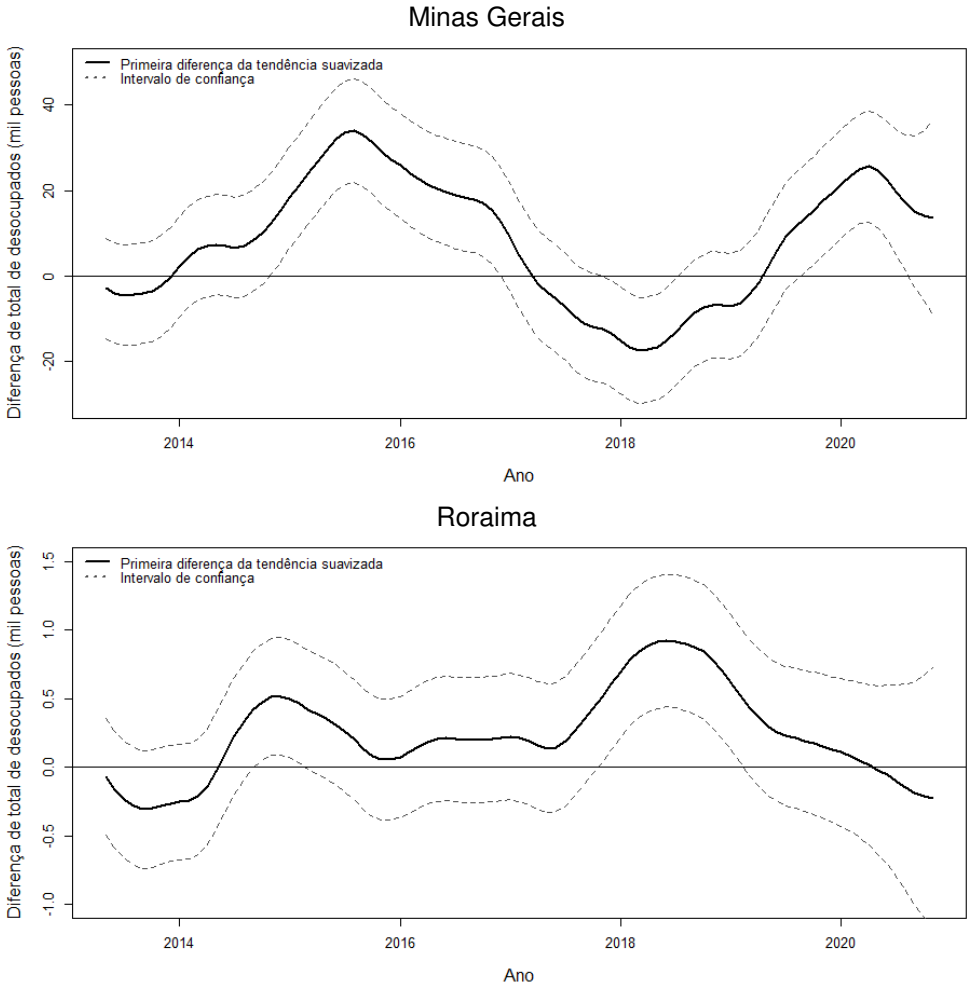
Ao analisar o comportamento ao longo do tempo da diferença da série de tendência e os respectivos intervalos de confiança de 95%, pode-se observar os meses nos quais a variação mensal foi estatisticamente diferente de zero. Optou-se por realizar essa análise na série suavizada, o que representa avaliar variações em uma série na qual é mais difícil detectar mudanças significativas mês a mês, dado o próprio procedimento de suavização. A Figura 7 apresenta as estimativas da variação mensal para o Brasil, Minas Gerais e Roraima.

Figura 7 – Estimativas mensais da primeira diferença da tendência suavizada baseadas no modelo univariado do total de desocupados e intervalos de confiança a 95% - Brasil, Minas Gerais e Roraima

(continua)



(conclusão)



Fonte – Elaboração Própria

Observa-se que para o Brasil e Minas Gerais, no período da crise econômica e política de 2015-2017 e na recente crise provocada pela pandemia de COVID-19, há evidências de diferenças significativas mês a mês (Figura 7). Roraima apresentou diferenças estatisticamente diferentes de zero em momentos distintos comparativamente ao Brasil e Minas Gerais.

A Tabela 3 apresenta os períodos com início, fim e duração de ocorrências de diferenças mensais significativas na tendência suavizada da série do total de desocupados. Também foram adicionados à Tabela 3 os períodos indicados como recessões econômicas pelo Comitê de Datação de Ciclos Econômicos (CODACE) para o Brasil, analisadas a partir da série do Produto Interno Bruto (PIB) trimestral dessazonalizado a preços de mercado (Fundação Getúlio Vargas, 2020). Essa datação é realizada trimestralmente e considerou que a economia brasileira entrou em recessão no segundo trimestre 2014.

Tabela 3 – Períodos e duração de diferenças significativas nas estimativas de tendência do total de desocupados e recessão econômica

Desagregação geográfica	Início	Fim	Duração (mês)
Diferenças mês-a-mês significativas			
Brasil	jan/15	jan/17	24
	mar/20	nov/20	9
Minas Gerais	dez/14	jan/17	25
	jan/18	jul/18	7
	out/19	set/20	12
Roraima	nov/14	mar/15	5
	dez/17	fev/19	15
Recessão econômica [1]			
Brasil	abr/14	dez/16	33
	jan/20	*	+30

Fonte - Elaboração própria com base nos dados básicos PNADC/IBGE e [1] Fundação Getúlio Vargas (2020). Nota:* O relatório disponibilizado em 29 de junho de 2020 comunicou o início da recessão e até julho de 2022 não houve indicação do final

Em termos do total de desocupados, um crescimento significativo mês a mês só foi verificado a partir de janeiro de 2015 comparado a dezembro de 2014. Esse é o período no qual se verifica o início de crescimento do total de desocupados que levou à alteração do nível da série a partir de 2017. Observa-se também que em março de 2020 já ocorreu uma mudança diferente de zero no total de desocupados no país e tal fenômeno se estendeu a todos os meses de 2020, com exceção de dezembro desse ano em que não houve alteração estatisticamente relevante. Isso corrobora o argumento em favor da atual necessidade de produção de indicadores mensais do mercado de trabalho dado que sua dinâmica no Brasil não se apresentou constante ao longo dos anos analisados no presente estudo.

A análise das diferenças temporais de curto prazo de indicadores-chave de mercado de trabalho foi também objeto de estudo em Lila & Freitas (2007). Os autores usaram dados da Pesquisa Mensal de Emprego (PME) do IBGE, referência na época para produção de estatísticas oficiais de mercado de trabalho, para avaliarem a pertinência de uma amostra mensal ou trimestral para uma pesquisa nacional domiciliar, que viria a substituir conjuntamente a PNAD e a PME. Os resultados desse estudo concluíram que, baseado nos dados das 6 regiões metropolitanas que eram alvo da PME, havia esparsos casos de variação mensal relevante. Com isso e também apoiado em questões operacionais, decidiu-se que o esquema amostral da PNADC seria trimestral.

Por outro lado, o estudo aqui apresentado e que utilizou os próprios dados da PNADC encontrou evidência distinta. A variação mensal foi relevante em longos períodos de tempo, inclusive com o caso do Brasil que entre abril de 2014 e dezembro de 2016 apresentou 33 meses consecutivos de variação

significativa (Tabela 3). Além disso, mesmo para economias estaduais com uma maior representatividade econômica, as dinâmicas regionais podem ser diferentes da nacional, tal como acontece em Minas Gerais.

A desocupação em Minas Gerais apresentou, ao contrário do Brasil, variações mensais diferentes de zero de janeiro até julho de 2018. Sendo ainda notada nesse período uma redução do total de desocupados nessa UF. A Figura 6 mostra que, no período após 2017, a queda do indicador em Minas Gerais foi mais acentuada que a vista para o Brasil. E a partir de outubro de 2019, o número de desocupados nessa UF voltou a subir, quando até setembro de 2020 combinou-se aos efeitos da pandemia de COVID-19 na economia.

Em contraponto, Roraima apresentou comportamento distinto do Brasil e de Minas Gerais no que se refere à significância estatística das variações mensais, bem como sua série de total de desocupados não guarda semelhança com a série nacional ou com a de Minas Gerais.

5. CONSIDERAÇÕES FINAIS

Além de ter como objetivo produzir estimativas mensais para o total de desocupados e estimativas da variação mensal para o Brasil e unidades da federação selecionadas utilizando a abordagem baseada em modelos de séries temporais, esse artigo investigou a integração de dados da PNADC com o registro administrativo de número de beneficiários do seguro-desemprego.

Os resultados do estudo indicaram que a correlação entre os componentes de tendência das séries não apresenta significância estatística. Assim, os modelos bivariados não foram capazes de reduzir os coeficientes de variação das estimativas, sendo esses inócuos para a melhoria da precisão. Portanto, resultados encontrados em experiências internacionais não se replicam para o caso brasileiro. Existem alguns descompassos entre as séries de total de desocupados e do número de beneficiários do seguro-desemprego que explicariam esse resultado distinto para o caso brasileiro. As divergências envolvem questões: (i) burocráticas, como um prazo máximo de 120 dias para requerer o benefício; (ii) conceituais, dado que a busca por trabalho pode não ocorrer por parte do segurado durante o período de recebimento do benefício; (iii) de elegibilidade, pela diferença de magnitude existente entre o número de desocupados e o número de beneficiários, evidenciando o papel da informalidade na economia brasileira e também o impacto da mudança dos critérios para recebimento do benefício, que teve um efeito de redução do número de beneficiários no mesmo momento que a desocupação cresceu no país durante uma crise econômica e política. Ressalta-se ainda que são

possíveis estudos futuros que incorporem outras formas de realizar a análise de intervenção como apresentado por Santos, Franco & Gamerman (2010).

Diante disso, todas as análises foram desenvolvidas com modelos univariados que se demonstraram vantajosos para a produção de estimativas mensais nacional e regionais do total de desocupados. Tais conclusões são semelhantes às encontradas por Gonçalves *et al.* (2022) para a taxa de desocupação. Destaca-se ainda o proveito de analisar estimativas de tendência que indicaram as trajetórias recentes e distintas do Brasil, de Minas Gerais e de Roraima, em detrimento das estimativas mensais baseadas no desenho amostral que são voláteis.

Adicionalmente, as estimativas de primeira diferença (que captam variações mensais) mostraram-se significativas por longos períodos desde o início da PNADC, evidenciando que as alterações econômicas, políticas, ou decorrentes da pandemia da COVID-19, impactaram, e continuam impactando, o mercado de trabalho.

Destaca-se que as séries de total de desocupados e suas respectivas variações mensais para Minas Gerais e Roraima foram distintas entre si e também diferentes da série nacional, ratificando o argumento da necessária produção de estimativas mensais para todas as UFs. E ainda, apesar da escolha por essas UFs, estudos futuros são indicados para os demais estados de forma a verificar se os resultados aqui apresentados persistem nesses outros contextos.

Assim, esse artigo demonstrou que os registros administrativos de seguro-desemprego não colaboram para a melhoria da precisão das estimativas amostrais de total de desocupados. Como destaque, o trabalho evidenciou a relevância de produção de estatísticas mensais ao constatar significativas e diferenciadas mudanças a curto prazo na desocupação no país e em dois de seus estados.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Banco Central do Brasil (2016). Efeito das mudanças das regras do seguro-desemprego. In: Boletim regional do Banco Central do Brasil. Brasília: Banco Central do Brasil, (pp. 110–112).

Banco Central do Brasil (2020). Estimativa para dados “mensalizados” da PNAD contínua. In: Relatório de Inflação. Brasília: Banco Central do Brasil, (pp. 38–40).

Binder, R. & Dick, J. (1989). Modelling and estimation for repeated surveys. Survey Methodology, 15, 29–45.

Boonstra, H.J.; van den Brakel, J. (2019). Estimation of level and change for unemployment using structural time series models. Survey Methodology, 45, 395–425.

van den Brakel, J.; Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. Survey Methodology, 35, 177–190.

Durbin, J.; Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. (2a ed.). Oxford: Oxford University Press.

Elliott, D.J.; Zong, P. (2019). Improving timeliness and accuracy of estimates from the UK labour force survey. Statistical Theory and Related Fields, 3, 186–198.

Fundação Getúlio Vargas. (2020). Comunicado de Datação de Ciclos Mensais Brasileiros.

Gonçalves, C.; Hidalgo, L.; Silva, D.B.N.; van den Brakel, J. (2022). Model-based unemployment rate estimates for the Brazilian labour force survey. Journal of the Royal Statistical Society: Series A (Statistics in Society), 185(4), 1707-1732.

Harvey, A.; Chung, C. (2000). Estimating the underlying change in unemployment in the UK. Journal of the Royal Statistical Society: Series A (Statistics in Society), John Wiley and Sons, 163, 303–309.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Hecksher, M. (2020). Valor impreciso por mês exato: microdados e indicadores mensais baseados na PNAD Contínua. Rio de Janeiro. (Nota técnica, 62).

Instituto Brasileiro de Geografia e Estatística. (2014a). Pesquisa Nacional por Amostra de Domicílios Contínua: Notas metodológicas: Versão 1. Rio de Janeiro.

Instituto Brasileiro de Geografia e Estatística. (2014b). Conjunto mínimo de indicadores padrão de qualidade a ser aplicado no MERCOSUL. Rio de Janeiro. (Textos para Discussão, 52).

Instituto Brasileiro de Geografia e Estatística. (2007). Pesquisa Mensal de Emprego. Rio de Janeiro, (Série Relatórios Metodológicos, 23).

Instituto Brasileiro de Geografia e Estatística. (2020). Sistema de Contas Nacionais: Brasil 2018. Rio de Janeiro.

Instituto Brasileiro de Geografia e Estatística. (2021). Sistema Nacional de Índices de Preços ao Consumidor: IPCA e INPC. Rio de Janeiro.

Instituto Brasileiro de Geografia e Estatística. (2022a). PNAD Contínua - Pesquisa Nacional por Amostra de Domicílios Contínua. Rio de Janeiro. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=o-que-e>.

Instituto Brasileiro de Geografia e Estatística. (2022b). Pesquisa Nacional por Amostra de Domicílios Contínua: Notas técnicas: Versão 1. 10. Rio de Janeiro.

Instituto de Pesquisa Econômica Aplicada (2016). Trabalho e renda. In: Política Sociais: acompanhamento e análise. Brasília: IPEA. pp. 309–360.

Lila, M. L.; Freitas, M. P. S. (2007). Estimação de intervalos de confiança para estimadores de diferenças temporais na Pesquisa Mensal de Emprego. Rio de Janeiro, (Textos para Discussão, 22).

Ministério do Trabalho e Previdência. (2021). Estatísticas do Seguro-Desemprego. Brasília.

Mourão, A. N. M.; Almeida, M. E.; Amaral, E. F. L. (2013). Seguro-desemprego e formalidade no mercado de trabalho brasileiro. Revista Brasileira de Estudos Populacionais, Rio de Janeiro, 30(1), 251–270.

Petris, G. (2010). An R package for dynamic linear models. Journal of Statistical Software, 36(12). 1–16.

Pochmann, M. (2015). Ajuste econômico e desemprego recente no brasil metropolitano. Estudos Avançados, São Paulo, 29(85) 7–19.

R Core Team. R: (2020). A Language and Environment for Statistical Computing. Vienna, Austria.

Ramos, C. A. (2003). Políticas de Geração de Emprego e Renda: justificativas teóricas, contexto histórico e experiência brasileira. Brasília. (Textos para Discussão, 277).

Santos, T. R.; Franco, G. C.; Gamerman, D. (2010). Comparison of classical and Bayesian approaches for intervention analysis. International Statistical Review, 78, 218–239.

Schiavoni, C.; Palm, F.; Smeekes, S.; van den Brakel, J. (2021). A dynamic factor model approach to incorporate big data in state space models for official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 184(1), 324-353.

Scott, A. J.; Smith, T. M. F. (1974). Analysis of repeated surveys using time series methods. Journal of the American Statistical Association, 69, 674–678.

Scott, A. J.; Smith, T. M. F.; Jones, R. G. (1977). The application of time series methods to the analysis of repeated surveys. International Statistical Review, 45, 13–28.

ISSN 2675-3243

volume 81

número 248

Janeiro/dezembro 2023