UNIVERSIDADE FEDERAL DO CEARÁ – UFC
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE – FEAAC
PROGRAMA DE ECONOMIA PROFISSIONAL – PEP

FÁBIO RENATO ARRUDA COELHO

MERCADO DE TRABALHO BRASILEIRO E OS IMPACTOS DA PANDEMIA

FORTALEZA
2024

FÁBIO RENATO ARRUDA COELHO

MERCADO DE TRABALHO BRASILEIRO E OS IMPACTOS DA PANDEMIA

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Orientador: Prof. Dr. Nicolino Trompieri Neto

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C616m Coelho, Fábio Renato Arruda.
Mercado de trabalho brasileiro e os impactos da pandemia / Fábio Renato Arruda Coelho. – 2024.
39 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Mestrado Profissional em Economia do Setor Público, Fortaleza, 2024.
Orientação: Prof. Dr. Nicolino Trompieri Neto.

1. Desemprego. 2. Marchine learning. 3. Pandemia. 4. Estratos geográfico. I. Título.

CDD 330

FÁBIO RENATO ARRUDA COELHO

MERCADO DE TRABALHO BRASILEIRO E OS IMPACTOS DA PANDEMIA

Dissertação submetida à Coordenação do Programa de Economia Profissional – PEP, da Universidade Federal do Ceará - UFC, como requisito parcial para a obtenção do grau de Mestre em Economia. Área de Concentração: Economia do Setor Público.

Aprovada em: **21 de fevereiro de 2024**.

BANCA EXAMINADORA

_____

Prof. Dr. Nicolino Trompieri Neto (Orientador)
Universidade de Fortaleza (UNIFOR)

_____

Prof. Dr. Cristiano da Costa da Silva
Universidade Federal de Pernambuco (UFPE)

_____

Prof. Dr. Diego Rafael Fonseca Carneiro
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Agradeço imensamente a Deus pelas conquistas que consigo.

Sou eternamente grato a meus pais Teófilo e Edmyrtes (*in memoriam)* por todos os ensinamentos.

Tenho enorme gratidão a vida por me proporcionar ser pai do Miguel, que me dá forças para todas as empreitadas.

Agradeço ao professor Nicolino pela orientação e banca pelas contribuições.

Por fim um agradecimento as funcionárias do Mestrado Profissional Geisa e Márcia.

RESUMO

A pandemia do Covid-19 devastou a saúde e a economia em inúmeros países, no Brasil não foi diferente. O período da primeira onda em 2020 foi marcado por uma restrição de locomoção que afetou drasticamente o emprego e o modo como as pessoas trabalhavam presente dissertação busca avaliar o impacto da pandemia nas economias locais brasileiras, mais precisamente os estratos geográficos retirado do IBGE. Foram utilizados dados do IBGE para desemprego e para dados relativos a questão sanitária se utilizou o Datasus. Através do método *marchine learning* recém-desenvolvido para construção contrafactual. As estimativas dos resultados, resultados indicam que os efeitos econômicos do choque da COVID-19 variam dramaticamente em todo o território brasileiro. As maiores perdas de emprego ocorreram em áreas caracterizadas por alta exposição a riscos de agregação social e fragilidades pré-existentes no mercado de trabalho. Essas descobertas demandam uma resposta política baseada no local para enfrentar a geografia econômica desigual da pandemia.

Palavras-chave: Desemprego. *Marchine learning*. Pandemia. Estratos geográfico.

ABSTRACT

The COVID-19 pandemic wrought havoc on global health systems and economies, with Brazil being no exception. The initial wave in 2020, characterized by the imposition of mobility restrictions, significantly impacted employment and altered occupational practices. The present dissertation seeks to evaluate the ramifications of the pandemic on the local economies of Brazil, specifically focusing on geographical strata delineated by the Brazilian Institute of Geography and Statistics (IBGE). Utilizing data from IBGE regarding unemployment and health issues, sourced from Datasus, this study employs an innovative machine learning methodology devised for the construction of counterfactuals. The findings suggest that the economic repercussions of the COVID-19 disturbance differed markedly across the Brazilian regions. The most substantial employment reductions were observed in areas with heightened risks of social congregation and pre-existing labor market vulnerabilities. These outcomes underscore the necessity for policy interventions tailored to the disparate economic landscapes shaped by the pandemic.

Keywords: Unemployment. Machine learning. Pandemic. Geographic strata**.**

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O mercado de trabalho em situações de ruptura tem grande relevância, principalmente após a experiencia da Pandemia do Covid-19. Nesse momento de isolamento o Brasil, e o Mundo, enfrentou desafios que afetaram tanto empregados quanto empregadores e desequilibrou de forma significativa o equilíbrio do mercado de trabalho.

Em virtude de evitar a proliferação do vírus diversas dinâmicas alteraram o mercado de trabalho. De início o impacto foram as medidas de contenção e distanciamento, que restringiram drasticamente a mobilidade. Assim, em decorrência dessas medidas, empresas fechadas, algumas permanentemente outras temporariamente, e houve um processo de demissão em massa. Cabe ressaltar que alguns setores foram mais afetados que outros em virtude da característica essencial de alguns serviços. Dessa forma setores de Turismo e Hotelaria foram drasticamente prejudicados, pois as medidas restritivas impossibilitaram quaisquer atividades nesse setor (Kapicka; Rupert, 2020).

Tal interrupção, de forma abrupta, na economia trouxe consigo uma elevada taxa de desemprego com consequências negativas para os trabalhadores e a sociedade como um todo. Houve escassez de emprego e queda geral da renda das famílias. Diante desse quadro houve uma certa automação de empregos específicos, ao passo que empregos na área de saúde e da tecnologia de informação ampliaram sua demanda. Dessa maneira mais do que o impacto imediato na economia e na renda, a pandemia proporcionou mudanças estruturais na forma de trabalhar (Blustein *et al*., 2020).

Porém os impactos da pandemia no desemprego não foram similares em regiões e países. Locais mais dependentes de setores como turismo e serviços foram mais afetados que regiões com atividades voltada para tecnologia da informação. A pandemia pode ter ampliado a desigualdade entre regiões, dado que localidades mais pobres sofreram mais com a queda na atividade econômicas, principalmente atividades informais (Cerqua; Leta, 2022).

O cenário de pandemia exige uma medida por parte do Estado através de políticas públicas. Buscando mitigar os efeitos da pandemia diversos países injetaram recursos na economia, seja para transferir renda diretamente as pessoas seja por meio de incentivos as empresas para atenuar as demissões.

Portanto o impacto da pandemia no desemprego pode variar de acordo com o grau de restrição e de infecção de cada país ou região, das medidas de políticas públicas adotadas e dos setores da economia. Cabe ressaltar que os efeitos da pandemia não foram apenas as causas

de curto prazo, como dito em parágrafos acima o cenário de restrição de mobilidade reformulou a prestação de alguns serviços, com ampliação de empregos no setor de tecnologia (Costa Dias *et al*., 2020).

Esse projeto de pesquisa contém, além dessa introdução, uma segunda seção que contextualiza o desemprego e o cenário pandêmico por qual passou o mundo. Na terceira seção se encontra uma breve revisão de literatura sobre trabalhos que analisaram os impactos na pandemia no mercado de trabalho; ao passo que na seção quatro e cinco são detalhadas a metodologia e a base de dados; por fim na sexta seção tem-se o cronograma da pesquisa.

2 DESEMPREGO, SETORES DE ATIVIDADE E PANDEMIA

O equilíbrio no mercado de trabalho e a concepção e as causas do desemprego estão longe de ser senso comum entre economistas e escolas de pensamento econômico. Os economistas clássicos e suas vertentes, de forma geral, defendem que o mercado de trabalho tende a se equilibrar com as forças de mercado e que não há desemprego involuntário. Por sua vez as teorias keynesianas de matriz heterodoxa advogam que existem desequilíbrios que geram desemprego involuntário e que se faz necessário alguma intervenção estatal.

Para os clássicos a ocorrência de desemprego voluntário é temporária, em decorrência de desequilíbrios naturais do mercado. O principal fator para ajustar o mercado é o salário dos trabalhadores, que são flexíveis. Portanto o desemprego que ocorre é devido de trabalhadores que escolhem o desemprego devido a salários aquém do esperam receber (Snowdon; Vane, 2005).

Uma ruptura com esse pensamento veio com a *Teoria Geral* de Keynes, que deu a Demanda Agregada grande papel na determinação no nível de emprego. Por essa visão poderia haver desemprego involuntário devido a ineficiência de demanda agregada. Em períodos de recessão a demanda agregada cai em virtude do investimento privado e da queda do consumo, o que deixa a economia com alto desemprego involuntário. Portanto não dependia apenas do mercado o equilíbrio do mercado de trabalho, mas também de intervenção estatal por meio de políticas fiscais e monetárias expansionistas (Simonsen, 1986).

Uma variante de visão heterodoxa é a noção de desemprego estrutural em que o desemprego é fruto do excesso de mão de obra em relação ao capital de modo que seja impossível alocar toda oferta de trabalho. De acordo com Simonsen[1] (1963) isso significa que a produtividade marginal se anule. Essa visão difere da abordagem keynesiana em relação a solução para esse excesso de oferta de trabalho, tal solução não ocorre com o aumento da demanda agregada, mas em progressivos aumentos de capital.

A partir da década de 1960, baseado em Philipps (1958), o desemprego passou a ser associado a taxa de inflação e a atividade econômica. Por esse mecanismo, Curva de Phillips, havia uma relação inversa entre inflação e desemprego de modo que quando a economia operava com pleno emprego havia uma forte pressão sobre a inflação. Essa teoria admitia que existe uma taxa natural de desemprego que não acelera a inflação (NAIRU).

[1] O autor descreve como contrassenso tal abordagem.

Posteriormente, na década de 1980, surgiram evidências de persistência na taxa de desemprego que mantinha a taxa de desemprego superior a NAIRU. Esse fenômeno macroeconômico de persistência da taxa de desemprego foi denominado de *histerese.* (Blanchard, 1986).

Em suma essas teorias abordam a questão do desemprego em função da ação livre do mercado e da intervenção estatal. Tais fatores são de caráter macroeconômico e possuem uma visão pouco aprofundada entre o comportamento dos trabalhadores e das firmas (Zylberstajn; Balbinotto, 1999). Portanto se faz necessário algumas considerações com fundamentação microeconômica desse tema.

A teoria da busca por emprego (*job search*) analisa o mercado de trabalho com base em como trabalhadores buscam por emprego avaliando os ganhos e custos desse processo. Os agentes, firmas e trabalhadores, possuem informações incompletas, de forma que o trabalhador não possui informação completa sobre a vaga. Nesse ambiente os trabalhadores possuem características profissionais diferentes e existe um custo de procura de emprego. Os trabalhadores buscam por emprego por maiores salários, mas enfrentam os custos de busca. Assim, recursos de assistência ao trabalhador colaboram para que trabalhadores possam esperar por mais tempo desempregado até que encontrem um emprego (Campêlo *et al*., 2018).

Por outro lado, a Teoria dos deslocamentos setoriais tenta explicar o comportamento no emprego em decorrência de mudanças na alocação de recursos produtivos. Ao longo do tempo a alocação de recursos muda entre os setores da economia, dessa maneira inovações tecnológicas e mudanças nas técnicas produtivas podem gerar mudanças na demanda por trabalho (Zylberstajn; Balbinotto, 1999).

Como observado existem diversos fatores macro e microeconômicos que impactam na taxa de desemprego e no equilíbrio do mercado de trabalho. Muitos desses fatores ocorrem com frequência na atividade econômica, no entanto quando ocorre um fenômeno de ruptura como uma pandemia diversos desses fatores podem ocorrer simultaneamente. As medidas restritivas da pandemia tiveram forte impacto na demanda agregada e impactaram setores e regiões de formas distintas.

Em fins de dezembro de 2019 a Comunidade Internacional foi alertada sobre diversos casos com características de pandemia na cidade de Wuhan, na China. Esses casos se tratava de uma nova Cepa do Corona vírus ainda não identificado em seres humanos. Em janeiro de 2000 a Organização Mundial da Saúde (OMS) declarou esse surto de Covid como uma Emergência de Saúde Pública de Importância Internacional (ESPII). Estava dada as condições

para que se declarasse uma pandemia, o que de fato ocorreu em 11 de março, quando a OMS declarou a COVID-19 como uma pandemia (OPAS, 2023).

Nesse momento vários países começam práticas de medidas restritivas para reduzir a propagação do vírus. Houve um aumento de números de casos e mortes em virtude da doença. Em abril e maio de 2020 o vírus continua a se espalhar e as medidas de restrições se enrijecem, somente a partir de junho alguns países começa a ensaiar uma flexibilização e abertura gradativa.

De forma geral o mundo quase como um todo reduziu drasticamente a sua produção e atividade econômica por um semestre em 2020. Apenas setores ligados a saúde e ao comercio de bens de primeira necessidade mantiveram funcionamento. O comercio eletrônico e profissões ligadas a tecnologia ganharam ímpeto nesse momento. Antes que se começasse um processo de vacinação diversos países passaram por uma segunda onda da doença que voltou a travar a economia (Cerqua; Leta, 2022).

A pandemia da COVID-19 teve um impacto significativo no cenário global de emprego, com certos setores enfrentando um risco mais elevado do que outros. Trabalhos que demandam proximidade física ou envolvem interação frequente com o público mostraram-se particularmente vulneráveis aos efeitos da pandemia. Essas ocupações frequentemente apresentam maior risco de exposição ao vírus, o que pode agravar disparidades já existentes no mercado de trabalho.

O ônus do risco de distanciamento social devido à COVID-19 recai desproporcionalmente sobre grupos vulneráveis da força de trabalho, como mulheres, trabalhadores mais velhos, não-nativos, pessoas com menor nível educacional e aqueles empregados em locais de trabalho de pequeno porte. Os resultados destacam a necessidade de respostas políticas imediatas e direcionadas para evitar perdas contínuas de emprego e o agravamento das desigualdades no mercado de trabalho e na sociedade em decorrência da pandemia.

O estudo de Pouliakas e Branka (2020) utiliza uma abordagem baseada em habilidades para identificar fatores individuais e profissionais com maior probabilidade de serem impactados pelas medidas de distanciamento relacionadas à COVID-19. Abrangendo os países da União Europeia e utilizando dados da Pesquisa Europeia de Habilidades e Empregos, o estudo cria um índice de risco de distanciamento. Esse índice fundamenta-se em descritores de habilidades que categorizam empregos de acordo com o nível de proximidade.

As estimativas mais conservadoras desse estudo indicam que aproximadamente 45 milhões de empregos nos 27 membros da União Europeia enfrentaram um risco elevado de interrupção durante a primeira onda da COVID-19. O estudo identificou os principais setores ou grupos de atividades com maior risco de perda de emprego. Abaixo, segue um quadro dos grupos de emprego e seus níveis de risco conforme indicados por esse estudo, que será base ao utilizado para a elaboração da modelagem econométrica dessa dissertação.

No caso brasileiro as medidas restritivas da pandemia vieram entre o fim de março e início de abril de 2020. Diversos governos e gestores subnacionais decretaram medidas restritivas que se arrastaram até metade do ano. O país ainda passou por uma segunda onda da pandemia entre fevereiro e março de 2021 (Costa, 2020).

Como comentado acima o Brasil passou por dois choques na atividade econômica em menos de um ano, o que afetou diretamente a atividade econômica e o nível de emprego. O Brasil vinha tentando se recuperar de uma recessão causada em 2015, em decorrência de uma crise política e econômica, que derrubara o PIB e o emprego no país. Assim em um intervalo de 5 anos o Brasil passou por dois processos de queda na economia e no emprego.

Ao passo que entre 2015 e 2016 a terapêutica para a crise foram medidas de austeridade, como reforma nas leis trabalhistas, aumento da taxa de juros e corte de gastos, em 2020 o remédio não podia ser tão amargo em virtude da situação de calamidade. Assim o Governo Federal impôs um plano com recursos para transferir renda direta e evitar demissões em massa.

O combate aos efeitos da pandemia foi uma expansão dos gastos e recuperação da demanda agregada. Com os recursos muitas empresas mantiveram os empregados e com a transferência de renda (auxílio emergencial) muitas famílias puderam manter um mínimo de consumo. O auxílio emergencial por meio do incentivo do consumo afetou positivamente a arrecadação dos estados brasileiros e ajudou a injetar renda no mercado informal.

No entanto, ainda que com as medidas de incentivo a economia do Governo Federal o desemprego no Brasil foi muito afetado e tem demorado a voltar para os índices anteriores a pandemia. Portanto é de suma importância entender os principais fatores impactantes do desemprego durante a pandemia no Brasil. Dessa forma será possível saber quais medidas o governo deve aplicar para reduzir esse indicador[2].

[2] Quando se fala em "medidas" não necessariamente se fala intervenção. Caso as consequências sejam atribuídas ao mercado de trabalho em si, cabe ao Governo apenas permitir as condições de mercado que ele se equilibre.

3 REVISÃO DE LITERATURA

A pandemia alterou o cenário econômico de todos os países do globo e teve forte impacto na retração da economia. Diante das medidas protocolares sanitárias os cenários econômicos e setoriais se modificaram. Tais mudanças alteraram formas de trabalhar de interagir dos indivíduos e reformularam setores de emprego. Empregos da área de tecnologia da informação ganharam espaço na economia e atividades informais e que trabalham com público diretamente foram afetadas negativamente. Os impactos sobre o desemprego ainda estão sendo observados em busca de soluções que atenuem esse problema. A seguir abordaremos alguns estudos que buscam compreender efeitos da pandemia sobre o desemprego.

Com o desenvolvimento de novas técnicas de estudos econométricos e ampliação da base de dados tem sido possível entender melhor a relação entre as variáveis econômicas. Foi com o uso de *Maching Learning* que Cerqua e Leta (2022) pesquisaram a desigualdade dos efeitos da pandemia do Covid-19 entre as regiões italianas. Por meio de um contrafactual os autores concluíram que os efeitos da pandemia foram divergentes de acordo com a região do país. Os maiores casos de desemprego se deram em localidades com riscos sociais e fragilidades, antes da pandemia, no mercado de trabalho. O trabalho também comparou a crise da pandemia com a recessão de 2008-2009[3] na busca de possíveis padrões similares.

O impacto de pandemias no crescimento econômico e no desemprego foi estudo por Rodríguez-Caballero e Vera-Valdés (2020). Diante dos efeitos devastadores da pandemia do Covid-19 os autores buscaram analisar os efeitos duradouros de pandemias no emprego e no PIB. Com dados centenários para Reino Unido foram testadas quebras estruturais baseadas no método de Bai e Pherron (2003). Os resultados indicam que após pandemias e rupturas na economia o crescimento econômico tem alta, no entanto há uma persistência na taxa de desemprego.

Nos países membros da União europeia o desemprego de jovens é um sintoma que persiste desde a década de 1980. Diante desse cenário Lambovska *et al*. (2021) optaram por estudar o impacto da pandemia no desemprego juvenil. Durante a pandemia o desemprego nessa faixa etária cresceu mesmo em países do bloco que não tinham tendencia de desemprego alto para jovens.

[3] Crise dos *subprimes*.

No caso brasileiro não há um volume de estudos que tratem diretamente do impacto da pandemia no desemprego, como métodos econométricos. No entanto o trabalho de Komatzu e Menezes Filho (2020) simulou dados para observar possíveis impactos. Perante a paralisação da economia o estudo detalhou que 37 milhões de pessoas dependia de emprego em atividades mais afetadas pela quarentena. Caso não houvesse medidas governamentais e as empresas demitissem em massa a taxa de desemprego poderia ter chegado a 28%.

Por sua vez (estudo do IPEA) enxergam uma complexidade na definição de atividades essenciais[4] no contexto da pandemia de COVID-19 no Brasil. Internacionalmente, essa definição foi baseada nas classificações oficiais de atividades específicas em cada localidade, proporcionando uma identificação precisa das empresas autorizadas a operar durante a crise. No entanto, no Brasil, a lista de atividades autorizadas foi definida por sucessivos decretos federais, sem utilizar a Classificação Nacional de Atividades Econômicas (CNAE) de forma consistente.

O estudo busca, então, associar os setores definidos como essenciais nos decretos com os setores da CNAE, reconhecendo a possibilidade de omissões e imprecisões na identificação real dos setores operando durante a pandemia. Cabe ressaltar que estados e municípios também estabeleceram suas próprias listas de setores essenciais, o qual o estudo não se debruça.

Ao comparar o percentual de emprego em setores considerados essenciais no Brasil com dados da União Europeia e dos Estados Unidos, verifica-se uma relação próxima entre os decretos federais iniciais e as medidas internacionais. Algumas discrepâncias são atribuídas às composições distintas do emprego em cada país.

Um ponto importante é que os dois primeiros decretos de março de 2020 foram mais restritivos em setores como construção civil e atividades profissionais, científicas e técnicas em comparação com os Estados Unidos e a União Europeia. Contudo, a partir do Decreto no 10.329/2020, houve uma significativa ampliação da lista de setores considerados essenciais, abrangendo mais de 70% do emprego formal no país a partir de maio de 2020.

[4] Para uma visão sobre as atividades desse estudo ver Anexo I.

4 FUNDAMENTOS BÁSICOS PARA ALGORITMOS BASEADOS EM ÁRVORES DE DECISÃO

A família de métodos baseados em árvores de decisão utiliza a estratégia de partição do espaço das variáveis explicativas de modo a obter um modelo com bom ajuste preditivo à variável de resposta avaliada. Além da flexibilidade na determinação da especificação ideal, estes métodos apresentam também vantagens para fins de análise dos resultados, haja visto que a cadeia de partições realizadas representa uma estrutura interpretável com base no conjunto de preditores adotados. Com o objetivo de apresentar a ideia intuitiva de modelos baseados em árvores de decisão, este subcapítulo faz uma síntese geral do algoritmo com base em Hastie *et al.* (2009).

Assuma um problema de predição para a variável de resposta contínua Y que é composto por duas variáveis explicativas X_1 e X_2. Para simplificar o processo, tome um modelo de árvore de decisão que somente realiza partições binárias sobre o espaço de variáveis explicativas e então modela a predição de Y em cada região a partir da respectiva média desta última variável. A determinação da variável explicativa (*spliting variable*) que irá particionar a árvore (ou seja, criar nós sobre a estrutura do modelo) e o ponto de corte da divisão (*split point*) são determinados com o objetivo de otimizar o ajuste preditivo de Y.

Com base no exposto, considere que no primeiro estágio a variável X_1 fora selecionada para realizar a partição e que o ponto de corte selecionado $X_1 = t_1$ define uma divisão relevante para o processo preditivo. Após a primeira iteração, novas regiões podem novamente serem particionadas, com o processo continuando até que alguma regra de finalização do processo seja ativada.

Observe que a região $X_1 \leq t_1$ é agora particionada segundo a variável X_2, com o ponto de corte $X_2 = t_2$, e o processo para este lado da árvore de decisão, encerra-se com duas regiões, quais sejam $R_1 = \{X_1 \leq t_1 \; e \; X_2 \leq t2\}$ e $R_2 = \{X_1 \leq t_1 \; e \; X_2 > t2\}$. De outro lado, na região $X_1 > t_1$, novamente adota-se uma nova partição sobre a variável $X_1 \leq t_3$, que inclui o terceiro nó terminal ($R_3 = \{t_1 < X_1 \leq t_3\}$). Para a região $X_1 > t_3$, define-se um novo ponto de corte através da variável X_2, qual seja $X_2 \leq t_4$. Assim, a nova partição finaliza com duas outras regiões constituídas ($R_4 = \{X_1 > t_3 \; e \; X_2 \leq t_4\}$; $R_5 = \{X_1 > t_3 \; e \; X_2 > t_4\}$).

Ao final, o resultado do processo é uma partição dos dados amostrais em cinco regiões $R_1, R_2, .., R_5$, e o modelo de regressão correspondente prediz Y com uma constante c_m dada pela média da variável de resposta em cada região:

$$\hat{f}(x) = \sum c_m I[(X_1, X_2) \in R_m]$$

Onde $I[.]$ denota uma variável indicadora com valor unitário se o par (X_1, X_2) está contido na *m-ésima* região e valor zero, caso contrário.

4.1 Modelo árvore de regressão (*regression tree*)

Uma vez apresentada a estrutura geral do modelo em árvore de decisão, é necessário estabelecer os critérios que definem o crescimento da árvore para constituir o modelo de regressão ideal, neste sentido, será apresentado nesta subseção o modelo *regression tree*. Suponha um modelo com p potenciais variáveis independentes (insumos) e uma variável dependente (resposta), que dispõe de N observações amostrais. Ou seja, para cada $i = 1,2,..,N$ existe um vetor $x_i = (x_{1i},..,x_{pi})$ e uma variável de resposta y_i.

O modelo árvore de regressão constitui-se de um algoritmo que deve determinar automaticamente qual das variáveis deve ser utilizada para discriminar os dados em uma dada região, qual o ponto de corte será selecionado e qual o formato que a árvore deve ter (número de nós gerados para particionar uma dada região).

Assume-se que o processo tenha tomado início com uma partição da árvore em M regiões $(R_1, R_2, .., R_M)$ e que a resposta em cada região seja modelada por uma constante c_m:

$$f(x) = \sum c_m I(x \in R_m)$$

Como critério de ajustamento do modelo, defina a minimização da soma do quadrado dos erros (SQE) de predição $\sum(y_i - f(x_i))^2$, de forma que a melhor valor para $\hat{c}_m$ será a média amostral de y_i na região R_m:

$$\hat{c}_m = média(y_i | x_i \in R_m)$$

Hastie *et al.* (2009) ressaltam, entretanto, que em regra, é inviável computacionalmente definir a melhor partição binária a partir da minimização da SQE, e indicam o uso do algoritmo *greedy* para encontrar uma solução global a partir de soluções locais ótimas. O algoritmo *greedy* visa resolver o problema de otimização em etapas, iniciando com um problema menor e o solucionado em cada estágio, a fim de encontrar uma solução global que se ajuste ao problema.

No caso do modelo árvore de regressão, com base na amostra completa, faça a divisão da $j-ésima$ variável em um dado ponto de corte s e defina um par de regiões:

$$R_1 = \{X | X_j \leq s\} \; e \; R_2 = \{X | X_j > s\}$$

Então, dentro do conjunto de variáveis e de possíveis pontos de corte, encontre o par (variável j-ponto de corte s) que resolve:

$$\min_{j,s} \left[\min_{c_1} \sum_{X_i \in R_1(j,s)} (y_i - c_i)^2 + \min_{c_2} \sum_{X_i \in R_2(j,s)} (y_i - c_i)^2 \right]$$

Para cada escolha j e s, o processo interno de minimização é resolvido por $\hat{c}_1 = média(y_i | x_i \in R_1(j,s))$ e $\hat{c}_2 = média(y_i | x_i \in R_2(j,s))$. Para cada variável *splitting*, o processo busca escanear através de todas as variáveis independentes, a determinação do melhor para (j,s) viável. Definido o ponto de divisão s, os dados são particionados em duas regiões e o processo é repetido em cada região. Então o processo é novamente repetido sobre todas as regiões resultantes.

Visto que no processo de partição não há um estabelecimento formal para o final do processo, uma questão que surge é até quando permitir o crescimento da árvore de decisão. Note que neste caso há um *trade-off*, na medida que poucas partições podem não capturar estruturas importantes (problema de *underfitting*), ao passo que árvores demasiadamente extensas tornam o algoritmo excessivamente dependente da amostra avaliada, e ter baixo poder representativo sobre dados fora da amostra (problema de *overfitting*).

Portanto, o tamanho da árvore é um hiperparâmetro de ajuste essencial para obter a eficiência do modelo, e deve ser definido em acordo com algumas regras de penalização. Uma estratégia popular para sua definição é o de podar o crescimento da árvore quando um número mínimo de nós é obtido.

Defina uma subárvore $T \subset T_0$ qualquer que pode ser obtida ao podar a árvore T_0, ao colapsar qualquer número de nós internos em sua composição. Dado a existência de m nós terminais, cada um representado por uma respectiva região R_m, seja $|T|$ o número de nós terminais na subárvore, têm-se:

$$N_m = \#\{x_i \in R_m\}$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

Determine o critério de custo da complexidade como:

$$C_\alpha(T) = \sum_{m-1}^{|T|} N_m Q_m(T) + \alpha|T|$$

E o objetivo é, para cada α, encontrar uma subárvore $T_\alpha \subseteq T$ que minimiza $C_\alpha(T)$. Note que parâmetro de ajuste $\alpha \geq 0$ pondera o *trade-off* entre o tamanho da árvore e sua qualidade de ajuste aos dados. Quanto maior for α, mais intensa será a penalização sobre modelos mais complexos, com maior número de nós, havendo uma propensão para a seleção árvores de decisão menores, e o inverso é válido.

Para cada α existe uma única subárvore T_α que minimiza $C_\alpha(T)$, em ordem de identificá-lo, o processo consiste em iniciar com a árvore original e iterativamente colapsar os nós internos que produzem o menor aumento em $\sum N_m Q_m(T)$ por nó e continue o processo até um único nó. A partir da subsequência de árvores obtidas, seleciona-se o de menor $C_\alpha(T)$.

4.2 Método de predição *random forest*

O modelo *random forest*, uma das ferramentas mais populares para realização de predições entre os algoritmos de *machine learning*, é uma técnica da classe dos métodos *ensembles*, os quais utilizam a combinação de resultados extraídos de um conjunto de modelos alternativos para assim extrair um novo resultado, que visa otimizar o poder preditivo obtido.

Do ponto de vista prático, o modelo é classificado como um método *ensemble Bagging*, onde os algoritmos são treinados de forma individual e independente e ao final são agregados através de algum método de ponderação.

Para construir tais predições, o modelo seleciona subconjuntos aleatórios de variáveis e as combina para construir os melhores preditores ajustados por meio do algoritmo árvore de regressão. O processo é realizado de forma iterada e utiliza o método de *bootstraping*, para gerar amostras independentes adotadas para treinar o algoritmo em cada árvore de regressão individual. Yoon (2020), sintetiza o procedimento básico do modelo da forma a seguir:

Etapa 1: para $m = 1,..M$**:**

i) Crie um conjunto amostral a partir da técnica *bootstrap*, de dimensão N a partir dos dados de treinamento;

ii) Construa uma *forest tree* aleatória, T_m, a partir dos dados gerados em (i) e repita as seguintes etapas para cada nó terminal da *forest tree* até atingir o mínimo tamanho de nó, n_min.

a) Selecione x variáveis aleatoriamente em um conjunto de p variáveis;

b) Colete a melhor variável e particione o ponto entre as x variáveis;

c) Particione o nó em dois nós subsequentes. A partição é definida de forma a minimizar o erro quadrático médio (MSE), calculado da forma:

$F_0(x) = \frac{1}{n}\sum(y_i - \gamma)^2$ Onde y_i é o valor observado e γ é o valor predito.

Uma limitação presente em algoritmos de classificação como o *regression tree*, é o problema de *overfitting*, que decorre do perfeito ajuste da árvore à subamostra selecionada, mas que não produz predições eficientes quando novos dados são adicionados. Em ordem de minimizar este problema, duas soluções podem ser implementadas no modelo *random forest*, quais sejam: realizar podas na árvore ao final do treinamento e/ou limitar o número de nós para divisão do algoritmo.

Etapa 2: Construa o produto do conjunto de árvores de regressões, $\{T_m\}_{\boldsymbol{m=1}}^{\boldsymbol{M}}$:

$$F_{rf}^{M} = \frac{1}{M}\sum T_m(x)$$

E o produto F_{rf}^{M} é calculado através da média do produto de todas as árvores. É importante notar que ao tomar a média de múltiplas predições reduz a variância e estabiliza a performance preditiva das árvores.

4.3 Validação cruzada

Um obstáculo relevante para a construção de um algoritmo de predição ótimo está no balanço entre impedir uma sensibilidade excessiva do modelo à base adotada para o treinamento (problema de *overfitting*), e restringir demasiadamente a etapa de treinamento de forma a gerar um modelo com baixo poder preditivo, resultando de pouco treinamento (problema de *underfitting*).

Os algoritmos de aprendizado de máquina são dependentes também de um número excessivo de hiperparâmetros (são os parâmetros que precisam ser definidos na etapa pré-

execução do modelo), os quais influenciam diretamente sobre a performance destes. No caso do modelo *random forest,* os hiperparâmetros a serem determinados são o número de nós em que a árvore pode ser particionada e a máxima profundidade da árvore.

Neste sentido, o processo de validação cruzada é uma técnica popular para selecionar um conjunto ótimo de hiperparâmetros. Esta técnica é baseada na divisão da base de dados de treinamento em $k - partes$, de forma a testar cada parte no ajuste do modelo. Seguindo a literatura, o número de partições é definido em $k = 10$, e a base de dados de treinamento é subdividida em 10 subconjuntos para treinar e ajustar o modelo. Entre os possíveis valores para o conjunto de hiperparâmetros, serão selecionados aqueles que produzem a menor média do EQM baseado no teste de 10 subconjuntos. O ajustamento dos hiperparâmetros será selecionado entre todas as combinações possíveis dos pares de valores (números de nós e profundidade máxima da árvore) pré-definidos e serão consideradas todas as variáveis do modelo.

5 BASE DE DADOS E ESTRATÉGIA EMPÍRICA

Estabelecer uma estimativa para o efeito da crise econômica ocasionada pelo advento da pandemia do Covid-19 sobre o mercado de trabalho dos estratos geográficos brasileiros é o objeto de estudo nesta dissertação. Este capítulo tem por objeto introduzir os dados que serão utilizados para conduzir a investigação e em seguida apresentar a estratégia empírica de modelagem, que permitirá obter uma estimativa contrafactual para a trajetória da população ocupada nos 146 estratos geográficos do Brasil, no cenário de ausência do choque econômico ocasionado pela pandemia do Covid-19.

5.1 Base de dados

Os dados que serão utilizados na execução do presente estudo podem ser segmentados em dois conjuntos distintos. No primeiro conjunto, serão consideradas variáveis relevantes para a predição da variação na população ocupada total (exceto administração pública), a fim de construir uma trajetória contrafactual desta variável para cada EG em 2020-2021, no cenário em que a crise da pandemia do Covid-19 não houvesse existido. Estas variáveis são apresentadas na Tabela 1:

Tabela 1 – Descrição da variável de resposta (Emprego) e das variáveis explicativas para o modelo de predição

Variável	Definição da variável	Período temporal	Fonte
Emprego	População ocupada no setor de mercado da economia	2012T1-2021T4	PNADC/IBGE
Emprego no setor industrial	População ocupada no setor industrial	2012T1-2021T4	PNADC/IBGE
Emprego no setor de serviços	População ocupada no setor de serviços	2012T1-2021T4	PNADC/IBGE
Especialização produtiva	Variável binária com valor igual à unidade para o subsetor com maior participação no emprego	2012T1-2021T4	PNADC/IBGE
Região	Variável binária indicando a região em que o estrato é localizado (Centro-Oeste, Norte, Nordeste, Sudeste, Sul)	2012T1-2021T4	PNADC/IBGE
População residente	Logaritmo natural da população residente projetada	2012T1-2021T4	PNADC/IBGE

Continua

Conclusão

Tabela 1 – Descrição da variável de resposta (Emprego) e das variáveis explicativas para o modelo de predição

Variável	Definição da variável	Período temporal	Fonte
Taxa de desemprego	Proporção de indivíduos com idade entre 15 e 64 anos que estão na força disponível de trabalho (PEA), mas não estão ocupados	2012T1-2021T4	PNADC/IBGE
Taxa de atividade	Proporção de indivíduos com idade entre 15 e 64 anos empregados e desempregados	2012T1-2021T4	PNADC/IBGE

Fonte: Elaborado pelo autor.

No segundo conjunto, a variável de resposta será dada pelo efeito da pandemia do Covid-19 sobre a população ocupada total (exceto administração pública), constituída a partir do modelo de predição. Já o vetor de variáveis explicativas considerará um conjunto de indicadores potencialmente associados à magnitude do impacto da pandemia sobre o mercado de trabalho dos estratos geográficos. A descrição dos dados é apresentada na Tabela 2:

Tabela 2 – Descrição das variáveis selecionadas para o modelo *regression tree*

Variável	Definição da variável	Período temporal	Fonte
Variação do Emprego durante a pandemia	Diferença entre a projeção contrafactual para o emprego entre 2020 e 2021 e o emprego observado.	2020T4 e 2021T4	PNADC/IBGE
Taxa de desemprego	Proporção de indivíduos com idade entre 15 e 64 anos que estão na força disponível de trabalho (PEA), mas não estão ocupados.	2020T4 e 2021T4	PNADC/IBGE
Taxa de mortalidade Covid por 100.000 habitantes	Total de óbitos notificados por Covid divido pela população, multiplicado por 100.000.	2020 e 2021	DATASUS
Taxa de casos Covid por 100.000 habitantes	Total de casos notificados de Covid divido pela população, multiplicado por 100.000.	2020 e 2021	DATASUS
Proporção de trabalhados com alto risco de agregação social	Número de empregos expostos à risco médio e alto de agregação social ponderados pela população ocupada.	2020T4 e 2021T4	PNADC/IBGE
Proporção de trabalhos com alto risco integrado	Número de empregos expostos à risco integrado médio-alto e alto ponderado pela população ocupada.	2020T4 e 2021T4	PNADC/IBGE

Continua

Conclusão

Tabela 2 – Descrição das variáveis selecionadas para o modelo *regression tree*

Variável	Definição da variável	Período temporal	Fonte
Proporção de trabalhadores em regime de tempo parcial	Número de trabalhadores com jornada semanal inferior à 30 horas semanais ponderado pela população ocupada.	2020T4 e 2021T4	PNADC/IBGE
Número de acidentes por 1.000 habitantes	Número de acidentes fatais em rodovias ponderadas pela população residente, multiplicado por 1.000.	2020 e 2021	DATASUS
Razão de dependência	Proporção da população com idade inferior à 15 anos ou idade superior à 64 anos.	2020T4 e 2021T4	PNADC/IBGE
Número de leitos hospitalares por 1.000 habitantes	Número de leitos hospitalares divido pela população residente, multiplicado por 1.000.	2012T1-2021T4	DATASUS
Proporção de trabalhadores atuando em ocupações do setor de saúde	Número de trabalhadores no setor de saúde divido pela população ocupada.	2020T4 e 2021T4	PNADC/IBGE

Fonte: Elaborado pelo autor.

5.2 Inferência causal e algoritmos de *machine learning*

A estratégia de modelagem clássica em pesquisas de inferência causal consiste na determinação na utilização de um grupo de controle (indivíduos, regiões ou instituições) que não sofreu exposição à um determinado tratamento para a identificação de uma trajetória contrafactual de uma variável de impacto para o grupo exposto ao tratamento. Neste cenário, desde que o grupo de controle apresente boa aderência ao grupo de tratamento em período pré-tratamento, e que o tratamento seja o único fator que causa divergência entre os grupos no período pós-tratamento, então a diferença entre a trajetória observada pelos tratados e a trajetória contrafactual, é uma estimativa consistente do impacto do tratamento sobre a população ou sobre os tratados, a depender da estratégia de identificação.

Entretanto, note que a existência de substrato populacional que não foi exposto ao tratamento é uma condição necessária para este tipo de modelagem, o que pode não ser verificado na ocorrência de uma série de ocasiões, o que expõe uma limitação teórica natural. A ausência de unidades de controle impossibilita a adoção de métodos de inferência causal

tradicionais – tais quais o modelo de controle sintético (Abadie *et al*., 2010) ou o modelo de diferenças em diferenças – para a identificação da trajetória contrafactual.

Diante desta dificuldade teórica, as técnicas de análise preditiva de *machine learning* (ML) têm ganhado relevância em pesquisas de inferência causal. Varian (2014, 2016) foi o precursor ao lançar luz de que a construção de um cenário contrafactual é uma tarefa preditiva e indicar o ML como ferramenta ideal para este processo.

Em uma estrutura de dados em painel, o autor destacar a possibilidade de adotar as observações pré-tratamento para gerar um grupo de controle artificial e gerar a tendência contrafactual futura no cenário de ausência de tratamento. Ao computar a diferença entre a trajetória observada e a trajetória contrafactual é obtido uma medida de efeito do tratamento sobre os tratados. A estratégia, nomeada *machine learning control method* (MLCM), ganhou especial atenção na análise regional de impacto da pandemia do Covid-19, com aplicações na área de saúde (Cerqua, 2021) e de mercado de trabalho (Cerqua; Letta, 2022).

Com base na rotina proposta por Cerqua *et al.* (2021), a construção de um cenário contrafactual para a trajetória do emprego nos estratos geográficos brasileiros será realizada em sete etapas:

i) Os dados pré-pandemia (2017Q1-2019Q4) são aleatoriamente divididos em uma amostra de treinamento (composto por 80% dos estratos geográficos) e a amostra de teste (um conjunto disjunto a amostra de treinamento, com os 20% estratos geográficos);
ii) A amostra de treinamento é usada para treinar o modelo *random forest* e então é aplicada a técnica de validação cruzada *10-fold*;
iii) Realiza-se a análise da performance preditiva fora-da-amostra do modelo com base na amostra de treinamento construída na primeira etapa;
iv) Testa-se a acurácia do modelo sobre a amostra completa para 2019 e compara-se sua performance preditiva com abordagens alternativas (modelo MQO tradicional, análise antes e depois);
v) A rotina é repetida para a amostra completa pré-pandemia, e compara-se novamente a performance do modelo com abordagens alternativas. Então, é realizada a predição para a amostra de 2020-2021;
vi) Deriva-se o efeito do tratamento para todos os estratos geográficos como a diferença entre os resultados observados para 2020-2021 e os resultados contrafactuais gerados pelo MLCM;

vii) É realizado o mapeamento dos efeitos individuais do tratamento em nível de estratos geográficos para inferir o impacto da pandemia do Covid-19.

Com o fim do processo de identificação da trajetória contrafactual, é possível estimar a influência do choque econômico originado pelo Covid-19 sobre o mercado de trabalho dos respectivos estratos geográficos no período 2020-2021, através da diferença percentual entre a população ocupada projetada na ausência do Covid-19 (cenário contrafactual) e a população ocupada observada (cenário real).

Posto isso, o segundo momento do estudo irá avaliar quais fatores são relevantes na predição da magnitude de impacto da pandemia sobre o mercado de trabalho local. Para isso, será aplicado o modelo *regression tree* sobre todo o espaço amostral, e serão identificados os preditores mais relevantes associados à variável de interesse (variação percentual da população ocupada no cenário real em relação ao cenário contrafactual).

6 RESULTADOS

6.1 Performance preditiva do modelo *random forest*

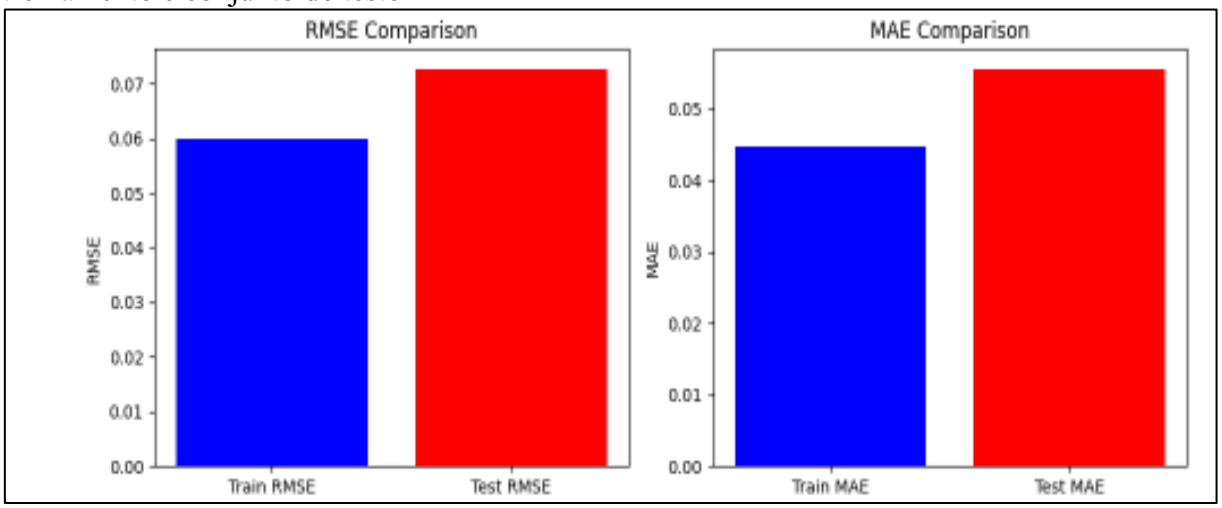
Com esta subseção, que abre a discussão dos resultados da análise empírica, apresenta-se as propriedades preditivas do modelo *random forest* em dois estágios sequenciais. O primeiro versa sobre o treinamento e validação do modelo preditivo, ao passo que o último traça um comparativo da capacidade preditiva do modelo calibrado vis-à-vis estratégias alternativas de *machine learning* e de econometria tradicional.

No que concerne à primeira etapa, após realizar a seleção do modelo *random forest* e o ajuste fino dos hiperparâmetros através da técnica de validação cruzada *10-fold*, com base no subconjunto de tratamento, foi avaliada a generalidade da especificação associada a partir da comparação do erro de previsão para o conjunto de treinamento e para o conjunto de teste.

Uma regra de bolso adotada na literatura consiste em estabelecer um valor limite de 30% para a perda de capacidade preditiva do modelo ao incorporar novas informações, as quais não foram utilizadas em seu processo de calibração (conjunto de teste). Para a avaliação da capacidade preditiva, foram adotadas as medidas de raiz do erro quadrático médio (RMSE) e do erro absoluto médio (MAE).

A Figura 1 e a Tabela 3 apresentam as estimativas de erro de previsão obtidas em cada caso considerando os dados pré-pandemia (2017-2019). Conforme observado, apesar de ocorrer um decréscimo na capacidade preditiva do modelo calibrado sobre o conjunto de teste em relação ao conjunto de treinamento, a queda de performance permaneceu dentro do limite estabelecido segundo ambos as medidas de erro de previsão, com um incremento na monta de 21,53% do erro de previsão com base no RMSE e de 24,41% com respeito ao MAE.

Figura 1 – Comparativo da Acurácia Preditiva do Modelo *Random Forest* sobre o conjunto de treinamento e conjunto de teste



Fonte: Elaborado pelo autor.

Tabela 3 – Perda de Capacidade Preditiva sobre o Conjunto de Teste

Diferença (RMSE)	**Δ% (RMSE)**	**Diferença (MAE)**	**Δ% (MAE)**
0.0129	21.53%	0.011	24.41%

Fonte: Elaborado pelo autor.
Nota: * A diferença no erro de medida i ($i = RMSE, MAE$) é igual ao erro de previsão sobre o conjunto de teste menos o erro de previsão sobre o conjunto de treinamento. Já a diferença parcial no erro de medida i ($i = RMSE, MAE$) é igual a diferença no erro de medida i sobre o erro de previsão sobre o conjunto de treinamento.

Após obter êxito no processo de validação, a etapa seguinte consiste na avaliação da performance preditiva do modelo selecionado em termos relativos a estratégias de estimação alternativas. Nesse caso, além da especificação ótima do modelo *random forest*, calibrou-se também o modelo *regression tree*, que consiste em um modelo de *machine learning* relativamente mais simples, além do modelo Lasso e do modelo de Mínimos Quadrados Ordinários (MQO). Neste caso, todos os modelos foram treinados com base nos dados de 2015 até 2018, ao passo que a previsão em cada modelo fora realizada para fora-da-amostra, considerando os quatro trimestres de 2019.

A Tabela 4 apresenta uma síntese dos resultados, os valores em percentual representam a proporção de incremento de erro de previsão com respeito ao modelo base (*random forest)*. De forma geral, confirma-se o prognóstico de melhor capacidade preditiva do modelo *random forest*, com o mesmo apresentando o menor erro de previsão segundo ambos os critérios. Com respeito ao RMSE, o modelo de regressão linear observou uma performance ligeiramente inferior (incremento de 3,1% no erro de previsão se comparado ao *random forest*), ao passo que o modelo *regression tree* magnitude de erro de previsão muito superior aos

demais. Ao considerar a medida MAE, há uma inversão entre o modelo Lasso e o MQO, com o primeiro tornando-se o segundo melhor modelo, enquanto o segundo apresentou um incremente relevante de erro de previsão (10,4%) com respeito ao *random forest.*

Tabela 4 – Comparação da Performance entre os Métodos de Previsão*

	RMSE	**Δ% ($RMSE$)**	**MAE**	**Δ% (MAE)**
Random Forest	0.065		0.048	
MQO	0.067	3,1%	0.053	10,4%
Lasso	0.069	6,1%	0.052	8,3%
Regression Tree	0.091	40,1%	0.069	43,7%

Fonte: Elaborado pelo autor.
Nota: * A diferença parcial no erro de medida i $(i = RMSE, MAE)$ é calculada a partir da razão entre a diferença do erro de previsão do método j $(j = MQO, Lasso, Regression\ Tree)$ e o método $random\ forest$, sobre o erro de previsão do método $random\ forest$.

6.2 Análise contrafactual da variação do emprego em 2020

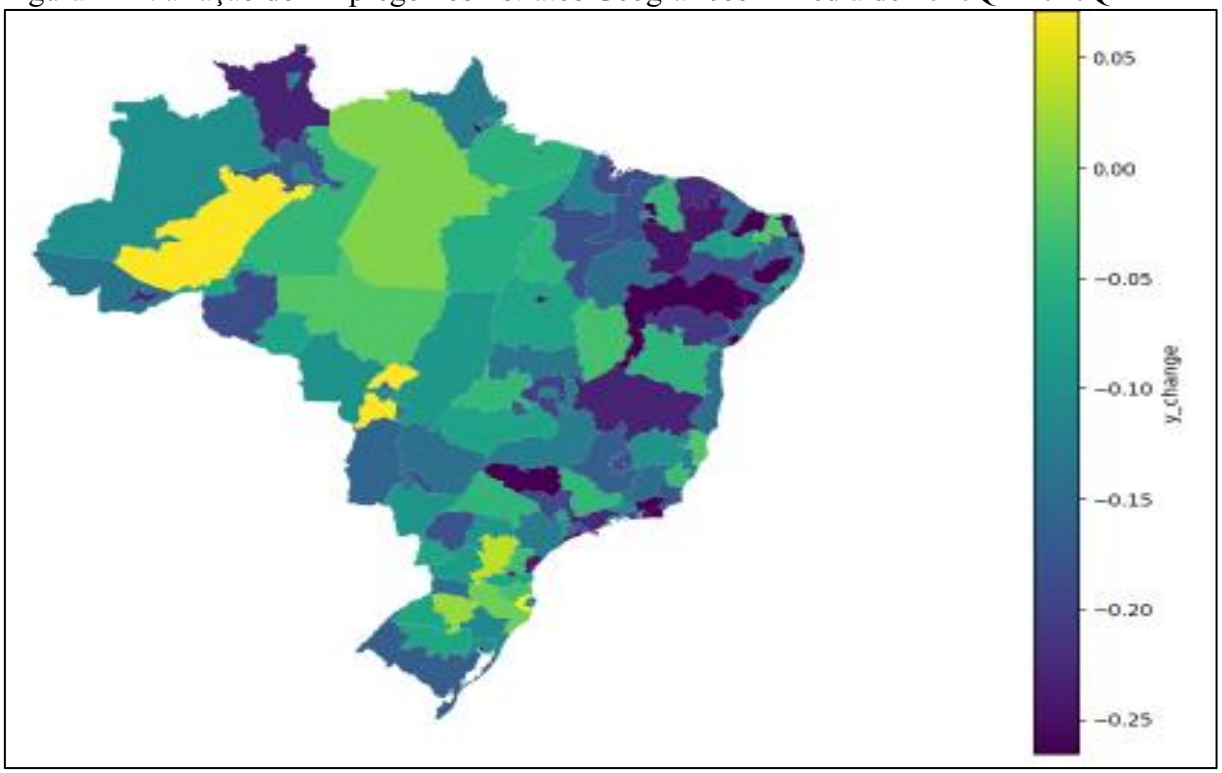
Determinado o *random forest* como modelo de previsão ideal para projeção da taxa de variação do emprego, nesta seção é apresentado o efeito médio do tratamento (pandemia do Covid-19) sobre as condições do mercado de trabalho enfrentadas em 2020 em nível de estratos geográficos. A variável de impacto, denominada daqui em diante de perda de emprego, é construída pela diferença entre a taxa de variação do emprego observada em 2020 e a taxa de variação de emprego contrafactual (que representaria a trajetória esperada em cada EG no cenário de inexistência da pandemia do Covid-19).

Em geral, ao comparar a trajetória observada e a sua contrapartida contrafactual, estima-se uma queda de 12,6 pontos percentuais (p.p.) no crescimento médio do emprego no Brasil ao longo dos três últimos trimestres de 2020. A Figura 2 exibe o mapa com a perda de emprego para os 146 EG, nota-se a existência de heterogeneidade tanto do ponto de vista inter-regional, com os EG do Nordeste apresentando uma queda relativamente mais intensa do emprego, quanto do ponto de vista intrarregional, existindo diferenças marcantes na perda de emprego mesmo entre localidades contíguas espacialmente.

Ao encontro do apontado, o efeito do tratamento nitidamente fora mais forte sobre os estratos geográficos localizados na Região Nordeste (redução média de 15,37 p.p. na variação do emprego), com intensidade crescente em EG não-litorâneos. Por outro lado, na comparação com o cenário contrafactual, o efeito foi menor na região Sul (queda de 7,2 p.p.). Discrepâncias no *gap* da variação do emprego em relação ao cenário contrafactual também

foram observados nas demais regiões, sugerindo a importância de questões específicas de cada localidade sobre o efeito da pandemia sobre o mercado de trabalho.

Figura 2 – Variação do Emprego nos Estratos Geográficos – Média de 2020Q2-2020Q4



Fonte: Elaborado pelo autor.

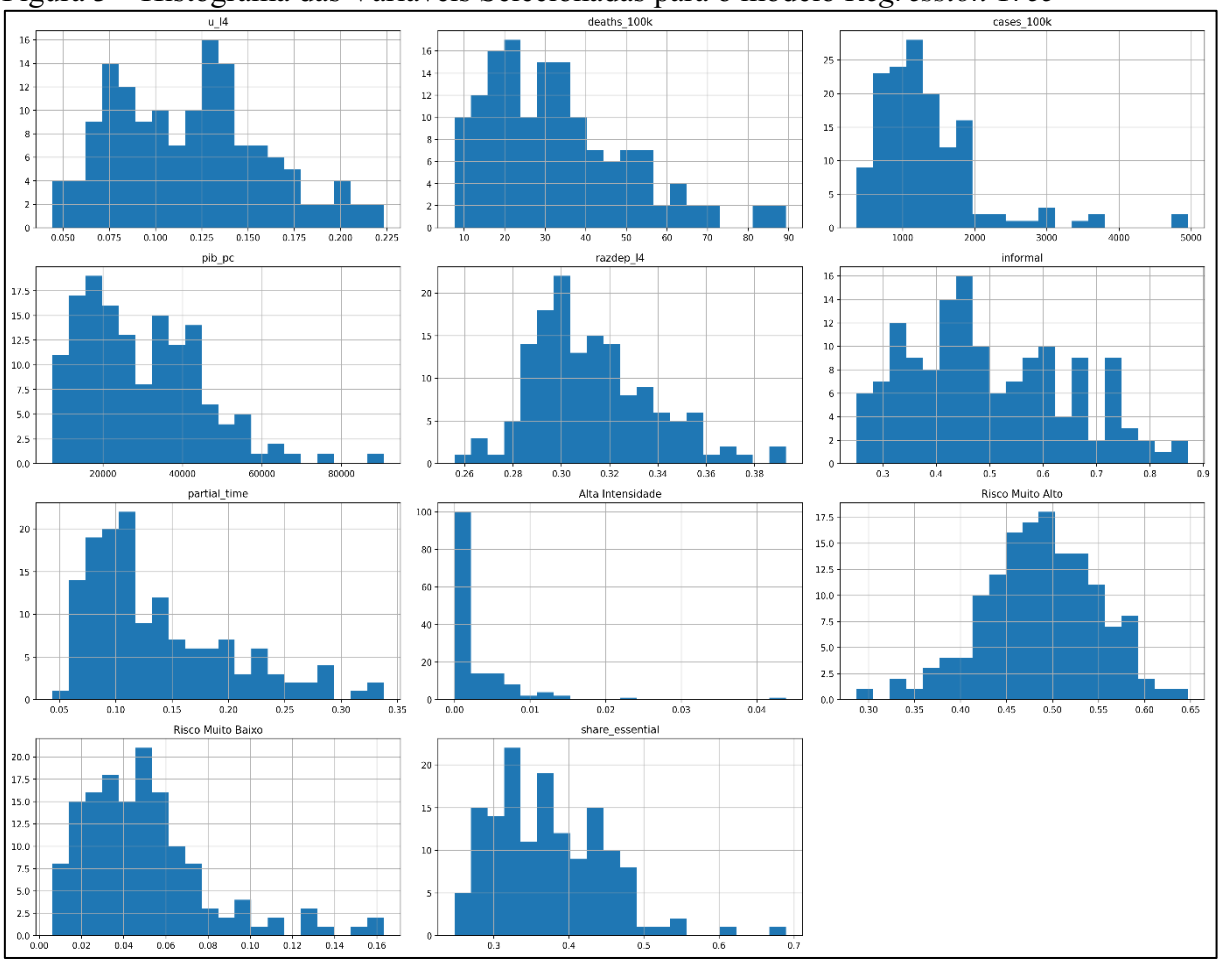
Haja visto que a intensidade da crise econômica originada pelo Covid-19 se distribuiu de forma heterogênea no espaço, a discussão sobre os fatores (sejam demográficos, econômicos ou epidemiológicos, por exemplo) associados a este padrão é uma questão natural de pesquisa.

A Figura 3 exibe um conjunto de histogramas subjacentes aos potenciais indicadores associativos selecionados. Antes de analisá-los, vale ressaltar à exceção das variáveis associadas ao Covid-19 (Total de Casos e de Óbitos por 100 mil habitantes), todas as demais variáveis foram calculadas com base nos dados de 2019, representando a condição do mercado de trabalho, de estrutura demográfica ou renda, por exemplo, em período anterior ao advento da pandemia do Covid-19. Neste sentido, o modelo irá buscar averiguar quais características das localidades atenuaram ou intensificaram o choque econômico gerado pela crise sanitária.

A exceção da variável de ocupações caracterizadas por intensidade tecnológica elevada (Alta intensidade), todas as demais apresentaram distribuições relativamente

alongadas, com a existência pontual de valores extremos na cauda direita das distribuições. A inspeção visual indica também grau elevado de assimetria à direita para as variáveis epidemiológicas – total de óbitos e casos por mil habitantes decorrentes do Covid-19 (*deaths_100k* e *cases_100k*, respectivamente), do PIB *per-capita* (*pib_pc*) e de características do mercado de trabalho, como a proporção de ocupações em tempo parcial (*partial_time*) e a proporção de ocupações com risco baixo de contaminação por Covid-19 (*Risco Muito Baixo*).

Figura 3 – Histograma das Variáveis Selecionadas para o modelo *Regression Tree*



Fonte: Elaborado pelo autor.

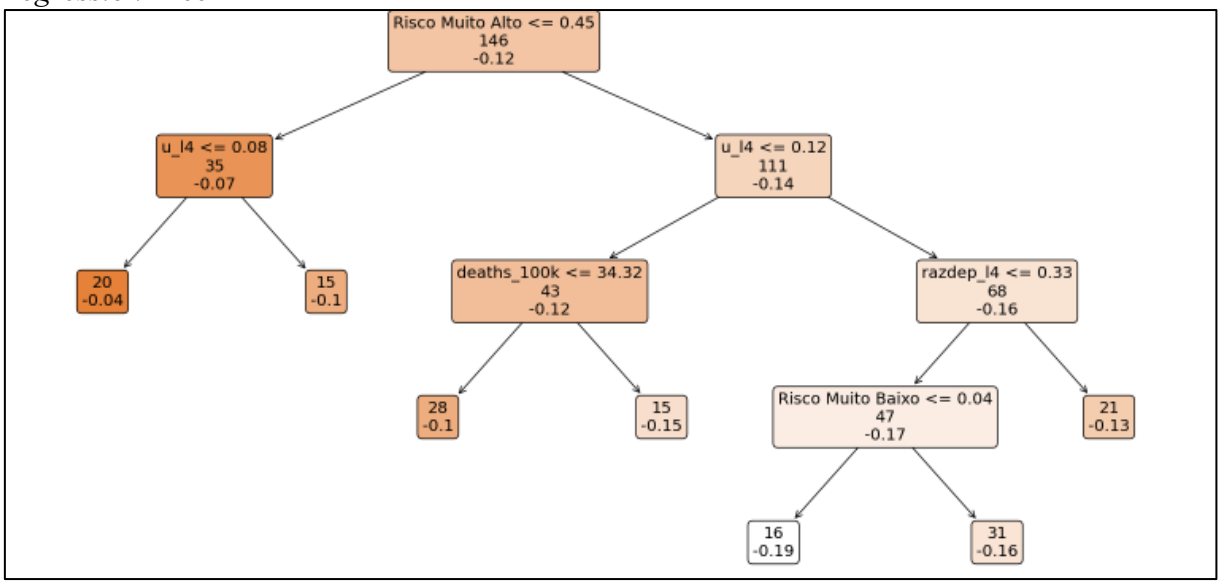
A Figura 4 reporta a análise associativa realizada a partir do modelo *regression tree* com base no conjunto de variáveis explicativas. Para fins de interpretação, três informações são apresentadas em cada casela da árvore de decisão, a do topo refere-se a variável discriminatória e o nível de segmentação, os nós à esquerda (direita) reportam a continuação da árvore em que a condição de desigualdade é atendida (não é atendida). Em seguida, no meio é indicado o número de estratos geográficos em cada casela de discriminação e por fim, a última linha apresenta o efeito médio do tratamento em cada casela (perda de emprego).

A árvore de regressão indicou que quatro variáveis associadas à intensidade da perda de emprego ao longo do ano de 2020, quais sejam o Risco de Exposição ao Covid (Risco Muito Alto, ou Risco Muito Baixo), a Taxa de Desemprego (u_l4), a Total de Óbitos por 100 mil habitantes em decorrência do Covid-19 ($deaths\_100k$) e a Razão de Dependência ($razdep\_l4$). Neste sentido, a heterogeneidade no impacto econômico da pandemia sobre as localidades foi associada a uma confluência de fatores de distintas dimensões, desde a esfera demográfica e epidemiológica até as condições pretéritas de fragilidade do mercado de trabalho.

Na margem, a árvore de regressão indicou que os EG que compartilhavam de proporção elevada de ocupações com risco elevado de contaminação (acima de 45%), nível de desemprego igual ou maior à 12%, razão de dependência inferior ou igual à 33% e proporção inferior à 4% de ocupações com risco baixo de contaminação, sofreram uma perda de emprego na casa de 19 p.p., que denota um valor 7 p.p. acima da média observada.

Em outro extremo, os EG com proporção inferior à 45% de ocupações com risco elevado de contaminação e com taxa de desemprego igual ou inferior à 8% em 2019, experenciaram uma ligeira perda de 4 p.p. na taxa de variação do emprego em relação à análise contrafactual.

Figura 4 – Determinantes da Variação do Emprego nos Estratos Geográficos – Modelo *Regression Tree*



Fonte: Elaborado pelo autor.

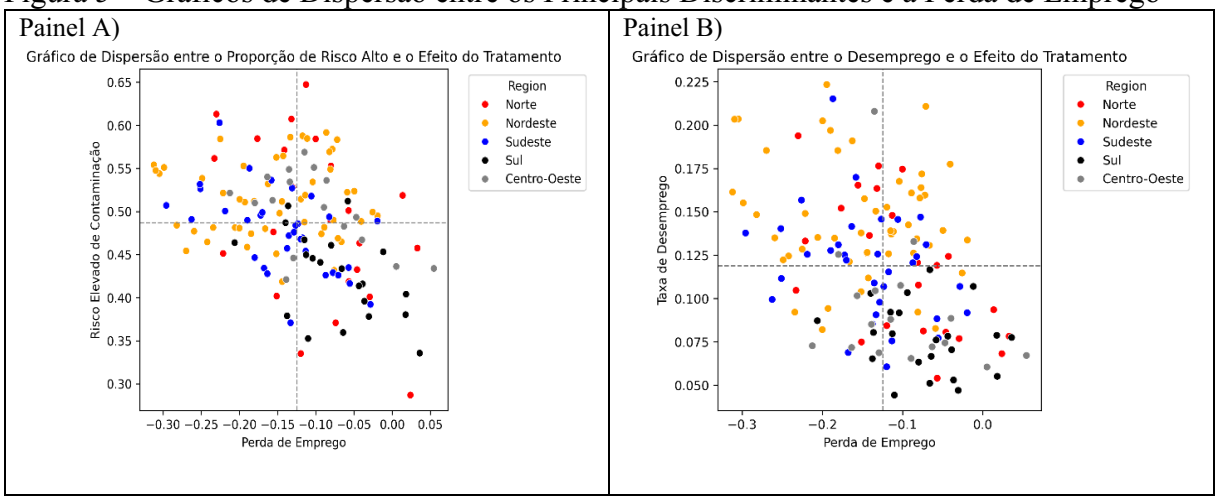
Estas evidências vão ao encontro ao reportado em estudos correlatos que avaliaram os efeitos da pandemia sobre o mercado de trabalho da Itália, país que foi assolado de forma severa pela primeira onda da pandemia do Covid-19 em 2020, ressaltando a exposição às

agregações e o risco de proximidade como principais fatores discriminantes do efeito do choque econômico do Covid-19 sobre o mercado de trabalho local (Barbieri *et al.*, 2021; Cerqua *et al.*, 2020).

A Figura 5 exibe os gráficos de dispersão entre os dois principais fatores associados à perda de emprego decorrentes da pandemia do Covid-19. Em ambos os casos, pode-se notar uma prevalência de estratos geográficos da Região Nordeste no quadrante de risco de contaminação (taxa de desemprego) acima da média, e perda de emprego acima da média (em termos absolutos). No outro espectro, nota-se uma forte concentração de estratos geográficos da Região Sul no quadrante de risco de contaminação (taxa de desemprego) abaixo da média e de perda de emprego abaixo da média (em termos absolutos).

Neste sentido, no caso do Brasil, acrescenta-se que a assimetria as condições mercado de trabalho ainda no período anterior ao advento da pandemia do Covid-19 como condicionante relevante para o efeito da pandemia sobre o próprio mercado de trabalho, retratando um ciclo vicioso. O nível de desemprego, especialmente em regiões menos desenvolvidas, permanecia elevado ainda em reflexo da crise econômica e fiscal de 2014-2016, com discrepâncias marcantes entre as regiões. Os estratos geográficos com desemprego elevado também eram caracterizados por elevado nível de informalidade, o que potencialmente elevou a exposição destes ao choque econômico.

Figura 5 – Gráficos de Dispersão entre os Principais Discriminantes e a Perda de Emprego



Fonte: Elaborado pelo autor.

7 CONSIDERAÇÕES FINAIS

Esta dissertação indicou que o efeito da pandemia do Covid-19 sobre o mercado de trabalho nacional se deu forma heterogênea no espaço. Ao comparar com a tendência contrafactual, ressaltou-se um efeito mais intenso sobre regiões que já apresentavam mercado de trabalho mais fragilizado em período prévio, contribuindo, assim, para um aumento nas inequidades regionais.

Ressalta-se também que o impacto da pandemia sobre o mercado de trabalho refletiu também condições não-econômicas, mas sim associadas especialmente à vulnerabilidade dos postos de trabalho com respeito ao risco de contaminação, bem como ao espraiamento da pandemia sobre o território local.

Neste sentido, dois apontamentos fazem-se interessantes, primeiro, no que concerne ao perfil de ocupações, os resultados não indicaram a proporção de ocupação consideradas essenciais como variável preponderante, mas sim a proporção de ocupações com risco de contaminação, tal resultado sugere que a queda mais intensa nestes locais pode ter sido gerado por uma combinação entre redução na demanda por trabalho (haja visto que estas ocupações associam-se aos setores mais expostos às restrições de operação realizadas à época) e na oferta de trabalho (na medida em que a percepção de risco elevado de contaminação, tornaria os trabalhadores menos propensos às ocupações).

Em seguida, visto que a variável de óbitos por Covid-19, e não a de casos de Covid-19, apresentou associação com o efeito do tratamento sobre o mercado de trabalho, abre-se a possibilidade para o argumento de que o grau de redução das atividades locais respondeu de forma relativamente mais intensa aos casos fatais, e não necessariamente ao espraiamento da pandemia em termos de ocorrência da doença. Entretanto, é importante destacar que ambas as variáveis são fortemente correlacionadas.

Por fim, do ponto de vista de políticas públicas, as evidências aqui obtidas ressaltam a necessidade de desenhos específicos do ponto de vista inter-regional (eixo Norte e eixo Sul) e intrarregional (áreas rurais e urbanas) para o combate de crises econômicas originadas por questões sanitárias. Haja visto que estas localidades apresentam condições distintas em termos de matriz produtiva, estrutura demográfica e grau de informalidade, faz-se necessário refletir sobre adoção de políticas não-horizontais, de forma a minorar os efeitos econômicas em acordo com as circunstâncias específicas.

REFERÊNCIAS

ABADIE, Alberto; DIAMOND, Alexis; HAINMUELLER, Jens. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. **Journal of the American statistical Association**, v. 105, n. 490, p. 493-505, 2010.

BLANCHARD, O. J.; SUMMERS, L. H. Hysteresis and the European unemployment problem. **NBER macroeconomics annual**, v. 1, p. 15-78, 1986.

BLUSTEIN, David L. et al. Unemployment in the time of COVID-19: A research agenda. **Journal of Vocational Behavior**, v. 119, 2020.

CERQUA, A.; DI STEFANO, R.; LETTA, M.; MICCOLI, S. Local mortality estimates during the COVID-19 pandemic in Italy. **Journal of Population Economics**, v. 34, n. 4, p. 1189-1217, 2021.

CERQUA, Augusto; LETTA, Marco. Local inequalities of the COVID-19 crisis. **Regional science and urban economics**, v. 92, 2022.

COSTA DIAS, Monica *et al*. The challenges for labour market policy during the Covid-19 pandemic. **Fiscal Studies**, v. 41, n. 2, p. 371-382, 2020.

COSTA, Simone da Silva. Pandemia e desemprego no Brasil. **Revista de Administração Pública**, v. 54, p. 969-978, 2020.

HASTIE, Trevor *et al*. **The elements of statistical learning:** data mining, inference, and prediction. New York: springer, 2009.

KAPICKA, Marek; RUPERT, Peter. Labor markets during pandemics. **Manuscript, UC Santa Barbara**, 2020.

KEYNES, J. M. **The General theory of employment, interest and money**. Republisheq by Harcourt Brace Jovanovich, 1936.

KOMATSU, B.K.; MENEZES-FILHO, N. **Simulações de impactos da COVID-19 e da renda básica emergencial sobre o desemprego, renda, pobreza e desigualdade**. São Paulo: Insper Centro de Gestão e Políticas Públicas, 2020. (Policy Paper, 43).

LAMBOVSKA, Maya; SARDINHA, Bogusława; BELAS JR, Jaroslav. Impact of the COVID-19 pandemic on youth unemployment in the European Union. **Ekonomicko-manazerske spektrum**, v. 15, n. 1, p. 55-63, 2021.

ORGANIZAÇÃO PAN-AMERICANA DE SAÚDE-OPAS. **Histórico da pandemia de COVID-19**. Disponivel em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19#:~:text=Em%2031%20de%20dezembro%20de,identificada%20antes%20em%20seres%20humanos>. Acesso em: 18 mai. 2023.

PHILLIPS, A. W. The relationship between unemployment and the rate of change of money wages in the United Kingdom 1861-1957. **Economica**, v. 25, n. 100, p. 283–299, 1958.

RODRÍGUEZ-CABALLERO, C. Vladimir; VERA-VALDÉS, J. Eduardo. Long-lasting economic effects of pandemics: Evidence on growth and unemployment. **Econometrics**, v. 8, n. 3, p. 37, 2020.

SIMONSEN, Mário Henrique. Salários, dualismo e desemprego estrutural. **Revista Brasileira de Economia**, v. 17, n. 4, p. 27-75, 1963.

_____. Cinquenta anos de Teoria geral do emprego. **Revista Brasileira de Economia**, v. 40, n. 4, p. 301-334, 1986.

SNOWDON, Brian; VANE, Howard R. **Modern macroeconomics:** its origins, development and current state. Edward Elgar Publishing, 2005.

VARIAN, Hal. Machine Learning and Econometrics. **Slides package from talk at University of Washington**, 2014.

VARIAN, Hal R. Causal inference in economics and marketing. **Proceedings of the National Academy of Sciences**, v. 113, n. 27, p. 7310-7315, 2016.

ZYLBERSTAJN, Hélio; NETO, Giácomo Balbinotto. As teorias de desemprego e as políticas públicas de emprego. **Estudos Econômicos**, São Paulo, v. 29, n. 1, p. 129-149, 1999.

ANEXOS

ANEXO A – QUADRO 1

Quadro 1 – Setores de Atividade e nível de risco de contaminação do Covid-19

Setores de Alto Risco: 1. **Alojamento e serviços de alimentação:** Apresentam um risco muito alto, pois frequentemente envolvem interação próxima com clientes e colegas de trabalho, tornando difícil manter o distanciamento social. 2. **Atacado e varejo, vendas, trabalho em lojas:** Outro setor com risco muito alto, já que os trabalhadores nessas áreas têm contato direto com clientes, aumentando o risco de exposição ao vírus. 3. **Serviços sociais e pessoais:** Esses serviços também são classificados como de alto risco devido à proximidade com pessoas, o que pode aumentar a exposição ao vírus.
Setores com Risco Variável (Alguns, Alto ou Moderado): 1. **Educação ou serviços de saúde:** Podem apresentar diferentes níveis de risco, dependendo das circunstâncias e práticas adotadas. Por exemplo, na área da saúde, os profissionais podem estar mais expostos devido ao contato próximo com pacientes infectados, enquanto na educação, a exposição pode variar dependendo do ensino remoto ou presencial. 2. **Agricultura, horticultura, silvicultura ou pesca:** Existe um risco variável, pois, embora não envolvam interação próxima com muitas pessoas, podem ter ambientes de trabalho compartilhados ou transporte em grupo. 3. **Indústrias culturais (artes, entretenimento) e transporte ou armazenagem:** São categorizados como apresentando risco variável, pois podem envolver diferentes níveis de contato e interação.
Setores com Baixo Risco: 1. **Serviços financeiros, seguros ou imobiliários, administração pública e apoio:** Esses setores são considerados de baixo risco, pois geralmente envolvem menor contato direto com o público ou com colegas de trabalho. 2. **Construção, manufatura:** São classificados como de baixo risco, pois os trabalhadores podem ter menos interação próxima durante suas atividades laborais.
Setores com Muito Baixo Risco: 1. **Fornecimento de gás ou eletricidade, mineração:** São considerados de muito baixo risco, provavelmente devido à natureza das atividades, que frequentemente não exigem interação física próxima. 2. **Serviços profissionais e científicos, tecnologia da informação e comunicação:** São setores que envolvem menor exposição ao contato físico direto, resultando em um nível de risco muito baixo em relação à exposição ao COVID-19.

Fonte: Pouliakas e Branka (2020).