CEAG E MS

GUIA PARA AVALIAÇÃO DE POLÍTICAS PÚBLICAS

Projeto: Estruturação de Monitoramento na gestão estratégica de projetos e cooperações do Ministério da Saúde (MS)

Centro de Estudos Avançados de Governo e Administração Pública
Universidade de Brasília

2026

CEAG-UnB Centro de Estudos Avançados de Governo e Administração Pública

UnB

BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA

SUS 35 ANOS

MINISTÉRIO DA SAÚDE

GOVERNO DO BRASIL DO LADO DO POVO BRASILEIRO

CEAG E MS

GUIA PARA AVALIAÇÃO DE POLÍTICAS PÚBLICAS

Projeto: Estruturação de Monitoramento na gestão estratégica de projetos e cooperações do Ministério da Saúde (MS)

Centro de Estudos Avançados de Governo e Administração Pública
Universidade de Brasília

2026

PROJETO DE PESQUISA
Estruturação de Monitoramento na gestão estratégica de projetos e cooperações do Ministério da Saúde (MS)

EQUIPE ACADÊMICA E TÉCNICA – UNIVERSIDADE DE BRASÍLIA (UNB)
Prof. Dr. Luiz Guilherme de Oliveira
Prof.a Dra. Doriana Daroit
Prof.a Dra. Susan Elizabeth Martins Cesar de Oliveira
Prof.a Dra. Gabriela Borges Antunes
Prof.a Dra. Fátima de Souza Freire
Prof. Dr. Arnaldo Mauerberg Junior
Prof. Dr. Antônio Nascimento Junior
Prof. Dr. Leonardo Cavalcanti da Silva
Prof. Dr. Paulo Carlos Du Pin Calmon
Prof. Dr. Roberto Góes Ellery Júnior
Prof. Dr. Victor Gomes e Silva
Prof. Dr. Wladimir Ganzelevitch Gramacho

ADMINISTRATIVO
Ms. Simone Braga Farias
Me. Wilson Alves Borba Junior

BOLSISTAS DO PROGRAMA DE DOUTORADO
Paula Daniella Prado Ramos
Solana Irene Loch Zandonai
Fabiana Bandeira dos Santos

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O objetivo desse documento é oferecer um conjunto de técnicas que possam ser utilizadas para avaliar políticas públicas, especialmente no âmbito do Ministério da Saúde. O público-alvo são técnicos do Ministério da Saúde que estejam envolvidos com avaliação de impacto das políticas executadas pelo ministério. Embora alguns aspectos do planejamento de uma avaliação de impacto sejam bastante técnicos, este documento busca fornecer ao leitor um arcabouço geral para participar com sucesso de uma ampla gama de avaliações de impacto.

As avaliações de impacto estimam o efeito de uma política sobre um conjunto de resultados de interesse. Exemplos incluem estimar o impacto de uma campanha de vacinação em massa na redução da incidência de doenças como sarampo ou poliomielite em uma determinada população, avaliar o impacto de uma campanha de conscientização sobre higiene (como lavagem das mãos) na diminuição de doenças diarreicas em comunidades vulneráveis e medir o efeito de um programa de distribuição gratuita de medicamentos para hipertensão ou diabetes na melhora da qualidade de vida e na redução de complicações associadas. O objetivo final é ser capaz de avaliar as mudanças no(s) resultado(s) de interesse que são atribuíveis ao programa em avaliação, para que possamos determinar pelo que o programa é responsável e

quão eficaz ele foi em alcançar seus objetivos. Dito de outra forma, o problema de avaliar o impacto de uma política é o problema da inferência causal.

A tarefa fundamental da inferência causal é determinar se uma relação de causa e efeito existe entre uma intervenção (como um programa ou tratamento) e um resultado observado. Isso envolve distinguir os efeitos diretamente atribuíveis à intervenção daqueles que poderiam ter ocorrido devido a outros fatores externos ou condições preexistentes. Em essência, busca-se responder à pergunta: "O que teria acontecido com os indivíduos ou a população se a intervenção não tivesse sido implementada?" Esse cenário contrafactual, porém, é impossível de observar diretamente, o que torna a inferência causal um desafio analítico complexo.

Para superar essa dificuldade, os métodos de inferência causal, como experimentos controlados randomizados ou técnicas estatísticas avançadas (como pareamento ou variáveis instrumentais), são utilizados para estimar o efeito causal. Esses métodos tentam criar uma comparação válida entre o grupo que recebeu a intervenção e um grupo equivalente que não a recebeu, isolando assim o impacto real da intervenção. O objetivo é garantir que as conclusões sobre a eficácia de um programa ou política sejam baseadas em evidências confiáveis, permitindo decisões informadas sobre sua continuidade, ajuste ou expansão.

A próxima seção trata das etapas preliminares ao processo de avaliação discutindo o processo de elaboração das questões e construção de indicadores. A terceira sessão apresenta vários métodos usados na literatura apontando vantagens e desvantagens de cada método, sem entrar em questões formais. A quarta seção discute os tópicos relativos à apresentação dos resultados da avalição. A quinta seção discute o método de diferenças em diferenças, o objetivo é fazer uma transição da discussão informal da terceira seção para a apresentação mais formal das seções seguintes. Em particular, a seção discute a equivalência entre cálculo de médias e uso de regressões lineares. Na sexta seção é feita uma apresentação formal do método de diferenças em diferenças para dado sem painel, a seção também discute o método de controle sintético. A sétima seção apresenta o estimado de diferenças em diferenças sintético e termina com um exemplo ilustrando a aplicação dos métodos de controle sintético, diferenças em diferenças e diferenças em diferenças sintético. A oitava seção apresenta as considerações finais.

Esse texto não tem como objetivo esgotar as questões técnicas envolvendo cada abordagem, para isso existe vasta literatura disponível. Em vez disso, o texto trata de apresentar estimadores usados na literatura e de decisões-chave que precisam ser tomadas durante a etapa de planejamento da avaliação de impacto, em particular nos papéis que tanto o responsável por encomendar a avaliação quanto o especialista técnico encarregado de conduzi-la devem desempenhar. As seções cinco a sete trazem abordagem mais técnicas relacionadas aos estimadores de controle sintético, diferenças em diferenças e diferenças em diferenças sintético.

2. ETAPAS PRELIMINARES

2.1 ELABORANDO AS QUESTÕES

A elaboração de questões relevantes para uma avaliação de impacto de políticas públicas utilizando inferência causal é um passo crucial para garantir que os resultados sejam úteis e informativos. O governo, como principal responsável pela formulação e implementação das políticas, desempenha um papel central ao definir os objetivos gerais da avaliação. Ele deve identificar claramente o que deseja compreender, por exemplo, se uma política de transferência de renda reduz a pobreza ou se um programa de saúde melhora indicadores de bem-estar. Essas questões precisam ser específicas, mensuráveis e alinhadas aos propósitos da política, refletindo as prioridades do poder público e as necessidades da sociedade.

Os especialistas encarregados de elaborar a avaliação, como pesquisadores ou analistas de dados, têm a responsabilidade de traduzir os objetivos do governo em perguntas tecnicamente viáveis para a inferência causal. Isso envolve determinar quais resultados de interesse (como taxas de escolaridade ou mortalidade) serão medidos e como o efeito causal da política será isolado de outros fatores. Eles devem propor questões

que permitam estimar o contrafactual, o que teria ocorrido na ausência da política, e sugerir métodos apropriados, como ensaios randomizados ou análises quasi-experimentais. O diálogo entre especialistas e governo é essencial para garantir que as questões sejam realistas, considerando limitações de dados e recursos disponíveis.

O público-alvo da política, ou seja, os indivíduos ou comunidades diretamente afetadas por ela, também deve ser considerado na formulação das questões. Suas experiências e perspectivas ajudam a identificar aspectos da política que podem não ser imediatamente evidentes para o governo ou especialistas. Por exemplo, uma questão relevante pode surgir ao perguntar se os beneficiários percebem mudanças em sua qualidade de vida ou acesso a serviços, o que pode direcionar a avaliação para resultados práticos e não apenas teóricos. Envolver o público-alvo, por meio de consultas ou pesquisas, assegura que as questões reflitam impactos reais e não apenas metas abstratas.

Além disso, as questões devem ser elaboradas de forma a equilibrar os interesses do governo, a expertise técnica dos especialistas e as necessidades do público-alvo. Por exemplo, enquanto o governo pode querer saber se uma política é custo-efetiva, os especialistas podem priorizar a robustez metodológica, e o público-alvo pode se preocupar com a acessibilidade da política. Uma questão bem elaborada, como "Qual é o impacto de um programa de capacitação profissional na empregabilidade de jovens em áreas urbanas vulneráveis?", consegue atender a essas diferentes demandas, fornecendo respostas úteis para todos os envolvidos.

Por fim, a construção dessas questões é um processo iterativo que exige colaboração contínua. O governo deve estar disposto a ajustar suas expectativas com base nas limitações apontadas pelos especialistas, enquanto estes precisam adaptar seus métodos às prioridades políticas e às vozes do público-alvo. O sucesso de uma avaliação de impacto depende de questões que não apenas permitam inferências causais sólidas, mas também gerem evidências práticas para aprimorar políticas públicas, beneficiando tanto os tomadores de decisão quanto a população atendida. Assim, o alinhamento entre esses três atores, governo, especialistas e público, é o fundamento para uma avaliação relevante e eficaz.

FIGURA 1
Atores relevantes para a avaliação de impacto

Fonte: elaboração própria.

2.2 CONSTRUÇÃO DOS INDICADORES

A construção de indicadores-chave de resultado para a avaliação de políticas públicas é um processo essencial para medir o sucesso e o impacto dessas iniciativas. Esses indicadores devem ser cuidadosamente elaborados para refletir os objetivos da política e permitir uma análise clara e objetiva. O governo, como principal formulador da política, precisa definir quais mudanças deseja observar, por exemplo, a redução da taxa de desemprego ou o aumento da cobertura vacinal. Para isso, os indicadores devem seguir características específicas, como o modelo SMART[1]: serem específicos (focados em um aspecto claro), mensuráveis (quantificáveis), alcançáveis (viáveis), relevantes (alinhados aos objetivos) e temporais (limitados a um período definido).

Uma característica essencial dos indicadores é sua consistência com a teoria que sustenta a política pública. Isso significa que eles devem estar fundamentados em uma lógica causal explícita, conhecida como "teoria da mudança", que explica como a intervenção deve gerar os resultados esperados. Por exemplo, se uma política de treinamento profissional visa aumentar a empregabilidade, o indicador-chave, como a porcentagem de participantes empregados após seis meses, deve refletir essa relação causal. Especialistas em avaliação, como estatísticos, economistas ou cientis-

1 Do inglês: Specific, Measurable, Achievable, Relevant and Time-bound.

tas sociais, desempenham um papel crucial ao garantir que os indicadores sejam teoricamente sólidos e tecnicamente robustos, evitando métricas vagas ou desconexas do propósito da política.

Outro aspecto importante é a necessidade de limitar o número de indicadores. Embora seja tentador medir muitos resultados para capturar todos os efeitos possíveis, um conjunto reduzido e bem selecionado facilita a coleta de dados, a análise e a comunicação dos resultados. O governo e os especialistas devem colaborar para priorizar indicadores que capturem os impactos mais significativos, evitando sobrecarga de informação. Por exemplo, em um programa de saúde materno-infantil, pode-se optar por apenas dois indicadores principais: a taxa de mortalidade infantil e a porcentagem de partos assistidos por profissionais, em vez de dezenas de métricas secundárias.

O público-alvo da política também influencia a construção dos indicadores, pois suas necessidades e realidades ajudam a definir o que é mais relevante. Um indicador SMART deve ser sensível às mudanças que afetam diretamente os beneficiários, por exemplo, em uma política de habitação, medir o número de famílias realocadas para moradias seguras é mais significativo para o público do que apenas o total de recursos gastos. Envolver representantes do público-alvo no processo, por meio de consultas ou grupos focais, pode garantir que os indicadores reflitam benefícios tangíveis e não apenas metas administrativas.

Por fim, a construção de indicadores-chave exige um equilíbrio entre ambição e praticidade. O governo pode buscar resultados amplos, mas os especialistas devem garantir que os indicadores sejam viáveis com os dados disponíveis e os prazos estabelecidos. Revisões periódicas são recomendadas para ajustar os indicadores conforme a política evolui ou novos desafios surgem. Quando indicadores bem elaborados, que respeitem as características SMART, são consistentes com a teoria e limitados em número, tornam-se ferramentas poderosas para avaliar o impacto de políticas públicas, oferecendo clareza para os tomadores de decisão e benefícios reais para a sociedade.

3. MÉTODOS DE AVALIAÇÃO

O conceito de contrafactual é fundamental na inferência causal, pois representa o cenário hipotético que descreve o que teria acontecido com um indivíduo ou grupo caso uma intervenção específica não tivesse sido aplicada. Em termos simples, é a resposta à pergunta: "Qual seria o resultado se a política ou programa não tivesse ocorrido?" Essa ideia é crucial porque, na realidade, só podemos observar o que aconteceu após a intervenção (o fato), mas não o que teria ocorrido sem ela (o contrafactual). Estabelecer esse cenário alternativo permite estimar o verdadeiro efeito causal de uma ação, isolando-o de outros fatores.

Na prática, como o contrafactual não pode ser diretamente observado, os pesquisadores utilizam métodos para aproximá-lo. Em experimentos controlados randomizados, por exemplo, o grupo de controle serve como uma *proxy* do contrafactual, pois é composto por indivíduos que não recebem a intervenção, mas são comparáveis ao grupo de tratamento em termos de características iniciais. Já em estudos observacionais, técnicas como pareamento (*matching*) ou modelos de regressão são empregadas para construir um contrafactual estimado, ajustando variáveis que poderiam influenciar o resultado. A qualidade dessa aproximação é o que determina a validade da inferência causal.

Um desafio central ao trabalhar com o contrafactual é garantir que ele seja realista e represente fielmente as condições que existiriam na ausência da intervenção. Isso exige que os pesquisadores controlem variáveis confundidoras, fatores externos que poderiam afetar o resultado independentemente da política ou programa avaliado. Por exemplo, ao avaliar o impacto de um programa de saúde na redução de doenças, mudanças econômicas ou sazonais podem influenciar os resultados. Se o contrafactual não levar esses fatores em conta, o efeito atribuído à intervenção pode ser superestimado ou subestimado, comprometendo a análise.

Em políticas públicas, o uso do contrafactual é especialmente valioso para justificar investimentos e ajustes em programas. Ao comparar os resultados observados com o cenário contrafactual estimado, os tomadores de decisão podem entender o que a política realmente mudou – por exemplo, quantos empregos foram criados por um programa de capacitação que não teriam surgido sem ele. Apesar de sua importância, a construção de um contrafactual confiável exige dados robustos, métodos rigorosos e, muitas vezes, suposições teóricas claras, o que torna a inferência causal uma tarefa complexa, mas indispensável para avaliações eficazes.

A utilização de experimentos controlados aleatórios na avaliação de políticas públicas enfrenta significativas dificuldades políticas e morais, que muitas vezes limitam sua aplicação. Politicamente, a randomização pode gerar resistência de gestores e comunidades, pois implica oferecer uma intervenção a apenas parte da população-alvo, enquanto outros são deliberadamente excluídos para formar o grupo de controle, o que pode ser percebido como injusto ou discriminatório. Moralmente, essa exclusão levanta dilemas éticos, especialmente quando a política envolve bens essenciais, como saúde, educação ou alimentação. Negar acesso a um grupo em nome da ciência pode ser considerado inaceitável, sobretudo se os benefícios da intervenção já são presumidos. Esses conflitos frequentemente levam à pressão por soluções alternativas, como implementações graduais ou estudos observacionais, que, embora menos rigorosos para inferência causal, atendem melhor às demandas de equidade e aceitação pública.

3.1 EXPERIMENTO CONTROLADO

O método de experimento controlado, frequentemente chamado de ensaio controlado randomizado (ECR), é uma abordagem amplamente utilizada para inferência causal. Consiste em dividir uma população em dois grupos: o grupo de tratamento, que recebe a intervenção (como um programa ou política), e o grupo de controle, que não a recebe. A alocação dos indivíduos a esses grupos é feita de forma aleatória, garantindo que as diferenças observadas nos resultados entre os grupos possam ser atribuídas à intervenção, e não a fatores pré-existentes. Esse método busca replicar o "contrafactual" usando o grupo de controle como referência.

As principais hipóteses estatísticas do ECR incluem a aleatoriedade da atribuição, que assegura que os grupos sejam comparáveis em características observáveis e não observáveis antes da intervenção. Outra hipótese é a independência entre os grupos, ou seja, o tratamento de um grupo não deve afetar o outro (conhecido como ausência de "efeitos de *spillover*"). Além disso, assume-se que a intervenção é a única diferença sistemática entre os grupos, permitindo que as diferenças nos resultados, como taxas de escolaridade ou saúde, sejam interpretadas como o efeito causal da política. Testes estatísticos, como o teste t ou regressões, são usados para verificar se essas diferenças são significativas.

Na prática, a implementação de ECRs em ciências sociais enfrenta diversas dificuldades. Uma delas é a questão ética: em políticas públicas, pode ser moralmente problemático negar a intervenção a um grupo de controle, especialmente se ela envolve benefícios essenciais, como acesso a saúde ou educação. Por exemplo, avaliar um programa de alimentação escolar excluindo algumas crianças pode gerar críticas e resistência. Isso leva à busca por alternativas, como listas de espera ou implementação gradual, mas essas soluções nem sempre mantêm a pureza do desenho experimental.

Outra dificuldade é a contaminação entre grupos, especialmente em contextos sociais onde indivíduos interagem. Se membros do grupo de controle são influenciados pelo grupo de tratamento, por exemplo, ao compartilhar informações ou recursos, o efeito isolado da intervenção fica comprometido. Em políticas públicas, como campanhas de conscientização, esse "efeito de *spillover*" é difícil de evitar, pois as men-

sagens podem se espalhar além do grupo-alvo. Isso exige ajustes no desenho do estudo ou o uso de unidades maiores, como comunidades inteiras, o que aumenta os custos e a complexidade.

A representatividade da amostra também é um desafio. Para que os resultados de um ECR sejam generalizáveis, a população estudada deve refletir as características do público-alvo da política em larga escala. No entanto, experimentos em pequena escala, como em uma única cidade, podem não capturar variações regionais ou culturais relevantes. Além disso, a implementação real de políticas públicas muitas vezes diverge do ambiente controlado de um experimento, introduzindo variáveis externas, como mudanças econômicas ou políticas, que afetam os resultados e dificultam a inferência causal.

Por fim, os custos e o tempo necessários para conduzir um ECR são obstáculos significativos. Em ciências sociais e políticas públicas, onde os efeitos podem levar anos para se manifestar (como em programas educacionais), os experimentos exigem recursos substanciais e paciência dos tomadores de decisão. Apesar dessas dificuldades, o ECR permanece um padrão-ouro para inferência causal quando bem implementado, pois oferece evidências robustas sobre a eficácia de intervenções, ajudando governos e especialistas a tomar decisões baseadas em dados sólidos, ainda que com adaptações contextuais.

3.2 MÉTODOS NÃO-EXPERIMENTAIS

Os métodos não experimentais para inferência causal são amplamente utilizados quando experimentos controlados randomizados não são viáveis, seja por questões éticas, logísticas ou financeiras, especialmente na avaliação de políticas públicas. Esses métodos, baseados em dados observacionais, incluem técnicas como o pareamento (*matching*), variáveis instrumentais, diferenças-em-diferenças (*difference-in-differences*) e regressão descontínua (*regression discontinuity*). O objetivo é estimar o contrafactual controlando variáveis confundidoras que poderiam influenciar o resultado. Por exemplo, ao avaliar o impacto de um aumento do salário-mínimo, o método de diferenças-em-diferenças compara a evolução do emprego em regiões afetadas e não afetadas antes e depois da política, isolando seu efeito.

Embora esses métodos ofereçam flexibilidade, sua principal dificuldade reside na dependência de suposições fortes e na qualidade dos dados disponíveis. No pareamento, por exemplo, é necessário assumir que todas as variáveis relevantes foram observadas e equilibradas entre os grupos comparados, o que nem sempre é realista, pois fatores não observáveis podem gerar vieses. Da mesma forma, o uso de variáveis instrumentais exige encontrar um instrumento válido, uma variável que afeta a intervenção, mas não o resultado diretamente, o que pode ser tecnicamente desafiador. Apesar dessas limitações, os métodos não experimentais são valiosos para analisar políticas já implementadas ou em contextos em que a randomização é impraticável, fornecendo evidências causais quando combinados com análises robustas e interpretações cuidadosas.

3.2.1 Estimador de Diferenças em Diferenças (DiD)

O estimador de diferenças em diferenças (*difference-in-differences*, ou DiD) é um método não experimental amplamente utilizado para inferência causal na avaliação de políticas públicas. Ele compara a evolução de um resultado de interesse ao longo do tempo entre dois grupos: um grupo afetado pela intervenção (tratamento) e outro que não foi (controle). A lógica do método é subtrair a diferença nos resultados antes e depois da intervenção no grupo de controle da diferença observada no grupo de tratamento, isolando assim o efeito causal da política. Por exemplo, ao avaliar o impacto de uma nova lei trabalhista sobre o emprego, o DiD analisa como as taxas de emprego mudam nas regiões onde a lei foi implementada em comparação com regiões onde não houve mudança, controlando tendências temporais comuns.

Uma das principais hipóteses do DiD é a “tendência paralela” (*parallel trends assumption*), que presume que, na ausência da intervenção, os grupos de tratamento e controle teriam seguido trajetórias semelhantes no resultado de interesse. Essa suposição é crucial, pois permite atribuir qualquer divergência pós-intervenção à política avaliada. Para verificar essa hipótese, os pesquisadores frequentemente analisam os dados pré-intervenção, examinando se as tendências nos dois grupos eram consistentes antes da mudança. Se essa condição não for atendida, talvez devido a diferenças econômicas preexistentes entre os

grupos, o estimador pode gerar conclusões enviesadas, superestimando ou subestimando o efeito da política.

O DiD é particularmente útil em políticas públicas porque aproveita dados longitudinais já disponíveis, como registros administrativos ou pesquisas populacionais, eliminando a necessidade de experimentos caros ou randomização. Uma aplicação comum é na avaliação de reformas educacionais: suponha que um estado introduza um programa de reforço escolar. O DiD pode comparar as notas dos alunos nesse estado com as de outro estado sem o programa, antes e depois da implementação, para estimar o impacto no desempenho escolar. Essa abordagem é flexível e pode ser ajustada com regressões para incluir variáveis de controle, como renda ou tamanho da escola, aumentando a precisão da análise.

Apesar de suas vantagens, o DiD enfrenta desafios práticos e teóricos. Um problema frequente é a presença de choques externos que afetam os grupos de maneira desigual após a intervenção, como crises econômicas ou mudanças legislativas paralelas, que podem confundir os resultados. Além disso, a escolha do grupo de controle é crítica: ele deve ser suficientemente semelhante ao grupo de tratamento em características relevantes, mas não afetadas pela política. Em contextos complexos, como políticas nacionais com múltiplos fatores em jogo, esses requisitos podem ser difíceis de satisfazer, exigindo testes de robustez e validação cuidadosa dos resultados.

Em resumo, o estimador de diferenças em diferenças é uma ferramenta poderosa para avaliar políticas públicas, oferecendo uma alternativa viável aos métodos experimentais quando a randomização não é possível. Suas aplicações vão desde a análise de impactos de subsídios agrícolas até a avaliação de programas de saúde, como a introdução de campanhas de vacinação. Com dados adequados e suposições bem fundamentadas, o DiD permite aos tomadores de decisão entenderem os efeitos reais de suas ações, contribuindo para o desenho de políticas mais eficazes e baseadas em evidências, ainda que demande atenção rigorosa às suas limitações metodológicas.

3.2.2 Estimador de Variáveis Instrumentais

O estimador de variáveis instrumentais (VI) é uma técnica estatística utilizada para lidar com o problema de endogeneidade em modelos de regressão, que ocorre quando uma variável explicativa está correlacionada com o termo de erro. Essa correlação pode surgir devido a omissão de variáveis relevantes, erro de medição ou simultaneidade entre as variáveis explicativas e a variável dependente. O método de VI resolve esse problema ao utilizar uma variável instrumental, que deve estar correlacionada com a variável explicativa endógena, mas não com o erro do modelo, permitindo a obtenção de estimadores consistentes e não viesados.

Na avaliação de políticas públicas, a endogeneidade é um desafio comum, pois muitas políticas não são implementadas de forma aleatória. Por exemplo, a alocação de recursos educacionais pode depender do nível socioeconômico da população, tornando difícil identificar o efeito causal do investimento na educação sobre os resultados dos alunos. O uso de VI pode ajudar a superar esse problema ao encontrar um fator externo que afete a variável explicativa sem estar diretamente relacionado ao resultado analisado. Um exemplo clássico é o uso da distância de uma escola como variável instrumental para medir o impacto da educação sobre os salários.

Outro exemplo de aplicação do estimador de VI em políticas públicas está na área da saúde. Suponha que um pesquisador queira estimar o efeito de um programa de vacinação sobre a taxa de mortalidade infantil. Se a adesão ao programa for influenciada por fatores individuais não observáveis, como preocupações com a saúde, a estimativa direta pode estar enviesada. Uma solução seria utilizar a distância até a unidade de saúde mais próxima como variável instrumental, assumindo que ela afeta a probabilidade de vacinação, mas não influencia diretamente a mortalidade infantil.

Apesar de sua utilidade, a escolha de uma boa variável instrumental é um dos principais desafios do método. Se a variável instrumental for fraca, ou seja, pouco correlacionada com a variável explicativa endógena, os resultados podem ser imprecisos. Além disso, a validade do instrumento precisa ser cuidadosamente testada, pois se houver alguma correlação com o termo de erro, os resultados podem ser enviesa-

dos. Dessa forma, o estimador de VI é uma ferramenta poderosa na avaliação de políticas públicas, mas seu uso requer uma escolha criteriosa de instrumentos e verificações estatísticas rigorosas.

3.3.3 Regressões com descontinuidade

A metodologia de Regressão com Descontinuidade (*Regression Discontinuity Design* – RDD) é uma técnica econométrica amplamente utilizada para estimar efeitos causais em situações em que há um critério de corte que determina a participação em um programa ou política pública. O princípio básico do RDD é que indivíduos muito próximos desse ponto de corte são comparáveis, permitindo que se avalie o impacto da intervenção como se fosse um experimento quase-aleatório. Isso ocorre porque aqueles logo acima e logo abaixo do limite enfrentam condições similares, exceto pela exposição ao tratamento, possibilitando uma inferência causal mais robusta.

O RDD pode ser aplicado em diversas áreas da avaliação de políticas públicas. Por exemplo, em programas educacionais, políticas que concedem bolsas de estudo ou acesso a escolas de qualidade com base em uma nota mínima em um exame são candidatas ideais para essa abordagem. Comparando o desempenho de estudantes que ficaram logo acima e logo abaixo do corte, é possível estimar o efeito da bolsa ou do acesso diferenciado na aprendizagem, empregabilidade e outros resultados de interesse. Como esses alunos tendem a ser muito semelhantes em características observáveis e não observáveis, a descontinuidade no desfecho pode ser atribuída ao programa em questão.

Outro campo de aplicação do RDD é na avaliação de programas sociais e de transferência de renda. Suponha que um benefício seja concedido apenas a famílias com renda abaixo de um determinado valor. Ao comparar indicadores socioeconômicos das famílias que ficaram ligeiramente abaixo e ligeiramente acima desse limite, pode-se medir o efeito do programa na redução da pobreza, no consumo ou no bem-estar dos beneficiários. Esse método é particularmente útil porque evita viés de seleção associado a decisões individuais de participação no programa.

Na área da saúde, a RDD tem sido usada para avaliar o impacto de políticas que estabelecem critérios objetivos para acesso a determinados tratamentos ou serviços médicos. Um exemplo é a elegibilidade para determinado medicamento gratuito baseada na idade ou em um

nível específico de gravidade da doença. Se os pacientes logo acima e logo abaixo desse limiar forem comparáveis, a diferença nos seus desfechos de saúde pode ser atribuída ao acesso ao tratamento, permitindo uma avaliação mais precisa da eficácia da política pública.

Apesar das vantagens, a metodologia de RDD também apresenta desafios. Um dos principais é garantir que os indivíduos não manipulem o critério de corte para se beneficiar da política, o que poderia comprometer a validade do experimento quase-natural. Além disso, a generalização dos resultados pode ser limitada, pois a estimativa obtida se aplica apenas aos indivíduos próximos ao ponto de corte e pode não ser representativa de toda a população elegível para a política.

Em conclusão, as regressões com descontinuidade são uma ferramenta poderosa para avaliação de políticas públicas, fornecendo estimativas causais robustas em contextos em que há um critério de elegibilidade bem definido. No entanto, sua aplicação exige atenção a possíveis manipulações do ponto de corte e à validade externa dos resultados. Quando bem implementada, essa abordagem permite que formuladores de políticas tomem decisões mais informadas sobre a efetividade de programas e intervenções governamentais.

3.3.4 Matching

O método de *matching* é uma técnica estatística amplamente utilizada na avaliação de políticas públicas para estimar efeitos causais quando não há um experimento aleatorizado. O objetivo principal desse método é comparar indivíduos que participaram de uma determinada política pública (grupo tratado) com indivíduos não participantes (grupo de controle), garantindo que ambos sejam o mais semelhantes possível em relação a características observáveis. Dessa forma, o *matching* busca reduzir o viés de seleção, permitindo uma comparação mais justa entre os grupos e uma melhor estimativa do impacto da política analisada.

Uma das principais vantagens do *matching* é sua capacidade de construir um grupo de controle mais apropriado, evitando comparações inadequadas entre indivíduos com características muito distintas. Ao parear indivíduos com base em variáveis como idade, nível de escolaridade, renda ou localização, o método assegura que diferenças nos resultados não sejam atribuídas a fatores externos, mas sim ao efeito

da política pública. Além disso, ao contrário de métodos experimentais, o *matching* pode ser aplicado em bases de dados observacionais, tornando-o uma alternativa viável quando experimentos aleatorizados não são possíveis ou éticos.

Outro ponto forte do método é a flexibilidade na escolha da técnica de pareamento. Existem diversas abordagens de *matching*, como *propensity score matching* (PSM), que pareia indivíduos com base na probabilidade de participação no programa, e *nearest neighbor matching*, que encontra indivíduos com características mais próximas dentro do grupo de controle. Essa diversidade de técnicas permite que pesquisadores escolham a abordagem mais adequada para seu contexto específico, melhorando a validade da análise.

Apesar de suas vantagens, o método de *matching* também apresenta limitações importantes. Uma das principais fragilidades é que ele só controla por características observáveis, ou seja, fatores não mensurados ou não disponíveis nos dados podem continuar gerando viés na estimativa do impacto. Isso significa que, se houver variáveis não observadas que influenciam tanto a participação na política quanto os resultados analisados, a estimativa ainda pode ser enviesada. Além disso, o *matching* pode ser sensível à especificação do modelo e à escolha das variáveis utilizadas para pareamento, exigindo um cuidado rigoroso na seleção de dados.

Por fim, outro desafio do *matching* é a necessidade de encontrar um número suficiente de indivíduos comparáveis no grupo de controle. Em alguns casos, pode ser difícil encontrar pares adequados para todos os participantes do programa, especialmente quando a política pública atende a um grupo muito específico. Isso pode reduzir a precisão dos resultados ou até mesmo impossibilitar a aplicação do método. Apesar dessas limitações, quando utilizado corretamente e combinado com outras técnicas, o *matching* é uma ferramenta poderosa para avaliação de políticas públicas, fornecendo estimativas mais confiáveis e auxiliando na formulação de intervenções mais eficazes.

3.3.5 Regressões múltiplas

As regressões múltiplas são amplamente utilizadas na avaliação de políticas públicas como uma ferramenta estatística para estimar re-

lações entre variáveis e medir o impacto de programas governamentais. Esse método permite analisar o efeito de uma variável explicativa sobre um resultado de interesse enquanto controla por outras variáveis que podem influenciar a relação. Por exemplo, ao avaliar o impacto de um programa de transferência de renda sobre o desempenho escolar, é possível incluir variáveis como idade, escolaridade dos pais e localização geográfica para isolar melhor o efeito da política em questão.

Uma das principais vantagens das regressões múltiplas é a sua flexibilidade e aplicabilidade a uma ampla gama de contextos e dados observacionais. Como o método permite controlar por diversas variáveis simultaneamente, ele ajuda a reduzir problemas de variáveis omitidas que poderiam enviesar os resultados. Além disso, as regressões múltiplas podem ser usadas com diferentes tipos de dados, incluindo séries temporais, dados em painel e cortes transversais, tornando-se uma ferramenta versátil na análise de políticas públicas.

Outra grande força do método é a facilidade de interpretação e a disponibilidade de técnicas para testar a robustez dos resultados. Estatísticas como o R^2, testes de significância e intervalos de confiança ajudam a avaliar a qualidade do modelo e a precisão das estimativas. Além disso, a inclusão de interações entre variáveis permite investigar se o impacto da política varia entre diferentes grupos da população, fornecendo informações mais detalhadas para a formulação de políticas mais direcionadas.

Apesar de suas vantagens, as regressões múltiplas apresentam limitações significativas. Uma das principais fraquezas é a dificuldade de estabelecer relações de causa e efeito apenas com base em correlações. Se uma variável explicativa estiver correlacionada com o erro do modelo, o estimador pode estar enviesado devido a problemas de endogeneidade. Isso pode ocorrer devido à omissão de variáveis relevantes, simultaneidade entre variáveis ou erro de medição, o que compromete a validade das conclusões.

Outro desafio das regressões múltiplas é a possibilidade de multicolinearidade, que ocorre quando duas ou mais variáveis explicativas estão altamente correlacionadas entre si. Isso pode tornar difícil distinguir os efeitos individuais de cada variável e levar a coeficientes instáveis e estatisticamente insignificantes. Além disso, a escolha inadequada de variáveis pode gerar modelos mal especificados, resultando em inferências equivocadas sobre o impacto da política pública analisada.

Por fim, a qualidade das estimativas obtidas em regressões múltiplas depende fortemente da qualidade dos dados disponíveis. Se os dados forem incompletos, imprecisos ou não representativos, os resultados podem ser distorcidos e levar a conclusões erradas. Portanto, embora as regressões múltiplas sejam uma ferramenta poderosa para a avaliação de políticas públicas, elas devem ser aplicadas com rigor metodológico, preferencialmente combinadas com outras abordagens para garantir maior validade e robustez na identificação de efeitos causais.

3.3.6 Controle Sintético

O método de controle sintético é uma técnica econométrica utilizada na avaliação de políticas públicas para estimar o impacto de uma intervenção quando não há um grupo de controle ideal disponível. Ele funciona criando uma combinação ponderada de unidades não tratadas (como estados, municípios ou países) para formar um "controle sintético" que simula o comportamento da unidade tratada caso a política não tivesse sido implementada. Essa abordagem tem sido amplamente utilizada para avaliar políticas como leis ambientais, reformas educacionais e programas de incentivo econômico.

Uma das principais vantagens do controle sintético é sua capacidade de construir um grupo de comparação mais apropriado do que métodos tradicionais, como diferenças simples ou diferenças-em-diferenças, especialmente quando a política pública afeta uma única unidade ou poucas unidades. Em vez de comparar diretamente a unidade tratada com um único grupo de controle, o método utiliza uma combinação de múltiplas unidades ponderadas para criar um contrafactual mais robusto. Isso reduz o viés causado por diferenças estruturais entre o grupo tratado e o controle.

Outra força do método é a transparência e a replicabilidade da construção do grupo sintético. Como os pesos atribuídos às unidades de controle são determinados de maneira formal e baseada em características observáveis antes da política ser implementada, há menor risco de manipulação subjetiva na seleção do grupo de comparação. Além disso, o método permite a visualização clara da evolução das tendências antes e depois da intervenção, facilitando a comunicação dos resultados para formuladores de políticas e a sociedade.

No entanto, o controle sintético também apresenta limitações. Um dos principais desafios é a necessidade de um conjunto amplo de unidades não tratadas para construir um controle adequado. Se houver poucas unidades disponíveis para compor o grupo sintético, o modelo pode não conseguir criar uma boa aproximação do contrafactual, tornando a estimativa pouco confiável. Além disso, o método é sensível à escolha das variáveis utilizadas para construir a combinação ponderada, podendo gerar estimativas instáveis caso variáveis importantes sejam omitidas.

Do ponto de vista estatístico, um dos principais problemas do controle sintético é a inferência. Como geralmente há apenas uma unidade tratada e um grupo sintético construído para comparação, as técnicas estatísticas tradicionais para estimar a incerteza dos resultados (como testes de significância) nem sempre se aplicam diretamente. Para contornar esse problema, muitas avaliações utilizam placebos ou testes de permutação para verificar a robustez dos resultados, mas a interpretação dos intervalos de confiança ainda pode ser desafiadora. Apesar dessas limitações, o controle sintético é uma ferramenta poderosa para avaliar políticas públicas em cenários onde outras abordagens não são viáveis, fornecendo estimativas mais confiáveis do impacto de intervenções governamentais.

3.4 OUTROS MÉTODOS

Algumas vezes não é possível usar os métodos descritos nessa seção. Isso decorre de fatores como limitações na coleta de dados, falta de condições para implementar um ensaio controlado randomizado, ou mesmo falta de tempo hábil para aplicar técnicas estatísticas apropriadas. Nesses casos alguns autores sugerem usar métodos que, apesar de limitados do ponto de vista estatístico, podem dar alguma indicação a respeito dos impactos das políticas analisadas.

3.4.1 Comparação “antes e depois”

O método de comparações antes e depois é uma abordagem simples e amplamente utilizada na avaliação de políticas públicas. Ele consiste em comparar os resultados observados antes da implementação de uma política com os resultados obtidos após sua execução, buscando identificar possíveis mudanças atribuídas à intervenção. Por exemplo, ao avaliar o impacto de um programa de capacitação profissional sobre o emprego, um pesquisador pode comparar a taxa de emprego dos participantes antes e depois do curso para verificar se houve melhora nos índices de empregabilidade.

Uma das principais vantagens desse método é sua facilidade de aplicação. Como ele não exige um grupo de controle, pode ser utilizado em diversas situações em que não há dados disponíveis para comparações com indivíduos ou localidades não afetadas pela política. Além disso, os dados necessários para essa análise geralmente estão disponíveis em registros administrativos ou pesquisas periódicas, reduzindo custos e tempo para a realização da avaliação.

Outra vantagem do método é sua utilidade na análise de impactos imediatos e no monitoramento da evolução de indicadores ao longo do tempo. Ele pode ser útil para gestores públicos que precisam de informações rápidas sobre os primeiros efeitos de uma política e para ajustes em sua implementação. Além disso, se houver dados coletados ao longo de vários períodos, pode-se observar tendências e verificar se a política teve um impacto duradouro.

Apesar de sua simplicidade, essa abordagem apresenta sérias limitações metodológicas. O principal problema é a dificuldade em estabelecer uma relação de causa e efeito entre a política pública e as mudanças observadas nos indicadores. Como muitos outros fatores podem influenciar os resultados ao longo do tempo, não há garantia de que a diferença observada antes e depois seja realmente causada pela intervenção e não por outros eventos simultâneos.

Um problema estatístico fundamental dessa abordagem é a falta de um grupo de controle, o que impede que se isole o efeito real da política. Se a economia estiver crescendo e a taxa de desemprego diminuir naturalmente, por exemplo, um programa de capacitação pode parecer eficaz mesmo que seu impacto real tenha sido pequeno. Esse

problema é conhecido como viés de fatores externos ou viés de tempo, pois eventos externos podem influenciar os resultados independentemente da política pública.

Outra questão estatística relevante é a reversão à média, que ocorre quando resultados extremos tendem a se aproximar da média ao longo do tempo, independentemente da intervenção. Isso pode levar a interpretações equivocadas sobre a eficácia de uma política. Por exemplo, se um programa de segurança pública for implementado após um pico de criminalidade, uma eventual queda nos crimes pode ocorrer naturalmente e não devido à política.

Por fim, o método de comparações antes e depois pode levar a conclusões equivocadas se houver mudanças no perfil da população ou na forma como os dados são coletados. Se os critérios de medição dos indicadores forem alterados ao longo do tempo, a comparação pode ser distorcida. Assim, embora seja um método acessível e útil para análises preliminares, ele deve ser usado com cautela e, sempre que possível, complementado por outras metodologias mais robustas, como o uso de grupos de controle ou métodos econométricos que controlem por fatores externos.

3.4.2 Diferença simples

O método de diferenças simples é uma abordagem utilizada na avaliação de políticas públicas que compara os resultados entre um grupo que recebeu a intervenção e um grupo que não recebeu. Essa comparação permite estimar o impacto da política pública analisando se há uma diferença nos indicadores entre os dois grupos após a implementação da política. Por exemplo, para avaliar um programa de incentivo à educação, pode-se comparar a taxa de conclusão escolar entre alunos que receberam e não receberam o benefício.

Uma das principais vantagens do método de diferenças simples é sua simplicidade e facilidade de interpretação. Ele permite obter uma estimativa direta do efeito da política sem a necessidade de dados históricos ou modelos econométricos sofisticados. Além disso, quando aplicado corretamente, pode oferecer uma primeira indicação do impacto da intervenção, sendo útil para gestores públicos que precisam de respostas rápidas sobre a efetividade de programas governamentais.

No entanto, esse método apresenta sérias limitações, especialmente no que diz respeito ao viés de seleção. Se os grupos tratados e não tratados forem sistematicamente diferentes antes da implementação da política, a estimativa obtida pode estar enviesada. Por exemplo, se apenas escolas de alta qualidade participarem de um programa de financiamento educacional, qualquer melhora no desempenho dos alunos pode ser mais resultado das características dessas escolas do que do financiamento em si. A falta de controle por essas diferenças pré-existentes pode levar a conclusões equivocadas sobre a efetividade da política.

Do ponto de vista estatístico, um dos principais problemas da abordagem de diferenças simples é a impossibilidade de separar corretamente os efeitos causais da política de outros fatores que podem influenciar os resultados. Variáveis não observadas, como diferenças culturais ou socioeconômicas entre os grupos comparados, podem afetar os resultados e gerar um viés de confundimento. Dessa forma, para que o método seja mais confiável, é essencial garantir que os grupos comparados sejam suficientemente semelhantes, o que pode ser feito com técnicas mais sofisticadas, como pareamento (*matching*) ou a abordagem de diferenças-em-diferenças, que controla melhor esses problemas.

4. APRESENTAÇÃO DOS RESULTADOS DA AVALIAÇÃO

Ao reportar uma avaliação de política pública baseada em inferência causal, é fundamental garantir que os resultados sejam apresentados de forma clara, precisa e transparente. Isso significa explicar a metodologia utilizada, justificar a escolha do método de inferência causal (como variáveis instrumentais, diferenças-em-diferenças ou regressão com descontinuidade) e discutir as principais limitações do estudo. O relatório deve incluir informações detalhadas sobre os dados utilizados, as fontes de informação e os critérios para a seleção da amostra, garantindo que os leitores possam compreender a robustez da análise.

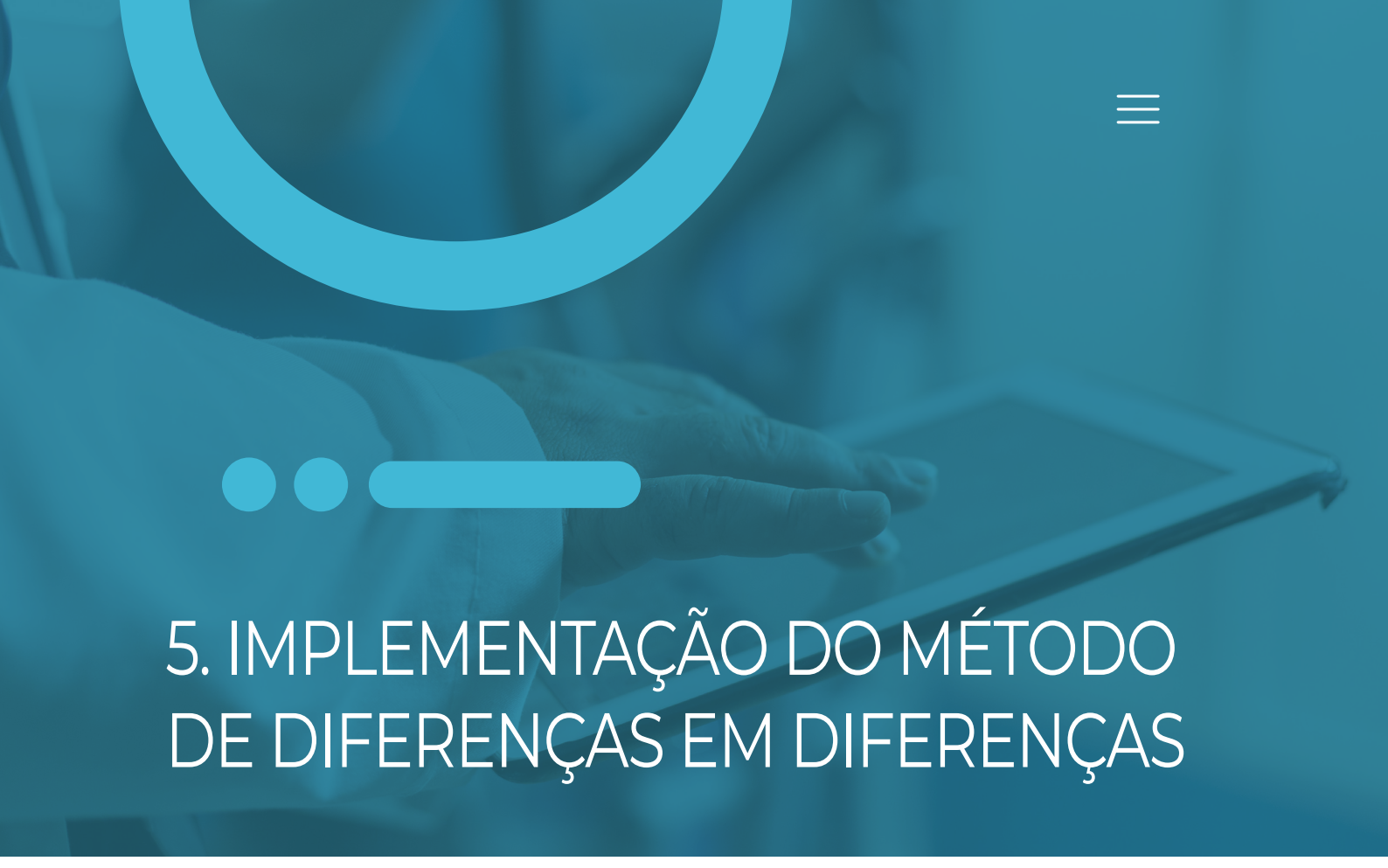
Uma consideração essencial ao comunicar os resultados é o cuidado com a interpretação das conclusões, especialmente no contexto político. Políticas públicas muitas vezes estão associadas a interesses governamentais e institucionais, e uma avaliação que sugira impactos negativos pode gerar resistência por parte de gestores e tomadores de decisão. Para evitar distorções, é importante apresentar os achados de maneira neutra e baseada em evidências, destacando tanto os efeitos positivos quanto os desafios ou impactos limitados da política. A transparência na comunicação dos resultados aumenta a credibilidade da avaliação e facilita o uso das evidências para a formulação de políticas mais eficazes.

Questões éticas também são fundamentais na divulgação de uma avaliação de política pública. Se a pesquisa envolveu dados sensíveis ou informações pessoais dos participantes, é essencial garantir a confidencialidade e a privacidade dos indivíduos analisados. Além disso, a interpretação dos resultados deve ser feita com responsabilidade, evitando conclusões precipitadas ou afirmações exageradas que possam levar a decisões equivocadas. Recomenda-se a adoção de diretrizes éticas reconhecidas, como aquelas estabelecidas por comitês de ética em pesquisa e organismos internacionais que regulam estudos com impacto social.

Outro ponto crucial na comunicação dos resultados é a consideração dos efeitos sobre o público-alvo da política. Uma avaliação que evidencia impactos negativos pode gerar preocupações e resistência entre os beneficiários, especialmente se houver risco de cortes ou modificações na política em questão. Assim, ao reportar os achados, é importante contextualizar os resultados e, quando possível, sugerir ajustes ou melhorias na política, em vez de apenas apontar falhas. Esse cuidado pode contribuir para um debate mais construtivo sobre a efetividade da intervenção, facilitando sua aceitação e aprimoramento.

Além disso, ao apresentar os resultados para diferentes audiências é recomendável adaptar a linguagem e os formatos de comunicação. Para gestores públicos, um sumário executivo com os principais achados e recomendações práticas pode ser mais eficaz do que um relatório técnico detalhado. Já para o público acadêmico, um relatório mais aprofundado, com metodologia e análise estatística detalhada, pode ser necessário. A clareza na comunicação ajuda a garantir que os resultados sejam compreendidos e utilizados de forma apropriada.

Por fim, a avaliação de políticas públicas deve sempre estar alinhada ao objetivo de contribuir para a melhoria das ações governamentais e para o bem-estar da população. Relatar os resultados com rigor metodológico, responsabilidade ética e sensibilidade política permite que a inferência causal seja utilizada de maneira produtiva, promovendo o aprimoramento das políticas públicas e a tomada de decisões baseadas em evidências.



5. IMPLEMENTAÇÃO DO MÉTODO DE DIFERENÇAS EM DIFERENÇAS

O método consiste em comparar a diferença entre a média da variável de interesse entre os grupos de tratamento e controle antes e depois da implementação da política. Seja $Y_{i0}|T$ o valor observado da variável de interesse, Y, para a unidade i que pertence ao grupo de tratamento. Seja $Y_{i0}|C$ o valor da variável de interesse para a unidade i do grupo de controle, que não recebe o tratamento, no período anterior à implementação da política. Finalmente sejam $Y_{i1}|T$ e $Y_{i1}|C$ os valores da variável de interesse para os grupos de tratamento e controle após a implementação da política.

O método de diferenças em diferenças consiste em calcular diferença entre a média da variável de interesse antes e depois da implementação da política para cada um dos grupos e depois calcular a diferença entre a diferença das médias de cada grupo. Especificamente calcula-se:

Diferença da média da variável de interesse antes e depois da política para o grupo de tratamento:

$$\bar{Y}_1|T - \bar{Y}_0|T$$

Diferença da média da variável de interesse antes e depois da política para o grupo de controle:

$$\bar{Y}_1|C - \bar{Y}_0|C$$

Diferença entre as duas diferenças acima:

$$Efeito\ do\ tratamento = (\bar{Y}_1|T - \bar{Y}_0|T) - (\bar{Y}_1|C - \bar{Y}_0|C).$$

O efeito do tratamento pode ser estimado por meio de regressão. Para isso é necessário definir uma variável *dummy*[2] para tratamento, a variável será igual a um se a unidade for do grupo de tratamento e zero se o estudante for do grupo de controle, e uma variável *dummy* para o tempo, um para o período posterior à política e zero para o período anterior à política. Chame a primeira variável *dummy* de I(trat) e a segunda de I(temp). A equação estimada será da forma:

$$Y_{it} = \beta_0 + \beta_1 I(trat_{it}) + \beta_2 I(temp_{it}) + \beta_3 I(trat_{it}) \times I(temp_{it}) + \varepsilon_{it}$$

A partir desta equação as médias condicionais podem ser calculadas como:

$$E(Y_{i1}|T) = E(I(trat_{ti}) = 1, I(temp) = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$
$$E(Y_{i0}|T) = E(I(trat_{ti}) = 1, I(temp) = 0) = \beta_0 + \beta_1$$
$$E(Y_{i1}|C) = E(I(trat_{ti}) = 0, I(temp) = 1) = \beta_0 + \beta_2$$
$$E(Y_{i0}|C) = E(I(trat_{ti}) = 0, I(temp) = 0) = \beta_0$$

Desta forma o efeito do tratamento, ou a diferença da diferença (DiD), pode ser estimado como:

$$DiD = \big(E(Y_{i1}|T) - E(Y_{i0}|T)\big) - \big(E(Y_{i1}|C) - E(Y_{i0}|C)\big)$$
$$DiD = (\beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 + \beta_1) - (\beta_0 + \beta_2 - \beta_0)$$
$$DiD = \beta_3$$

2 Uma variável ***dummy*** é aquela que captura deslocamentos no modelo quantitativo. Quantitativamente ela pode assumir valores de zero ou um.

Repare que o estimador “antes e depois” pode ser calculado como:

$$E(Y_{i1}|T) - E(Y_{i0}|T) = \beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1) = \beta_2 + \beta_3$$

Por sua vez, o estimador de diferença simples pode ser calculado como:

$$E(Y_{i1}|T) - E(Y_{i1}|C) = \beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_2) = \beta_1 + \beta_3$$

Com esse procedimento podemos estimar o efeito do tratamento, realizar testes de significância e construir intervalos de confiança usando análise de regressão. A Tabela 1 resume os passos para o cálculo da diferença em diferença:

TABELA 1
Cálculo do estimador DiD

	Período posterior	Período anterior	Diferença
Tratamento	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Controle	$\beta_0 + \beta_2$	β_0	β_2
Diferença	$\beta_1 + \beta_3$	β_1	β_3

Fonte: elaboração própria.

Para ilustrar a aplicação do método DiD bem como das etapas que permitem realizar a comparação entes e depois e o método de diferenças simples será o usado o exemplo em Card e Krueger (1994). O artigo usa o método DiD para avaliar o impacto do aumento do salário-mínimo em Nova Jérsei (NJ) e é uma das principais referências para análise de causalidade com dados observados.

Em 1992 o estado americano de NJ aumentou o salário-mínimo de 4,25 dólares por hora para 5,05 dólares por hora. A questão é saber se esse aumento levou a uma queda no emprego, especificamente no

emprego de quem recebe valores próximos do salário-mínimo. Para responder essa questão os autores usaram dados de empregos em cadeias de *fast-food*. Especificamente, foram utilizados dados de salário e de emprego em tempo total e em tempo parcial para 291 cadeias de *fast-food* em Nova Jérsei.

Para começar vale checar se o aumento de salário-mínimo de fato afetou as lojas de fast-food, ou seja, se as lojas pagavam menos que o novo salário-mínimo e se passaram a pagar um valor igual ou maior após o aumento. Os dados mostram que antes do aumento apenas 9% dos salários nessas lojas eram menores do que 5,05 dólares por hora e depois do aumento 99,7% passaram a ser maiores ou iguais a esse valor. Desta forma, é possível concluir que a lei teve impacto.

O próximo passo é avaliar se por ter de pagar um maior salário as cadeias de *fast-food* ajustaram o emprego, especificamente queremos saber se os restaurantes trocaram empregados de tempo integral por empregados de tempo parcial para reduzir custos. O método antes e depois consiste em comparar a variável de interesse, proporção de empregados em tempo integral no total de empregados, antes da mudança e depois da mudança.

Os dados mostram que a proporção de empregados de integral era de 29,7% antes do aumento do salário-mínimo e passou a ser 32,1% depois do aumento, ocorreu um aumento na proporção de emprego. A diferença de 2,4% é a estimativa do efeito da política pelo método antes e depois.

Naturalmente teríamos de avaliar se esse efeito é significativo do ponto de vista estatístico antes de concluir que o aumento do salário-mínimo levou a um aumento no emprego em tempo integral, essa questão será discutida mais à frente. Antes tratemos de uma questão mais relevante para inferência causal. Como garantir que aumento na variável de interesse, proporção de empregados em tempo integral, ocorreu por conta da política, aumento do salário-mínimo, ou por conta de outros fatores?

É possível que no período os Estados Unidos estivessem passando por um período de expansão da economia e que o emprego em tempo integral fosse ainda maior na ausência do aumento do salário-mínimo. Essa é a principal limitação de comparar antes e depois, fica difícil determinar se o efeito é da política ou de outras variáveis. Podemos tentar incorporar outras variáveis no modelo, mas sempre existem variáveis

não observadas e não podemos descartar o risco de esquecer alguma variável. Vamos explorar esses problemas ilustrando o método de comparação simples.

Suponha que em vez de comparar o antes e depois de Nova Jérsei nossa estratégia seja comparar o emprego em tempo integral em Nova Jérsei e na Pensilvânia (PA), um estado vizinho com áreas urbanas próximas à Nova Jérsei e que não ocorreu aumento no salário-mínimo, depois da política de aumento do salário-mínimo.

Considerando o período posterior à política a proporção de empregados em tempo integral na Pensilvânia era de 27,2% e em Nova Jérsei era de 32,1%, a diferença é de 4,9%. Essa comparação simples da variável de interesse no grupo tratado, NJ, e no de controle, PA, sugere que o aumento do salário-mínimo não causou redução de contratações em tempo integral. Se as escolhas do grupo de tratamento e controle tivesse sido aleatória, como ocorre em um experimento controlado randomizado, essa diferença seria um bom estimador do efeito da política, mas não é o caso. É possível que existam outros fatores que expliquem a diferença no emprego entre PA e NJ.

Um desses fatores é a presença de cada cadeia em cada estado. Na amostra da Pensilvânia, o *Burger King* responde por 46,3% das lojas e em Nova Jérsei responde por 40,5%. É possível que a diferença observada entre emprego de tempo integral nos dois estados seja por conta de uma prática específica do *Burger King* de trabalhar com mais empregados em tempo parcial. Variáveis desse tipo, que afetam a variável de interesse estão presentes antes do tratamento são chamadas de confundidoras e dificultam sobremaneira a inferência causal. É por conta da existência dessas variáveis que aparece a máxima de que correlação não é causalidade.

No caso das cadeias é possível avaliar o impacto delas na estimativa avaliando a proporção do emprego em tempo integral em cada cadeia, isso se chama controlar pela variável. A Tabela 2 mostra a variável de interesse em cada cadeia e em cada estado.

TABELA 2

Proporção do emprego em tempo integral por cadeia de *fast-food*

Cadeia	NJ	PA	Diferença
Burger King	35,8%	32,1%	3,7%
KFC	32,8%	23,6%	9,2%
Roys	28,3%	21,3%	7%
Wendys	26%	24,8%	1,2%

Fonte: elaboração própria.

Repare que em todas as cadeias a proporção de empregos em tempo integral é maior em Nova Jérsei do que na Pensilvânia, isso sugere que cadeias não são um problema na análise. Porém é possível imaginar outras variáveis que expliquem a diferença, questões específicas da economia de cada estado, questões legais ou mesmo aspectos culturais podem influenciar a variável de interesse. O fato é que a presença de variáveis confundidoras comprometem avaliações que tomam por base a diferença simples entre o grupo de tratamento e controle quando a distribuição do tratamento não é aleatória.

Em resumo, a estratégia de comparar antes e depois pode ser comprometida por fatores externos que mudam no tempo. A estimativa por diferenças simples não depende do tempo, mas pode ser comprometida por fatores externos que diferenciem o efeito do tratamento nos dois grupos, ou seja, diferenças pré-tratamento podem influenciar o resultado da análise. O estimador de diferenças em diferenças aperfeiçoa o estimador antes e depois e o de diferenças simples tentando eliminar o viés causado por confundidores que mudam no tempo. A hipótese chave é que na ausência do tratamento a variável de resultado no grupo tratado seguiria uma tendência paralela a observada no grupo de controle.

No exemplo, a estratégia DiD consiste em supor que na ausência da elevação do salário-mínimo (tratamento), o emprego em tempo integral nos restaurantes de NJ seguiriam a mesma tendência observada na PA. Para estimar o DiD é preciso calcular a diferença antes e depois

nas unidades tratadas, no exemplo NJ, e na unidade de controle, no exemplo PA, e depois calcular a diferença dessas diferenças.

O primeiro passo para calcular o estimador de diferenças em diferenças é calcular a diferença antes e depois na unidade tratada e na unidade de controle. No exemplo a diferença entre a proporção de empregos em tempo integral entes e depois do aumento de salário-mínimo em Nova Jérsei será calculada nos dois estados. Na Pensilvânia a diferença foi de -3.8%, em Nova Jérsei, como já vimos, foi de 2,4%. Note que na Pensilvânia foi observada uma queda na proporção de empregados em tempo integral. O segundo passo é calcular a diferença entre essas duas diferenças, desta forma DiD = 2,4% - (-3,8%) = 6,2%. Ainda temos que discutir a significância estatística desses estimadores, mas o resultado sugere mais uma vez que o aumento do salário-mínimo não causou uma queda na proporção de empregos em tempo integral.

O estimador DiD se torna problemático quando temos razões para desconfiar que a tendência das unidades tratadas na ausência do tratamento (contrafactual) não seria paralela a tendência observada nas unidades de controle. No exemplo, seria o caso se fosse registrado um fenômeno que só ocorreu na PA e que poderia afetar o emprego em tempo integral.

Não é fácil determinar se essas tendências seriam ou não paralelas na ausência do tratamento, afinal não observamos o contrafactual, mas é possível fazer um esforço de pesquisa para "defender" essa hipótese. Por exemplo, com mais dados, talvez fosse possível testar se no passado o emprego na PA e em NJ seguiram tendências paralelas.

Para realizar inferência estatística, especificamente avaliar se os valores encontrados são estatisticamente diferentes de zero, vamos usar de regressão para obter o estimador de diferenças em diferenças. Conforme vimos acima, para isso devemos fazer uma regressão com a proporção de empregos em tempo integral como variável dependente, as variáveis independentes são uma *dummy* especificando os períodos anteriores e posteriores ao aumento do salário-mínimo, uma *dummy* especificando as unidades tratadas e de controle, no caso se a loja fica em PA ou NJ, e um termo de interação entre essas duas variáveis. Será estimada a equação:

$$Y = \beta_0 + \beta_1 Dtr + \beta_2 Dtm + \beta_3 Dtr \times Dtm + \varepsilon_{it}$$

Na equação acima a variável Y_t representa a proporção de empregos em tempo integral, a variável Dtr é a *dummy* de tratamento, 1 para NJ e 0 para PA, e a variável Dtm é a *dummy* de tempo, 1 para depois e 0 para antes. A Tabela 3 mostra o resultado da regressão.

TABELA 3
Regressão DiD

	Variável Dependente: Proporção de empregos em tempo integral
Constante, β_0	0,310*** (0,029)
Dummy tratamento, β_1	-0,013 (0,033)
Dummy antes e depois, β_2	-0,038 (0,042)
Termo de iteração, β_3	0,062 (0,046)

Nota: * p-valor < 0,1, ** p-valor < 0,05, ***p-valor < 0,01. **Fonte**: elaboração própria.

Comparando a Tabela 3 com a Tabela 1 temos que o estimador de diferença simples é igual a $\beta_1 + \beta_3 = -0{,}013 + 0{,}062 = 0{,}49 = 4{,}9\%$, número que corresponde ao que tínhamos encontrado quando calculamos a diferença entre a proporção de empregos em tempo integral na Pensilvânia e em Nova Jérsei. O estimador antes e depois é dado por $\beta_2 + \beta_3 = -0{,}038 + 0{,}062 = 0{,}024 = 2{,}4\%$, mais uma vez o número bate com o que tínhamos calculado anteriormente. Por fim, o estimador de diferenças em diferenças é dado por β_1, β_2 e β_3.

A vantagem de usar a abordagem de regressão é a facilidade de fazer inferência. Praticamente todo software que faz regressão oferece o erro padrão (que está entre parênteses na Tabela 3), a estatística *t* e o p-valor dessa estatística. Assim basta determinar o nível de significância desejada e observar uma dessas estatísticas. No exemplo apenas a

constante é significativa a 1%, as estimativas de β_1, β_2 e $\beta_3 \beta_1, \beta_2$ e β_3 não foram significativas nem a 10%. Portanto, não podemos afirmar que o efeito do aumento do salário-mínimo em Nova Jérsei é diferente de zero a um nível de significância de 10%, 5% ou 1%.

A abordagem de regressão também permite adicionar variáveis de controle de forma relativamente simples. Basta adicionar a variável na equação estimada. Por exemplo, para controlar por cadeia de *fast-food*, como fizemos, acima basta estimar a seguinte regressão.

$$Y = \beta_0 + \beta_1 Dtr + \beta_2 Dtm + \beta_3 Dtr \times Dtm + \beta_4 F + \varepsilon_{it}$$

Onde as letras representam as mesmas variáveis que na equação anterior e F é uma variável qualitativa especificando a cadeia de cada loja. Como a variável de controle é qualitativa o software vai estimar um coeficiente para cada cadeia (é como se fossem criadas uma *dummy* para cada cadeia menos a cadeia de referência). Controlando pelas cadeias o valor estimado para $\boldsymbol{\beta}_3$ permanece 6,2%.

6. DIFERENÇAS EM DIFERENÇAS EM CONTEXTOS GERAIS

A análise de regressão para estimação de diferenças em diferenças pode ser ampliada para situações mais complexas. É possível adaptar a técnica para casos com mais unidades tratadas e/ou de controle bem como para maiores períodos anteriores e posteriores ao tratamento. Por exemplo, podemos ter dados de vários estados que aumentaram o salário-mínimo e usar uma série longa de dados coletados antes e depois do aumento, ou seja, a estimação pode ser feita usando dados em painel.

A estimação antes e depois com dados em painel é derivada de uma equação que relaciona, no tempo t, a variável de resultado de interesse, y_{Ft}, contra todos os fatores relevantes para a formação daquela variável de interesse, ou seja, a soma de n elementos associados ao processo gerador dos dados, F, e uma dummy para os períodos pós-evento:

$$y_{Ft} = a_F + \sum_{i=1}^{N} \gamma_n F_{nt} + \propto d_t + \varepsilon_{Ft}$$

Onde a_F é uma constante (invariante ao tempo e associada ao estado); ε_{Ft}, , o termo de erro e d_t igual a 1 para os anos pós-evento e 0, caso contrário.

A dificuldade inerente a essa categoria de modelo é imputar todos os fatores relevantes que influenciaram o nível de resultado e que não possuem correlação com o evento. Se houver falha nesse processo, o impacto do evento pode ser estimado erroneamente. Caso contrário, o parâmetro estimado $\propto$ tem o papel de representar o efeito do evento para a variável de interesse que desejamos estudar (chamado *average treatment effect* – ATE).

Os modelos que calculam o impacto da política por método de diferenças em diferenças (DiD) são mais simples pela exigência de dados, porque, nesse caso, escolhe-se um grupo de controle, como representação dos fatores *F*, mencionados acima. Ou seja, esses grupos consistem em unidades com características semelhantes, mas cujos resultados analisados não tenham sido influenciados pelo evento. No caso mais simples, o conjunto de variáveis, *F*, mencionadas acima, deve ser igual para os dois grupos comparados (controle e de tratamento). Se isso for verdade, a diferença de resultado entre o grupo em que houve o evento e o grupo de controle pode ser suficiente para quantificar o impacto gerado pela política.

Para o grupo tratado o resultado pode ser expresso tal como na equação anterior, já o resultado no grupo de controle, y_{Ct}, , pode ser escrito como:

$$y_{Ct} = a_C + \sum_{i=1}^{N} \gamma_n F_{nt} + \varepsilon_{Ct}$$

Da diferença entre as equações (1) e (2) traduz-se o diferencial da variável resultado entre os grupos tratamento e o de controle em termos apenas da identificação do período após o evento. Sendo assim:

$$y_{Ft} - y_{Ct} = (a_F - a_C) + \propto d_t + (\varepsilon_{Ft} - \varepsilon_{Ct})$$

É importante observar que se existirem fatores que afetam apenas um dos grupos, sua inclusão será necessária; porém, seria desnecessária a inclusão de todos os fatores F relevantes incluídos na equação usada no antes e depois.

Como já apontado, a dificuldade da aplicação de modelos DID está, principalmente, na escolha adequada do chamado grupo de controle. Uma solução que visa minimizar esse tipo de problema é a comparação de resultados obtidos com grupos de controle distintos, inclusive a partir de grupos geográficos, grupos sintéticos etc. Portanto a discussão de grupos de controle é fundamental para a escolha do modelo adequado ao caso de estudo, pois grande parte do que será determinado como efeito líquido (ATE) depende do que é determinado como grupo de controle. À frente discutiremos alguns métodos adequados para dados que não são experimentos naturais (dados não experimentais) e com grupos de controle sintéticos. No primeiro caso se aplica a situações em que o grupo de tratamento não foi selecionado por um choque exógeno na implementação da política e no segundo caso se discute técnicas para se construir grupo de controle sintético, que é uma boa alternativa quando é difícil identificar o grupo de controle no período após a mudança/implementação da política.

Além dos diferenciais entre resultados nos grupos de controle e de tratamento, é possível incluir em modelos uma variável *dummy* para algumas características relevantes, o que permite estimar interceptos distintos. Isso é importante, porque se há choques específicos que não mudam ao longo do tempo e são atribuídos às características destas variáveis ou mercados, esse será capturado pelo coeficiente dessa variável.

A correta identificação do impacto de um programa ou política está baseada na hipótese de que os grupos de controle e tratamento possuem a mesma tendência. Se este não for o caso o estimador DiD não é consistente. No caso em análise pode ocorrer a situação em que uma unidade no grupo de controle se comporta de forma distinta das que estão no grupo de tratamento. Se isto ocorrer o estimador adequado poderia ser o DiD com ajustamento de tendências ou o de grupo sintético.

O controle para ajustamento de tendência é discutido por Blundell e Dias (2000, 2008). Considere que existe a possibilidade de que a hipótese de tendência comum entre grupos de controle e tratamento

não se sustenta, mas é possível supor que a seleção de tratamento é independente do efeito temporário indivíduo-específico:

$$E[u_{it}|d_i = dt'] = E[n_i|d_i = d] + q^d m_t$$

Na equação acima q^d é um escalar que permite efeitos agregados diferentes entre os dois grupos e d representa os grupos de tratamento e controle podendo tomar os valores discretos [0, 1].

Neste caso, quando se estima o modelo DID teremos:

$$E[\hat{\alpha}^{DID}] = \alpha^{ATE} + (q^1 - q^0)E[m_{t_1} - m_{t_0}]$$

Que não é capaz de estimar o parâmetro correto dado por ATE, a não ser que q^1 seja igual a q^0, que seria o caso do estimador DiD com tendência comum entre os dois grupos. O termo $(q^1 - q^0)E[m_{t_1} - m_{t_0}]$ na equação acima contabiliza pelo viés da estimativa na presença de choques agregados distintos.

Uma solução possível seria comparar as tendências dos dois grupos historicamente antes do período de intervenção. Os dados pré-tratamento podem ser úteis se existir outro intervalo de tempo em que mudança similar na tendência agregada tenha ocorrido. Este período poderia ser representado por (ta, tb). Para a estimação correta seria necessário estimar o modelo DID contendo o viés de tendência como na equação acima, estimar o impacto de diferentes tendências para o período (ta, tb) e recalcular o impacto final eliminando o viés devido ao choque agregado.

Especificamente, seria necessário estimar o viés descrito acima. Para realizar este procedimento se estima o diferencial da tendência agregada para o período pré-tratamento, (ta, tb). Esta estimativa é utilizada para se calcular o termo de viés, supondo que:

$$(q^1 - q^0)E[m_{t_a} - m_{t_b}] = (q^1 - q^0)E[m_{t_1} - m_{t_0}]$$

O impacto do tratamento pode ser isolado ao se comparar as estimativas DiD para os dois períodos, (ta, tb) e (t0, t1). Neste caso, o novo estimador entrega o efeito médio de tratamento:

$$\hat{\alpha} = \underbrace{\{[\bar{y}^F_{t_1} - \bar{y}^F_{t_0}] - [\bar{y}^C_{t_1} - \bar{y}^C_{t_0}]\}}_{DiD\ tradicional} - \underbrace{\{[\bar{y}^F_{t_b} - \bar{y}^F_{t_a}] - [\bar{y}^C_{t_b} - \bar{y}^C_{t_a}]\}}_{Controle\ diferença\ choque\ agregado} \quad (10)$$

tal que F é o grupo de tratamento, C o grupo de controle, t = 1 para o período pós-tratamento e t = 0 para o período pré-tratamento.

Uma generalização para a solução do problema do choque agregado descrito na equação acima poderia ser a análise com grupos de controle sintéticos. O método de controle sintético (SC), apresentado por Abadie e Gardeazabal (2003) e Abadie et al. (2010), tenta resolver esse problema comparando a tendência na unidade atingida pelo choque ou pela política com a tendência em uma unidade sintética composta a partir de diversas unidades observadas. Na definição em Abadie e Gardeazabal (2003) e Abadie et al. (2010) a unidade de controle sintético é uma média ponderada das unidades de controle disponíveis que melhor aproxima as características, inclusive de tendência, da variável tratada antes do tratamento.

A apresentação formal do método pode ser encontrada em Abadie et al. (2010). Considere que são observadas $j = 1, 2, \cdots, J+1$ unidades nos períodos $t = 1, 2, \cdots, T$ e que a primeira unidade tenha sido submetida a uma determinada intervenção, de forma que as demais unidades serão usadas para formar o controle sintético. Defina Y^N_{it} como os valores da variável de interesse para unidade i no período t caso a unidade não tivesse sido submetida à intervenção e Y^I_{it} caso a unidade tenha sofrido a intervenção. A unidade sintética deve ser capaz de reproduzir a unidade que será tratada não apenas na variável de interesse, mas em um conjunto de variáveis relevantes. Seja U_i um vetor $r \times 1$ de variáveis relevantes observadas para cada unidade, defina também o vetor $K = (k_1, \cdots, k_{T_0})$, onde T_0 é o período anterior à intervenção, como pesos de uma combinação linear para variável relevante antes da intervenção para as várias unidades $\bar{Y}^K_i = \sum_{s=1}^{T_0} k_s Y_{is}$, estas combinações podem ser usadas para controlar por características cujo os efeitos variam ao longo do tempo.

Para construir a unidade de controle sintético é preciso criar um vetor $(J \times 1)$ de pesos $W = (w_2, \cdots, w_{J+1})$ com $w_j \geq 0 \, \forall j$ e $\sum_{j=2}^{J+1} w_j = 1$ onde cada elemento do vetor representa o peso de uma unidade de controle observada. Abadie e Gardeazabal (2003) e Abadie et al. (2010) propõem escolher o vetor de pesos W^* tal que a unidade sintética de controle obtida melhor aproxime a unidade que passou pela intervenção com respeito a U_i e $M \leq T_0$ combinações lineares para variável de interesse antes da intervenção. Formalmente W^* é tal que $\sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_1} \approx \bar{Y}_1^{K_1} \cdots \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_M} \approx \bar{Y}_1^{K_M}$ e $\sum_{j=2}^{J+1} w_j^* U_j \approx U_1$, então:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* \, Y_{jt}$$

É o estimador de α_{1t} nos períodos posteriores à intervenção.

Para implementar o estimador de controle sintético numericamente é preciso definir uma distância entre a unidade de controle sintético e a unidade tratada, para isso basta agregar as características da unidade tratada na matriz $X_1 = \left(U_1', \quad \bar{Y}_1^{K_1}, \cdots, \bar{Y}_1^{K_M}\right)_{k \times 1}$ e os valores das mesmas variáveis para as unidades de controle em $X_1 = \left(U_j', \quad \bar{Y}_j^{K_1}, \cdots, \bar{Y}_j^{K_M}\right)_{k \times J}$. O vetor de pesos é calculado de forma a minimizar:

$$\|X_1 - X_0 W\|V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

onde $V_{k \times k}$ é simétrica positiva semidefinida. Abadie e Gardeazabal (2003) e Abadie et al. (2010) sugerem escolher $V_{k \times k}$ como:

$$argmin_{V \in \mathcal{V}} \left(Z_1 - Z_0 W^*(V)\right)' \left(Z_1 - Z_0 W^*(V)\right)$$

onde $\mathcal{V}$ é o conjunto de todas as matrizes diagonais positivas definidas.

7. DIFERENÇAS EM DIFERENÇAS SINTÉTICO

Geralmente métodos de diferenças em diferenças (DiD) são aplicados em casos em que se tem um número substancial de unidades que são expostas a mudança de política e pesquisadores desejam fazer hipótese de tendência paralela, que implica que podemos controlar pelos efeitos de seleção por meio de efeitos de tempo e de unidade. Por outro lado, controle sintético (SC) geralmente é usado em situações com apenas uma única ou poucas unidades expostas, o método procura compensar a falta de tendência paralela pela determinação de pesos às unidades ancorando na tendência pré-tratamento.

Synthetic Difference in Differences (SDID), proposto em Arkhangelsky, Athey, Hirshberg, Imbens e Wager (2021) combina as melhores características de DID e SC. Similarmente ao SC, SDID utiliza uma hipótese de tendência paralela ao calcular pesos para as unidades antes do tratamento. Como no DID, o método é invariante a deslocamentos aditivos nas unidades e permite inferência em painéis grandes. Deste modo, SDID tem potencial para suplantar as aplicações de DID e SC.

7.1 O ESTIMADOR SDID

O estimador SDID é um modelo de regressão ponderada pelos pesos de grupo ω_i , ou simplesmente unidades tratadas e não-tratadas, e pelos pesos de tempo, λ_t . Esta regressão é um estimador com dois efeitos fixos (TWFE) do modelo DiD.

$$\underset{\tau,\mu,\alpha,\beta,\gamma}{\arg min}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\mu-\alpha_i-\beta_t-X_{it}\gamma-W_{it}\tau)^2\hat{\omega}_j\hat{\lambda}_t\right\}$$

O método é aplicado em painel balanceado com *N* unidades e *T* períodos. A unidade resultado é representada por Y_{it} para a unidade *i* no período *t* e exposto ao tratamento binário $W_{it} \in \{0,1\}. X_{it}$. X_{it} é uma matriz de variáveis explicativas que pode ser incluído no modelo SDID. As hipóteses de identificação do efeito tratamento são:

1. As unidades de controle, N_{co}, nunca são expostas ao tratamento.
2. As unidades que recebem o tratamento, $N_{tr} = N - N_{co}$, são expostas ao tratamento comente após o período T_{pre} . Todas as unidades são expostas em bloco ao tratamento.

O procedimento começa procurando pesos que alinhem as tendências da variável resultado do grupo de controle e do grupo de tratamento. Exemplo:

$$\sum_{i=1}^{N_{co}}\hat{\omega}_i Y_{it} \approx \frac{1}{N_{tr}}\sum_{i=N_{co}+1}^{N} Y_{it} \quad \forall\, t = 1,\cdots,T_{pre}$$

A tendência paralela também é encontrada para períodos, ou seja, os pesos de tempo, λ_t , são calculados para equilibrar os períodos pré-tratamento com o período pós-tratamento.

7.1.1 Estimação dos pesos: $\hat{\omega}_i$ e $\hat{\lambda}_t$

O ponderador dos grupos é estimado como a solução de um problema de otimização de uma regressão *ridge* com restrições. Aqui supomos uma forma funcional para a regressão *ridge*, como em Hastie, Tibshirani, and Friedman (2017, p. 63). Os pesos são uma solução para:

$$\hat{\omega}_i = arg\ min\left\{\sum_{t=1}^{T_{pre}}\left(\omega_0 + \sum_{i=1}^{N_{co}}\omega_i Y_{it} - \frac{1}{N_{tr}}\sum_{i=N_{co}+1}^{N} Y_{it}\right)^2 + \phi\sum_{i=1}^{N_{co}}\omega_i^2\right\}$$

Sujeito a duas restrições:

$$\sum_{i=1}^{N_{co}}\omega_i = 1 \ \ e \ \ \omega_i \in \Re_+^{N_{co}}$$

Observe que os pesos são definidos para o período pré-tratamento. O parâmetro *ridge*, ϕ, controla o montante de redução dos parâmetros. Em um modelo de regressão linear com muitas variáveis correlacionadas, os coeficientes se tornam difíceis de estimar e exibem grande variância. Por exemplo, um coeficiente de uma variável com grande valor pode ser cancelado por um coeficiente negativo de magnitude similar. Pela imposição do termo de penalidade, a regressão *ridge* reduz este tipo de problema.

No estimador SDID, Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021), estabelecem ϕ como:

$$\phi = T_{pre}\left(T_{pos}N_{tr}\right)^{1/2}\zeta^2$$

onde ζ é dado por:

$$\zeta^2 = \frac{1}{N_{co}T_{pre} - 1}\sum_{i=1}^{N_{co}}\sum_{t=1}^{T_{pre}-1}(\Delta_{it} - \overline{\Delta})^2$$

Vale ainda que

$$\Delta_{it} = Y_{i,t+1} - Y_t, \text{ e } \overline{\Delta} = \frac{1}{N_{co}(T_{pre}-1)} \sum_{n=1}^{N_{co}} \sum_{t=1}^{T_{pre}-1} \Delta_{it}.$$

Na abordagem SDID, o parâmetro de regularização ζ é ajustado para capturar a mudança típica em um período na unidade não exposta ao tratamento no período anterior a intervenção.

Por sua vez, os pesos do grupo de tratamento são definidos como a média do número de unidades tratadas:

$$\omega_i = \frac{1}{N_{tr}} \; se \; i \; \in N_{tr}$$

Por fim, o peso de tempo é estimado pela solução do seguinte problema de otimização linear:

$$\hat{\lambda}_t = arg\ min \left\{ \sum_{i=1}^{N_{co}} \left(\lambda_0 + \sum_{t=1}^{T_{pre}} \lambda_t Y_{it} - \frac{1}{T_{pos}} \sum_{t=T_{pre}+1}^{T} Y_{it} \right)^2 \right\}$$

Sujeito a:

$$\sum_{t=1}^{T_{pre}} \lambda_t = 1$$

$$\lambda_t \in \Re_+^T$$

$$\lambda_t = \frac{1}{T_{pos}} \; se \; t \in T_{pos}$$

7.2 INFERÊNCIA

Sob condições apropriadas, o estimador é assintoticamente normal e centrado em zero. Tendo um estimador consistente para a variância assintótica, V_τ , podemos construir intervalos de confiança $\tau \in \hat{\tau}^{SDID} \pm z\alpha_{/2}\sqrt{\hat{V}_\tau}$ que permite realizar inferência adequada.

A literatura sugere duas alternativas para estimativa da variância: *Bootstrap* e *Jackkniffe*. A estimação por *Bootstrap* talvez seja natural, como explicitado por Bertrand, Duflo e Mullainathan (2004). O *bootstrap* é simples de implementar e entrega performance robusta em painéis grandes. O problema desta alternativa é que pode ser computacionalmente intensiva, pois é preciso reestimar todo o algoritmo para cada replicação *bootstrap*, especialmente para grandes bases o custo computacional é proibitivo (custo aqui é de tempo, energia e preço de uma máquina veloz). Abaixo segue o algoritmo *bootstrap* para SDID:

Uma alternativa a estimação por *bootstrap* é utilizar a estimação *jackknife* na regressão ponderada SDID. Com o método *jackknife* apenas é necessário rodar todo o procedimento uma única vez, pois é apenas necessário rodar a regressão WLS com os pesos de tempo e de grupo do algoritmo pleno. Uma condição para utilizar a variância *jackknife* é a de que o peso de tempo, $\hat{\lambda}$, de previsão suficiente para as unidades expostas ao tratamento. Este resultado depende da estrutura específica do estimador SDID e não se aplica a métodos similares, como o SC, por exemplo.

Em muitas aplicações o controle sintético possuir apenas uma unidade tratada. Para este fim, os autores propõem um terceiro alternativa: cálculos de placebo. A principal ideia do método de placebo é considerar o comportamento da estimação de controle sintético quando se troca a unidade que foi exposta ao tratamento por unidades que não foram expostas.

7.3 ESTIMADORES DID E SC NO ARCABOUÇO SDID

Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) apresentam DiD e SC dentro do arcabouço SDID. Quando os pesos de unidade e tempo não existem, ou seja, quando $\omega_i = 1$ e $\lambda_t = 1$, o estimador SDID colapsa para uma abordagem DiD-TWFE:

$$\underset{\tau,\mu,\alpha,\beta,\gamma}{arg\ min}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{jt}-\mu-\alpha_i-\beta_t-X_{it}\gamma-W_{it}\tau\right)^2\right\}$$

Quando se subtrai o efeito fixo e se utiliza pesos de unidades temos um estimador análogo ao SC:

$$\underset{\tau,\mu,\alpha,\beta,\gamma}{arg\ min}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{jt}-\mu-\beta_t-X_{it}\gamma-W_{it}\tau\right)^2\widehat{\omega}_i\right\}$$

7.4 PROGRAMA DE CONTROLE DE TABACO NA CALIFÓRNIA

Um exemplo clássico utilizado em diversos estudos é a avaliação do programa de controle de tabaco feito pelo estado da Califórnia em 1989. A Proposição 99 aumentou os impostos pagos sobre um pacote de cigarros em 25 centavos de dólar (veja descrição mais detalhada em Abadie et al, 2010). O impacto desta reforma foi calculado utilizando a evolução do consumo de pacotes per capita na Califórnia em comparação com 38 estados que não implementaram a mesma política.

Os dados utilizados na análise cobrem 39 estados entre 1970 e 2000. A adoção ocorreu na Califórnia em 1989, isto implica que:

$$T_{pre} = 19\ e\ T_{pos} = 12$$

Existem 38 estados para controle e uma única unidade tratada:

$$N_{co} = 38 \; e \; N_{tr} = 1$$

A estimação por DiD sintética começa com a estimação dos pesos dos grupos. Geralmente é preciso determinar uma base de dados ampla onde pode se encontrar as unidades adequadas para construir o grupo de controle sintético.

No método de controle sintético (Abadie et al, 2010) os pesos são determinados pela comparação de características dos estados. No artigo original sobre controle sintético as médias pré-tratamento são logaritmo do rendimento per capita, percentual de pessoas entre 15 e 24 anos, preço médio cobrado no varejo, consumo de cerveja per capita, vendas de cigarro per capita para 1988, 1980 e 1975. Veja a Tabela 4.

TABELA 4

Médias do Previsor de Vendas de Cigarros

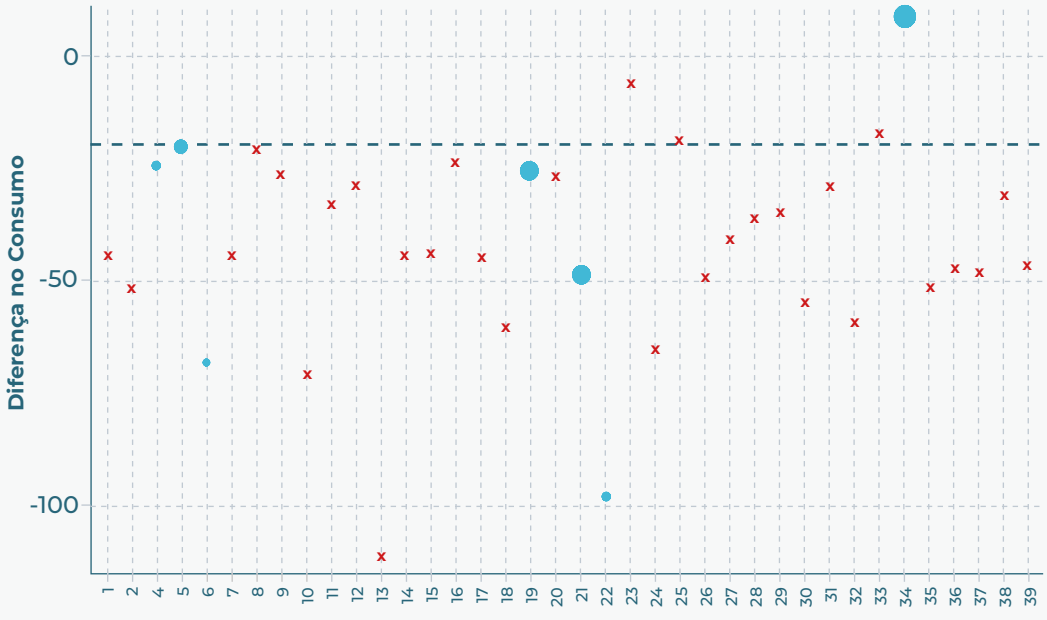
Variável	Real da Califórnia	Sintético da Califórnia	38 Estados de Controle
Ln (Rendimento per capita)	10,08	9,86	9,86
Percentual de pessoas de 15-24 anos	17,40	17,40	17,29
Preço de varejo	89,42	89,41	87,27
Consumo de cerveja per capita	24,28	24,20	23,75
Vendas de cigarros per capita 1988	90,10	91,62	114,20
Vendas de cigarros per capita 1980	120,20	120,43	136,58
Vendas de cigarros per capita 1975	127,10	126,99	132,81

Nota: Todas as variáveis, exceto as vendas de cigarros defasadas, são médias do período de 1980-1988 (o consumo de cerveja é a média de 1984-1988) Fonte: Abadie et al (2010).

As Figuras 2, 3 e 4 mostram os pesos na estimação usando controle sintético, diferenças em diferenças e diferenças em diferenças sintético. É possível reparar que no DiD todas as unidades possuem o mesmo peso e que o SDID utiliza mais unidades para criar o controle sintético do que o SC.

FIGURA 2

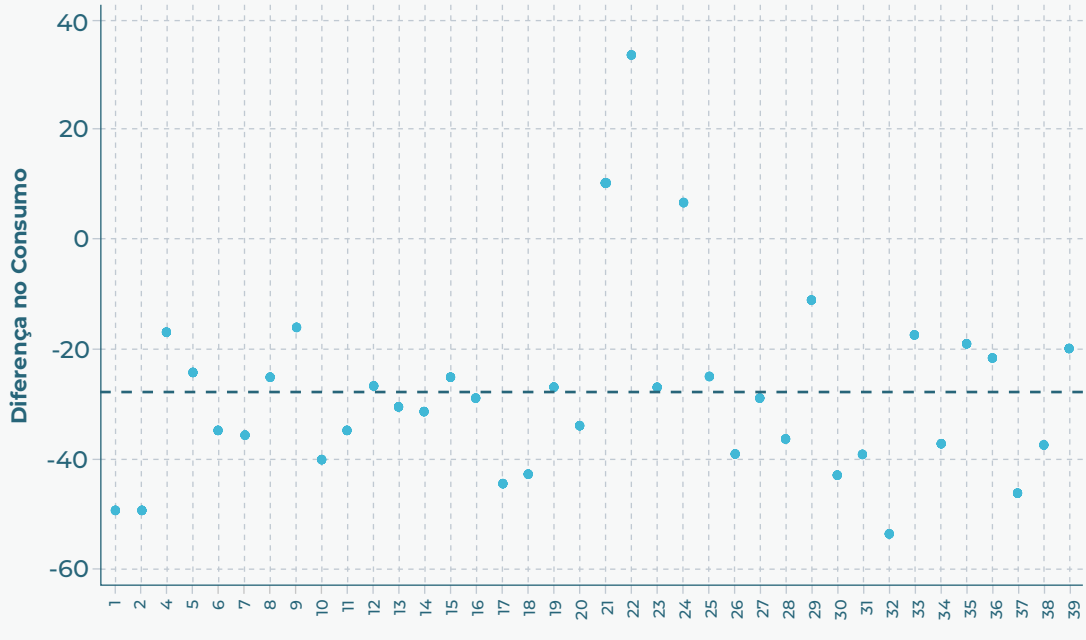
Pesos para os Métodos SC



Fonte: Abadie et al (2010).

FIGURA 3

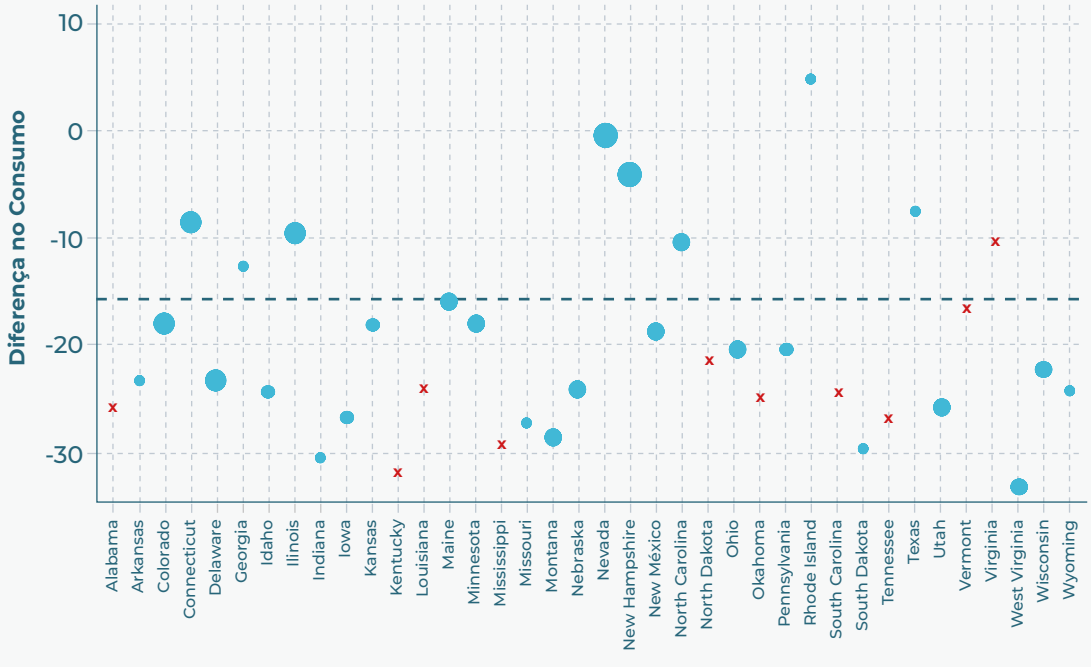
Pesos para os Métodos DID



Fonte: Abadie et al (2010).

FIGURA 4

Pesos para os Métodos SDID

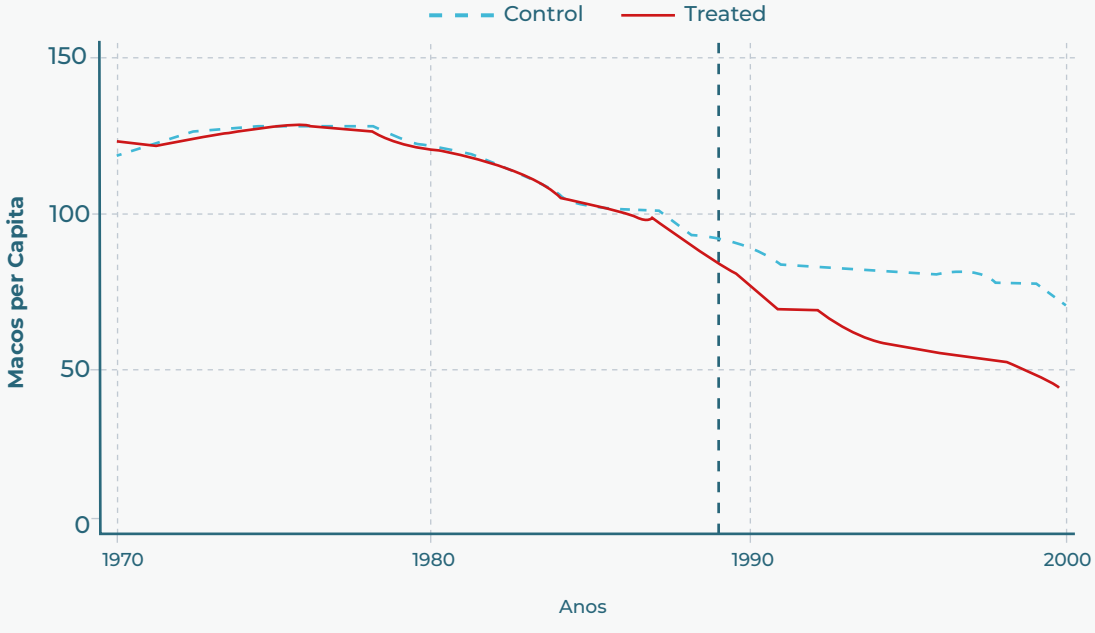


Fonte: Abadie et al (2010).

As Figuras 5, 6 e 7 mostram o consumo de maços de cigarros na California e na Califórnia sintética utilizando os métodos SC, DiD e SDID. Repare que o método SC faz a melhor aproximação no período pré-tratamento, porém o controle sintético criado no SDID apresenta comportamento mais próximo da hipótese de tendência paralela.

FIGURA 5

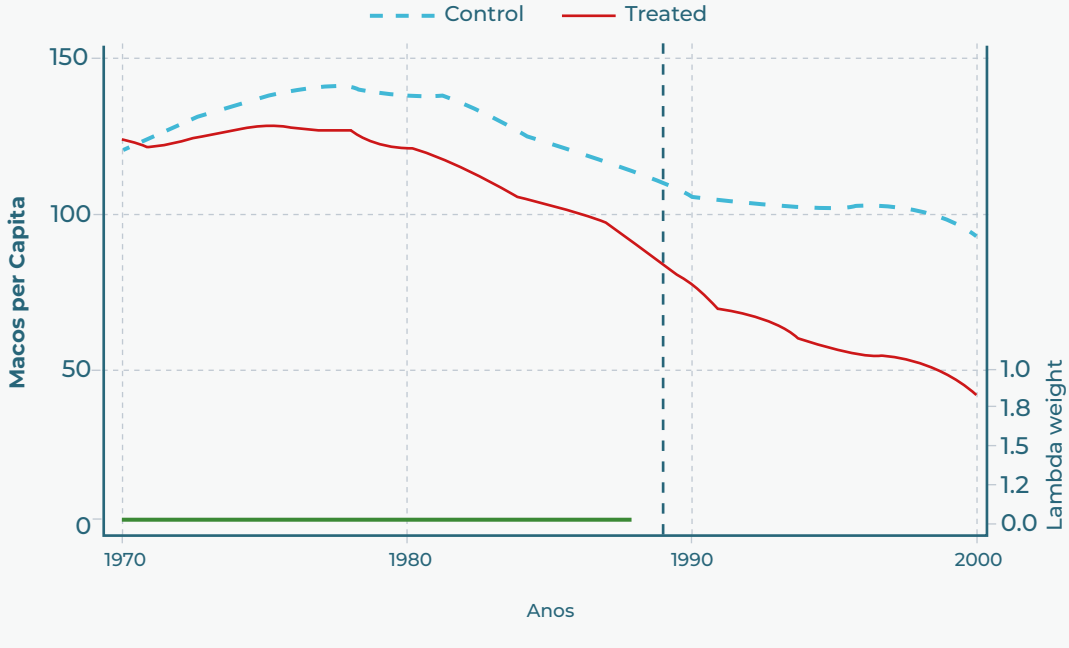
Tendências para os Métodos SC



Fonte: Abadie et al (2010).

FIGURA 6

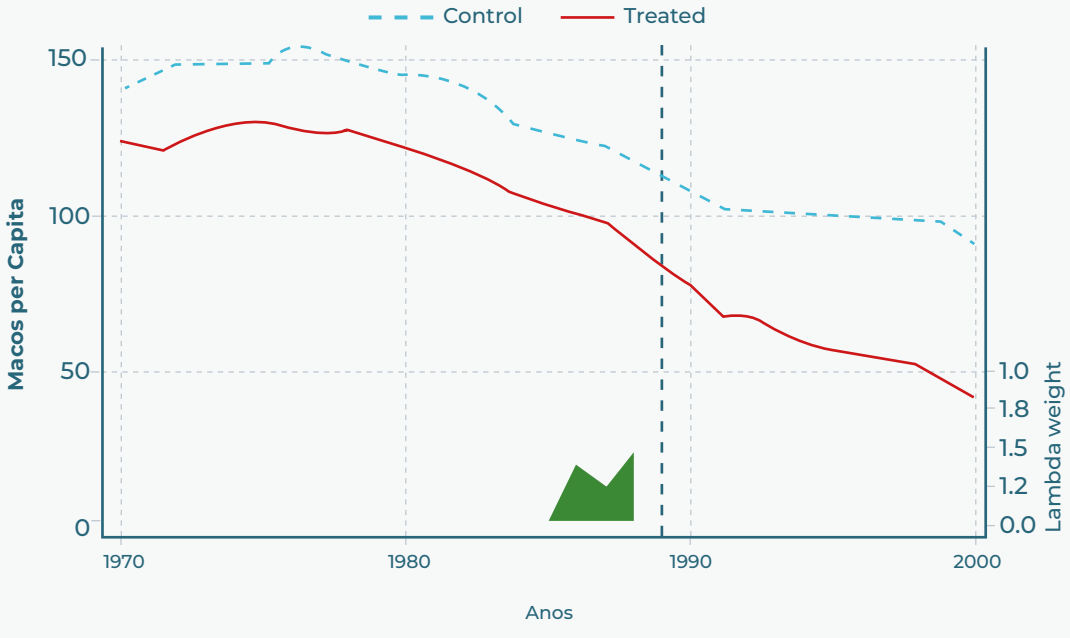
Tendências para os Métodos DID



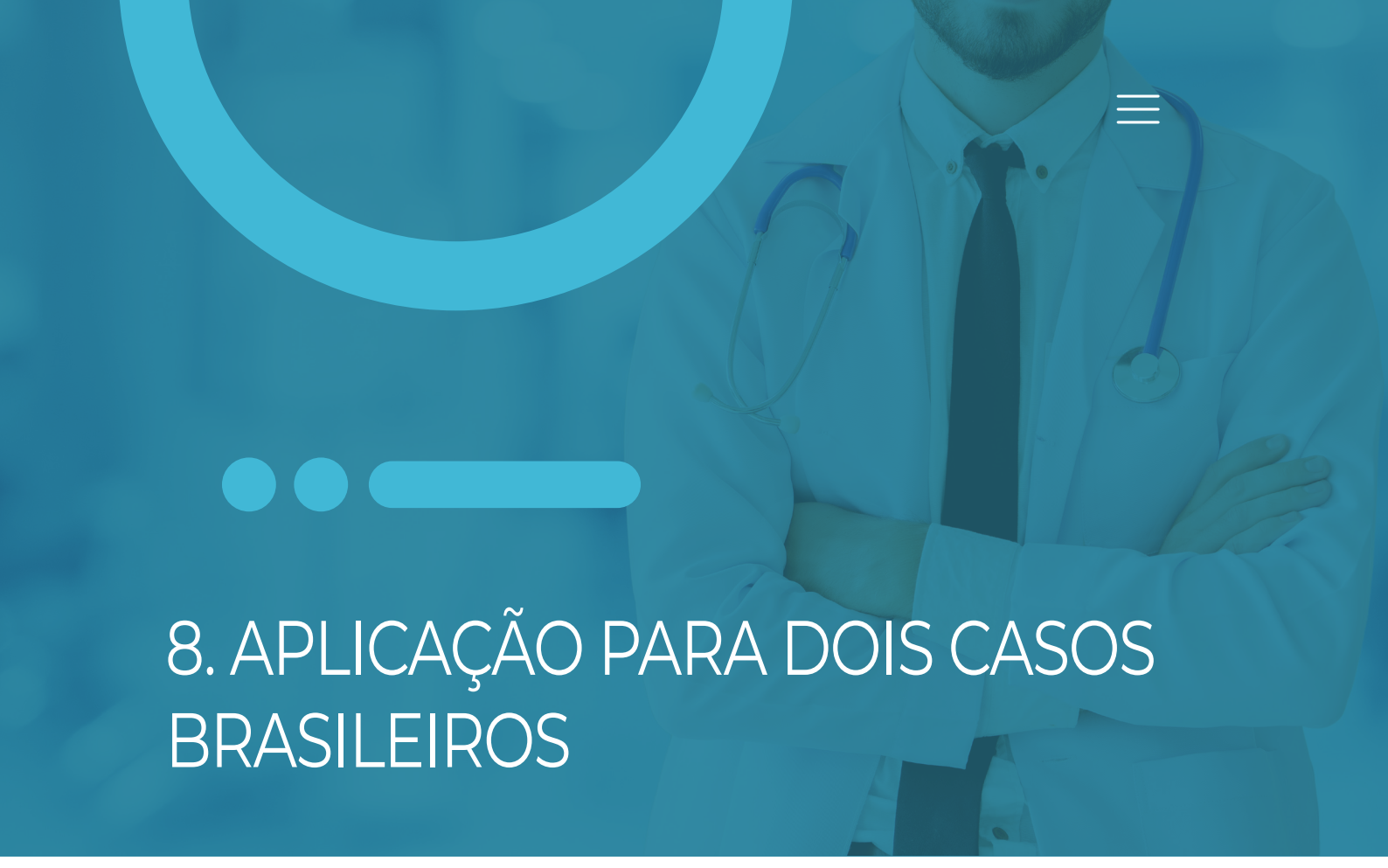
Fonte: Abadie et al (2010).

FIGURA 7

Tendências para os Métodos SDID



Fonte: Abadie et al (2010).



8. APLICAÇÃO PARA DOIS CASOS BRASILEIROS

Foi realizada a avaliação da política para dois casos do PROADI-SUS: o primeiro é o LEAN e o segundo é o “Saúde em nossas mãos”. Abaixo seguem a descrição dos programas.

O projeto Lean nas Emergências é uma iniciativa do Ministério da Saúde executada no âmbito do PROADI-SUS, originalmente liderada pelo Hospital Sírio-Libanês a partir de 2017, em cooperação com outros hospitais de referência (como a Beneficência Portuguesa de São Paulo e o Moinhos de Vento) para difundir e aplicar a filosofia Lean especificamente em serviços de urgência e emergência. O objetivo central do Lean nas Emergências é reduzir a superlotação nas urgências e emergências de hospitais públicos e filantrópicos e melhorar o tempo de atendimento e a eficiência operacional, sem comprometer a qualidade e segurança do cuidado. Isso se faz por meio da identificação e eliminação de desperdícios nos fluxos (tempo parado, retrabalhos, gargalos), racionalização de recursos, otimização de uso de espaço e insumos, e fortalecimento da gestão cotidiana.

O “Saúde em Nossas Mãos” é uma colaborativa do PROADI-SUS, articulada com o Ministério da Saúde por meio do Programa Nacional de Segurança do Paciente (PNSP), com o objetivo de reduzir infecções

relacionadas à assistência à saúde (IRAS) em unidades de Terapia Intensiva (UTIs) públicas brasileiras. A iniciativa se apoia no "*improvement model*" do *Institute for Healthcare Improvement* (IHI), baseado no método *Breakthrough Series* (BTS) – um modelo de melhoria da qualidade em larga escala que combina aprendizado colaborativo, testes rápidos de mudança e disseminação de boas práticas. A meta explícita do projeto era reduzir as densidades de incidência das três IRAS principais associadas a dispositivos invasivos: Pneumonia associada à ventilação mecânica (PAV), Infecção primária da corrente sanguínea associada a cateter venoso central (CLABSI/IPCSL), e Infecção do trato urinário associada a cateter vesical (CAUTI/ITU-AC).

8.1 LEAN NAS EMERGÊNCIAS

O projeto visa entregar a redução da superlotação de hospitais públicos e filantrópicos, agilizando o atendimento do paciente, consequentemente reduzindo o tempo de passagem pelo pronto socorro (tempo entre chegada e saída física do paciente deste setor), além de otimizar o uso dos leitos e aumentar a eficiência do centro cirúrgico do hospital. Também tem por objetivo ampliar para outros hospitais, o conhecimento sobre a metodologia Lean Healthcare aplicada à serviços de urgência, bloco cirúrgico e unidades de internação, e ser base de conhecimento e boas práticas através da Comunidade Lean nas Emergências(www.leannasemergencia.com.br), além de disseminar a metodologia de forma remota, através desta plataforma digital, para hospitais do interior e de pequeno ou médio porte. Como resultado, busca-se a melhoria da eficiência operacional dos hospitais nos processos das três grandes áreas assistenciais: (i) pronto-socorro, (ii) bloco cirúrgico e (iii) unidade de internação, melhorando a qualidade e segurança do processo de cuidado e a sustentabilidade dos hospitais. O Lean possui duas fases. A Lean Fase I teve a participação de 37 hospitais no primeiro semestre de 2023. Este projeto conta com consultores dos hospitais Beneficência Portuguesa, Hospital Moinhos de Vento e Hospital Sírio-Libanês. Em julho de 2023 iniciou o ciclo 8 do Lean Fase I com 39 hospitais. O Lean Fase II atuou em 8 hospitais, trata da transformação nos hospitais e é exclusivamente liderado pelo Hospital Sírio-Libanês. O ci-

clo 4 se inicia também em julho de 2023 e conta com 12 hospitais – veja as tabelas correspondentes.

TABELA 5

Hospitais Lean Fase I

Nome do Hospital	UF	CNES
Honpar Hospital Norte Paranaense	Paraná	2576341
Hospital Municipal Abelardo Gadelha da Rocha	Ceará	2562316
Hospital de Urgência de São Bernardo do Campo	São Paulo	2069776
Hospital e Pronto Socorro Municipal de Várzea Grande	Mato Grosso	2391635
Hospital Estadual de Formosa	Goiás	2534967
Hospital Estadual Mario Covas de Santo André	São Paulo	2080273
Hospital Márcio Cunha	Minas Gerais	2205440
Hospital Maternidade São Vicente de Paulo	Ceará	2564211
Hospital Metropolitano de Urgência e Emergência	Pará	3987884
Hospital Regional de Barbacena Dr. José Américo	Minas Gerais	3698548
Hospital Regional de Estância Jesse Fontes	Sergipe	6901743
Hospital Regional do Sudoeste Walter Alberto Pecoits F B	Paraná	6424341
Hospital Reginal Justino Luz	Piauí	4009622
Hospital Regional Tiberio Nunes	Piauí	2365146
Secretaria Estadual de Saúde do Amapá Hospital Estadual	Amapá	2021064
Sesdec RJ Hospital Estadual Getúlio Vargas	Rio de Janeiro	2270234
Hoesp	Paraná	4056752
Hospital Municipal Dr. Fernando Mauro Pires da Rocha	São Paulo	2786680
Hospital Dr. Leopoldo Bevilacqua	São Paulo	2077434
Hospital Ferreira Machado	Rio de Janeiro	2287579
Hospital Geral de Clínicas de Rio Branco	Acre	2001578
Hospital Municipal de Paulínia	São Paulo	2081059
Hospital Municipal Senhora Santana	Minas Gerais	2119420
Hospital Municipal Vereador José Storopolli	São Paulo	3212130
Hospital Pompeia	Rio Grande do Sul	2223546
Hospital Regional Dr. Jorge de Matos Cohen	Amazonas	3210243
Hospital Santa Casa de Misericórdia	Paraná	0014109
Hospital Santo Antonio Maria Zaccaria	Pará	2678403
Hospital Universitário Alzira Velano	Minas Gerais	2171988
Instituto de Cardiologia Hospital Viamão	Rio Grande do Sul	5223962
Santa Casa de Araçatuba Hospital Sagrado Coração de Jesus	São Paulo	2078775
Santa Casa de Caridade de Uruguaiana	Rio Grande do Sul	2248190
Santa Casa de Misericórdia de São José do Rio Preto	São Paulo	2798298
Sopai Hospital Infantil	Ceará	2526638

Fonte: elaboração própria a partir de dados disponibilizados pelo MS.

TABELA 6
Lean Fase II

Nome do Hospital	UF	CNES
Hospital e Pronto Socorro 28 de Agosto	Amazonas	2013649
Hospital de Clínicas de Uberaba UFTM	Minas Gerais	2206595
Hospital Geral de Itapecerica da Serra	São Paulo	2792176
Hospital Angelina Caron	Paraná	0013633
Hospital Eladio Lasserre	Bahia	0003980
Hospital Geral de Cotia	São Paulo	2792141
Hospital Regional Hans Dieter Schmidt	Santa Catarina	2436450
Santa Casa de Campo Grande	Mato Grosso do Sul	0009717
Hospital Regional José Simone Netto	Mato Grosso do Sul	2651610
Hospital de Cuiabá Dr. Leony Carvalho	Mato Grosso	9209352
Complexo Hospitalar de Mangabeira	Paraíba	2399628
Hospital Santa Isabel	Santa Catarina	2558246
Hospital Platão de Araújo	Amazonas	5169976
Hospital de Base Luis Eduardo Magalhães	Bahia	2385171
Hospital de Clínicas de Porto Alegre	Rio Grande do Sul	2237601
Hospital e Maternidade Silvio Avidos	Espírito Santo	2446030
Hospital Geral de Roraima	Roraima	2319659
Hospital Padre Albino	São Paulo	2889327
Imip	Pernambuco	0000434
Maternidade Escola Assis Chateubriand	Ceará	2481286

Fonte: elaboração própria a partir de dados disponibilizados pelo MS.

De acordo com o relatório de execução do projeto, observou-se redução de 38% no atraso na primeira cirurgia do dia, redução de 25% na taxa de cancelamento cirúrgico, redução de 22% no tempo de passagem do paciente pelos hospitais, e queda de 28% no escore chamado NEDOCS (*National Emergency Department Overcrowding Score*), composto por 07 variáveis e utilizado para avaliar nível de superlotação de serviços de urgência e emergência.

Indicadores de resultados[3]

A literatura de economia da saúde utiliza frequentemente algumas variáveis como resultado. Geralmente a literatura reporta índices de saúde de pacientes baseados em várias medidas intermediárias de saúde. "*These scores are designed to aid medical decision-making and provide benchmarking tools. For any given scoring method, health measures are categorized into ranges, with each range being assigned a number; the score is a weighted average of these score components*" (Svirbely e Sriram, 2001).

Athey e Stern (2002) utilizam um índice de saúde para pacientes cardíacos. Como eles não conseguiram identificar um único sistema "melhor" de pontuação para o grupo específico de pacientes (todos com diagnósticos cardíacos, com dados vitais medidos na chegada da ambulância), construíram vários índices, baseados nos principais escores desenvolvidos para cuidados críticos, utilizando quatro medidas brutas de status de saúde intermediária. Começaram criando duas variáveis indicadoras com base no fato de o paciente estar na região "de alto risco" em termos de uma única medida de saúde: PRESSÃO ARTERIAL ALTA (igual a 1 se a pressão arterial sistólica for inferior a 90) e PULSO ALTORISCO (igual a 1 se a taxa de pulso for inferior a 40). A correlação entre essas medidas é 0,68. Em seguida, calcularam um índice de status de saúde intermediário, o HINDEX, da seguinte forma. Primeiro, foi criado um conjunto de categorias para cada uma das quatro medidas brutas de saúde com base em (a) os pontos críticos de corte para PRESSÃO ARTERIAL, RESPIRAÇÃO e GLASGOW sugeridos por um sistema de pontuação líder (chamado de Sistema de Pontuação de Trauma Revisado - RTS) e (b) PULSOALTO RISCO. Em seguida, é realizada uma regressão probit da MORTALIDADE EM 48 HORAS sobre o conjunto completo dessas variáveis categóricas. O HINDEX é calculado como o valor previsto da MORTALIDADE EM 48 HORAS a partir dessa regressão (sua média é 0,035, igual à probabilidade de mortalidade da amostra). Assim, o HINDEX pode ser interpretado como a probabilidade de mortalidade em 48 horas de um paciente, condicionado a (a) seu estado de saúde

3 A revisão de literatura desta seção também serve para a implementação do programa seguinte, SNM.

no momento da chegada da ambulância e (b) o paciente recebendo um nível "médio" de cuidado subsequente à chegada da ambulância.

Outro exemplo, o trabalho de McCullough, Parente e Town (2016) usa três medidas principais de resultado:

- Mortalidade após 60 dias;
- Readmissão após 30 dias (condicional a sobrevivência), e;
- Tempo de permanência.

Os autores usam estas medidas de resultado para infarto agudo do miocárdio (AMI), insuficiência cardíaca congestiva (CHF), aterosclerose coronária (CA) e pneumonia (PN).

Indicadores para SUS

A partir da SIA/Datasus é possível construir indicadores de resultados para atividade ambulatorial. O tempo de permanência é possível ser calculado utilizando a data do início e do fim do procedimento utilizando o registro de produção ambulatorial. Produção ambulatorial também pode ser a indicação de óbito, a remoção e a alta hospitalar. Entretanto, na base de dados não existem dados para estes indicadores.

As variáveis indicadoras podem ser construídas para as CIDs associadas a AMI, CHF, CA e PN, permitindo identificar, por exemplo, a presença/ausência de cada condição em nível de atendimento; contudo, dada a possibilidade de informações inconsistentes no SIA[4], convém adotar uma estratégia complementar e mais robusta: contabilizar, mensalmente e de forma independente da CID. Essa medida alternativa reduz a sensibilidade a erros de classificação ou subnotificação de diagnósticos, viabiliza análises de tendência e volume assistencial ao longo do tempo.

Design

Lean Fase I: A fase de implementação teve sua conclusão em junho de 2023 para os hospitais do ciclo 07, e dezembro de 2023 para os hospitais no ciclo 08.

4 O problema com a SIA é que existem muitas lacunas de preenchimento no campo de CID.

Lean Fase II: No primeiro semestre de 2023, o Ciclo 03 da “Transformação Lean nos Hospitais “contou com a participação de 08 hospitais e no Ciclo 04 foram contemplados 12 hospitais que haviam participado do Lean nas Emergências em ciclos anteriores.

8.1.1 Análise Quantitativa

Análise quantitativa é realizada utilizando indicadores de produção ambulatorial do SIA/DATASUS. No procedimento de organização da base de dados, foram excluídos os procedimentos domiciliares e psicossocial. Para tratabilidade, a subsequente análise foi realizada apenas para o Estado de São Paulo por ter grande número de hospitais de referência.

Para o Estado de São Paulo são 10 milhões de observações, aproximadamente. Seguindo a literatura citada anteriormente são analisados os casos das seguintes CIDs:

- I21 Infarto agudo do miocárdio;
- I50 Insuf. cardíaca
- I70 Aterosclerose
- J12 Pneumonia viral NCOP
- J13 Pneumonia dev Streptococcus pneumonia
- J14 Pneumonia dev Haemophilus infuenza
- J15 Pneumonia bacter. NCOP
- J16 Pneumonia dev out microorg infecc espec NCOP
- J17 Pneumonia em doença COP

A partir da seleção dessas CIDs foram utilizadas um conjunto de três variáveis geradoras do serviço de emergências. São elas: óbito, transferência e tempo de permanência. Como a unidade de tratamento é o hospital, para cada uma delas foi construída uma variável prevista como função da idade dos pacientes das emergências. Para realizar o controle individual foram estimados modelos probit utilizando como variáveis de previsão a idade do paciente e gênero. Esse procedimento foi realizado para toda a base.

DID

Segue abaixo a estimativa utilizando o modelo clássico de diferença-em-diferenças:

TABELA 7
Estimativa utilizando modelo DID

Variável	Coeficiente	Erro padrão	Estatística t	Valor-p	Limite inferior 95%	Limite superior 95%
Óbito	-2.062332	1.137812	-1.81	0.07	-4.292564	0.1679005
Permanência	-2.437316	1.383099	-1.76	0.078	-5.148338	0.2737052
Transferência	-2.08994	1.366406	-1.53	0.126	-4.768241	0.5883609

Fonte: elaboração própria.

O modelo de diferença-em-diferenças simples (DID) apresenta os seguintes pontos principais. O coeficiente de “tratamento” (interação representando o efeito do tratamento) é negativo para óbito (−2,06), permanência (−2,44) e transferência (−2,09), o que está na mesma direção do que mostraremos no modelo SDID, mas não é estatisticamente significativo ao nível convencional de 5% (p entre ~0,07 e 0,13), indicando menor precisão.

SDID

Foi estimado é o efeito médio do tratamento sobre os tratados (ATT) da intervenção Lean 2, comparando os hospitais tratados com uma contrafactual sintética construída, para três resultados (desfechos): Óbito, Permanência e Transferência. Como as variáveis dependentes estão em log natural, os coeficientes são aproximadamente diferenças logarítmicas, e a transformação exponencial deles permite interpretar os efeitos em termos de variação percentual.

A inferência utilizada é baseada em placebo, pois são apenas duas unidades que recebem o tratamento. Seguindo a metodologia SDID,

para avaliar a robustez e a significância do efeito estimado, foi conduzida uma análise de placebo com 200 simulações. Isso foi feito atribuindo falsos tratamentos — seja realocando o "tratamento" para unidades não tratadas ou para tempos em que ainda não havia intervenção – e reestimando o ATT em cada simulação para construir uma distribuição empírica de efeitos sob a hipótese nula de nenhum impacto real.

Resultados do Efeito Líquido do Tratamento (Lean 2) – Modelo Diferença-em-Diferenças Sintética (SDID):

TABELA 8

Resultados do Efeito Líquido do Tratamento (Lean II) – SDID

Variáveis Resultado	Efeito médio do tratamento sobre os tratados (ATT)	Erro padrão	Estatística t	Valor-p	Limite inferior do intervalo de confiança (95%)	Limite superior do intervalo de confiança (95%)
Óbito	-2.09755	0.74917	-2.8	0.005	-3.5659	-0.62921
Permanência	-2.46028	0.90601	-2.72	0.007	-4.23604	-0.68452
Transferência	-2.08434	0.77796	-2.68	0.007	-3.60912	-0.55957

Fonte: elaboração própria.

Variável óbito. ATT = -2,09755. Isso significa que, após a intervenção, o número esperado de óbitos nos hospitais tratados ficou abaixo do contrafactual sintético. Em termos percentuais: exp(-2,09755) - 1 ≈ -0,877, ou seja, uma redução de cerca de 87,7% nos óbitos.

Intervalo de confiança (95%): de -3,5659 a -0,62921. Exponenciando: exp(-3,5659) ≈ 0,0283 (redução de 97,2%) até exp(-0,62921) ≈ 0,5337 (redução de 46,6%). Ou seja, com 95% de confiança, a intervenção está associada a uma queda nos óbitos entre aproximadamente 46% e 97%. Significância: t = -2,80, valor-p = 0,005. Estatisticamente significativo ao nível de 1%; rejeita-se a hipótese nula de efeito zero.

Variável permanência. ATT = -2,46028. Interpretação percentual: exp(-2,46028) - 1 ≈ -0,914, ou seja, redução de cerca de 91,4% no tempo de permanência.

Intervalo de confiança (95%): de -4,23604 a -0,68452. Exponenciando: exp(-4,23604) ≈ 0,0144 (redução de 98,6%) até exp(-0,68452) ≈ 0,5044 (redução de 49,6%). Assim, a redução plausível no tempo de permanência está entre cerca de 50% e 99%. Significância: t = -2,72, valor-p = 0,007. Também significante (p < 0,01).

Variável Transferência: ATT = -2,08434. Em termos percentuais: exp(-2,08434) - 1 ≈ -0,876, ou seja, redução de cerca de 87,6% nas transferências.

Intervalo de confiança (95%): de -3,60912 a -0,55957. Exponenciando: exp(-3,60912) ≈ 0,0272 (redução de 97,3%) até exp(-0,55957) ≈ 0,5717 (redução de 42,8%). Assim, a queda está provavelmente entre 43% e 97%. Significância: t = -2,68, valor-p = 0,007. Estatisticamente significativa.

Todos os três desfechos da produção emergencial apresentam efeitos negativos, que é o sinal esperado para eficiência, e estatisticamente significativos da intervenção Lean II, indicando uma forte queda relativa em óbitos, tempo de permanência e transferências nos hospitais tratados em comparação à contrafactual sintética. As magnitudes são grandes em termos percentuais (reduções na casa de 40–99%), o que sugere impacto substantivo da intervenção.

Magnitude vs. plausibilidade: reduções da ordem de 80–90% são expressivas; convém confrontá-las com os níveis absolutos pré-intervenção para entender o impacto em termos de unidades (por exemplo, quantos óbitos ou dias de permanência foram evitados). Cabe observar que esta é uma aplicação para apresentar a utilização do método.

O modelo SDID identifica que a intervenção Lean 2 está associada a reduções substanciais e estatisticamente significativas em óbitos, tempo de permanência e transferências. Contudo, dado que o período base é novembro de 2020 – ainda em contexto marcado por choques e adaptações decorrentes da pandemia de COVID-19 – e o tratamento ocorre apenas em novembro de 2023, parte da diferença observada pode refletir trajetórias assimétricas de recuperação entre hospitais tratados e de controle. Assim, embora os efeitos estimados sejam grandes, recomenda-se complementar a análise com verificações de robustez (testes placebo, avaliação de tendências prévias e controle por proxies de recuperação pós-COVID) para separar o efeito direto da intervenção Lean de possíveis descompassos de normalização das operações hospitalares.

8.2 SNM

Como discutido anteriormente, o SNM é um projeto que visa reduzir infecção hospitalar no Brasil. Um dos principais estudos que analisaram os resultados do programa foi o de Tuma et al. (2023). Segue a seguir um resumo deste artigo.

Infecções relacionadas à assistência à saúde continuam sendo um problema grave, com impacto significativo sobre morbidade, mortalidade e custos, mesmo quando existem medidas preventivas de baixo custo. O Ministério da Saúde do Brasil, por meio do Programa Nacional de Prevenção e Controle de IRAS e da Política Nacional de Segurança do Paciente, lançou a iniciativa colaborativa 'Saúde em Nossas Mãos', adotando o modelo *Breakthrough Series* (BTS) para reduzir infecções em unidades de terapia intensiva. O projeto foi implementado entre janeiro de 2018 e fevereiro de 2020, com análise de linha de base em 2017 para estabelecer taxas iniciais de infecção. Participaram hospitais públicos e sem fins lucrativos com estrutura mínima de qualidade, resultando na seleção de 120 UTIs, das quais 116 permaneceram na análise final. Essas unidades foram agrupadas em cinco hubs de mentoria.

A intervenção combinou componentes técnicos de prevenção de CLABSI, VAP e CA-UTI com metodologias de melhoria da qualidade. Equipes interdisciplinares foram treinadas, receberam coaching e participaram de seis sessões presenciais intercaladas por períodos de ação e suporte remoto mensal. As mudanças foram implementadas via ciclos *Plan-Do-Study-Act* (PDSA), com plataforma online de compartilhamento de materiais e monitoramento contínuo. Os indicadores incluíam desfechos de infecção (densidades de CLABSI, VAP e CA-UTI conforme definições oficiais) e adesão a *bundles* de prevenção, além de higiene das mãos. A adesão foi medida de forma '*all-or-nothing*' para os *bundles*, e a relação entre adesão e desfechos foi explorada com correlações de Pearson. Séries temporais foram modeladas com regressões que levaram em conta autocorrelação, utilizando abordagens tipo ARIMA e gráficos de controle para monitoramento das mudanças.

Os resultados foram substanciais. Observou-se redução de 43,5% na CLABSI, de 52,1% na VAP e de 65,8% na CA-UTI em comparação com a linha de base, com modelos temporais ajustando as tendências e mostrando significância estatística forte. Estima-se que mais de 5.000 infecções foram prevenidas no período analisado. A adesão aos *bundles*

teve correlação negativa significativa com as densidades de infecção: por exemplo, a inserção e manutenção de cateteres centrais esteve fortemente correlacionada com queda em CLABSI, a prevenção de VAP apresentou correlação robusta com redução de pneumonia associada à ventilação, e a avaliação diária da indicação e técnica de inserção/manutenção de cateter vesical explicaram a maior redução em CA-UTI. Alguns subcomponentes, como aspectos relacionados a curativos de CLABSI, não mostraram correlação significativa, possivelmente devido a limitações de suprimento ou mensuração. Já a higiene das mãos aumentou, mas sua correlação com os desfechos não foi estatisticamente significativa, sugerindo desafios de implementação isolada sem estratégias multimodais completas.

As reduções observadas foram superiores às tendências nacionais agregadas, indicando que a abordagem colaborativa produziu efeitos além do esperado de melhoria gradual. Com base em taxas de mortalidade atribuídas às infecções evitadas, estimou-se que aproximadamente 1.836 vidas foram potencialmente salvas até fevereiro de 2020. O sucesso foi atribuído à combinação de capacitação contínua, empoderamento das equipes, mentorias estruturadas e ciclos de *feedback* rápidos que permitiram ajustes locais.

Os autores destacam que o modelo BTS é viável em contextos de renda média para reduzir lacunas entre evidência e prática clínica, ao integrar aprendizado colaborativo, auditoria sistemática e adaptação local. A análise reconhece limitações importantes: o desenho não era originalmente um estudo acadêmico formal, não houve grupo controle estruturado, existiu potencial viés de seleção das UTIs participantes, faltaram dados sobre gravidade dos pacientes e composição da casuística, e a estimativa de vidas salvas utilizou taxas padronizadas sem ajuste fino para riscos basais. Heterogeneidade entre as unidades também limitou análises de subgrupos mais detalhadas.

Mesmo com essas limitações, a iniciativa demonstrou que intervenções de qualidade bem estruturadas, com monitoramento e suporte contínuo, podem gerar reduções amplas e clinicamente relevantes em infecções relacionadas à assistência, reforçando a importância da fidelidade na implementação dos *bundles* e da infraestrutura de melhoria da qualidade para sustentabilidade. O projeto serviu como um exemplo de como programas nacionais podem escalar práticas baseadas em evidência e salvar vidas em larga escala.

A seleção dos dados segue a literatura (Tuma et al, 2023), e gostaríamos de selecionar os dados de internação das CIDs descritas anteriormente. CLABSI (*central line–associated bloodstream infection*): T80.211 e variantes: *Bloodstream infection due to central venous catheter*. VAP (*ventilator-associated pneumonia*). J95.851 – Ventilator *associated pneumonia*. Por fim, CA-UTI (*catheter-associated urinary tract infection*). Códigos de complicação do cateter urinário: T83.51 e subitens, especialmente para cateter uretral de demora (*indwelling urethral catheter*).

Grande problema com estas classificações é a baixa notificação. Raramente as CIDs secundárias são preenchidas e a internação não ocorre pelas CIDs associadas ao efeito do SNM.

i. tempo de internação (dias_perm)
ii. Quantidade de dia de UTI no mês (uti_mes_to)

O exercício realizado para analisar impacto do SNM utilizou os hospitais do DF e ES, uma vez que o grupo de unidades tratadas é menor do que o possível controle. Os hospitais que fizeram parte do projeto estão na Tabela a abaixo. Portanto, a base de dados é composta por todos os hospitais na base SIH do DF e ES, para o mês de novembro de 2018, 2020, 2023. Como o foco é UTI foram apenas considerados hospitais com leitos para esta finalidade. Como realizado no LEAN, as variáveis objeto da análise de impacto quantitativa foram normalizadas por gênero e idade dos pacientes de UTI. A normalização é a previsão das variáveis de interesse do modelo de regressão individual (por paciente) controlados por idade, o quadrado da idade e o gênero dos pacientes. A base individual possui 114 mil observações, aproximadamente. Para analisar o impacto na produção hospitalar, a informação por paciente é reduzida para o uso médio de permanência de internação e uso de UTI por hospital.

TABELA 9

Hospitais do DF e ES participantes do SNM

Estado	Cidade	Hospital	CNES
DF	Brasília	Hospital da Criança de Brasília José Alencar (HCB)	6876617
DF	Brasília	Hospital São Mateus	6730914
DF	Brasília	Hospital Universitário de Brasília (HUB)	0010510
DF	Brasília	Hospital Regional de Santa Maria (HRSM)	5717515
DF	Brasília	Hospital de Base do Distrito Federal (HBDF)	0010456
ES	Vitória	Hospital Universitário Cassiano Antônio Moraes (HUCAM)	4044916
ES	Venda Nova do Imigrante	Associação Social Filantrópica Hospital Padre Máximo	2403331
ES	Serra	Hospital Estadual Dr. Dório Silva	2486199
ES	Cachoeiro de Itapemirim	Hospital Evangélico de Cachoeiro de Itapemirim (HECI)	2547821
ES	Cachoeiro de Itapemirim	Hospital Materno Infantil Francisco de Assis (HIFA)	2485729
ES	Guaçuí	Santa Casa de Misericórdia de Guaçuí	2447029
ES	Linhares	Hospital Rio Doce	2465833

Fonte: elaboração própria.

Na Tabela 10, abaixo, são apresentados os momentos da permanência e dos dias de UTI/mês dos hospitais participantes do SNM. Nesse momento da análise as variáveis não estão em log natural, enquanto nos modelos estão serão usadas com log natural. Naturalmente, permanência total é maior do que permanência na UTI/mês.

TABELA 10

Momentos das Variáveis Permanência e Dias de UTI no Mês, 2018 e 2023

Período	Variável	Obs	Média	Desvio Padrão	Min	Max
2018	Perm.	7775	5.316498	1.75018	3.397511	13.41343
2018	UTI/m	7775	0.430148	0.266729	0.120883	1.70364
2023	Perm.	10791	5.653676	1.935522	2.994402	13.34195
2023	UTI/m	10791	0.754864	0.449962	0.229304	3.077531

Fonte: elaboração própria.

Fazendo um teste de média entre 2018 e 2023 se observa que para permanência, a diferença média de aproximadamente 0,337 é altamente significativa ($t \approx 12{,}4$; $p \approx 4.4\times10^{-35}$). Para UTI/mês, a diferença média

de aproximadamente 0,325 é extremamente significativa ($t \approx 61,5$; $p < 10^{-40}$). Os intervalos de confiança não incluem zero em nenhum caso, confirmando diferenças robustas entre 2018 e 2023.

Em seguida é estimado o modelo diferença-em-diferenças.

Abaixo estão os resultados utilizando o estimador SDID para avaliar o impacto quantitativo do SNM. A avaliação foi feita para dois períodos no grupo de controle. Como explicado em Tuma et al, a aplicação do SNM começa em 2020 e termina em 2023, fazendo sentido analisar os dois cenários. No cenário que inclui 2020 no grupo de controle, se assume que os agentes implementam rapidamente a mudança proposta, enquanto quando se assume 2023, se parte da premissa de que o treinamento toma tempo para gerar os efeitos.

TABELA 11

Estimação de Modelo DID para Duas Variáveis de Tratamento e Dois Períodos de Controle

Período de tratamento	Variável dependente	ATT	Erro Padrão	t	P>|t|	95% CI inferior	95% CI superior
2020, 2023	Perm.	-0.01286	0.063397	-0.2	0.839	-0.13764	0.111918
2023		-0.04095	0.060973	-0.67	0.502	-0.16096	0.079062
2020, 2023	UTI/mês	-0.0536	0.091782	-0.58	0.56	-0.23425	0.127055
2023		-0.06272	0.093351	-0.67	0.502	-0.24646	0.121019

Fonte: elaboração própria.

Os resultados do modelo DID indicam que o parâmetro associado ao tratamento é muito próximo de zero e sem significância estatística em ambos os desfechos e janelas temporais. Para a permanência no período de tratamento 2020 e 2023, o coeficiente de -0,01286 com erro-padrão de 0,063397 gera t = -0,20 e p = 0,839, enquanto o intervalo de confiança de 95% varia de -0,13764 a 0,111918. Isso indica que a redução pontual de cerca de 1,29% no log de permanecia é compatível com efeito nulo dentro de uma incerteza que comporta aumento de até 11,19% ou queda de até 13,76%. Ao restringir o tratamento ao ano de 2023, o coeficiente fica em -0,04095 com erro-padrão de 0,060973, t = -0,67 e p = 0,502, e o intervalo de confiança de -0,16096 a 0,079062 continua en-

globando zero e ampla variação possível. Para a variável UTI por mês no período de 2020 e 2023, o coeficiente estimado é -0,05360 com erro-padrão de 0,091782, t = -0,58 e p = 0,560, e o intervalo de confiança de 95% se estende de -0,23425 a 0,127055. Isso sugere uma redução pontual de 5,36% no log de UTI/mês, mas sem evidência estatística, uma vez que a incerteza abarca tanto reduções substanciais de até 23,43% quanto aumentos de até 12,71%. Quando se analisa apenas 2023, o coeficiente passa a -0,06272 com erro-padrão de 0,093351, t = -0,67 e p = 0,502, e o intervalo de -0,246459 a 0,121019 reforça a falta de significância.

Em todos os cenários, os p-valores muito altos e os intervalos de confiança amplos indicam ausência de efeito detectável do tratamento. As magnitudes pontuais modestamente negativas (de cerca de -1,3% a -6,3%) não se sustentam diante da variabilidade dos dados.

A seguir são apresentadas as estimações do efeito tratamento utilizando o estimador Synthetic *Difference-in-Differences*, que é mais promissor, pois permite construir um grupo de controle sintético que garante tendências paralelas pré-intervenção.

TABELA 12

Resultados do Efeito Líquido do Tratamento (Lean II) – Modelo Diferença-em-Diferenças Sintética (SDID)

Período de tratamento	Variável	ATT	Erro-padrão	t	P>|t|	95% CI inferior	95% CI superior
2023	Perman.	0.02284	0.03897	0.59	0.558	-0.05355	0.09923
2023	UTI/mês	-0.00588	0.06362	-0.09	0.926	-0.13058	0.11882
2020, 2023	Perman.	-0.02088	0.03300	-0.63	0.527	-0.08556	0.04379
2020, 2023	UTI/mês	-0.04927	0.05720	-0.86	0.389	-0.16139	0.06285

Fonte: elaboração própria.

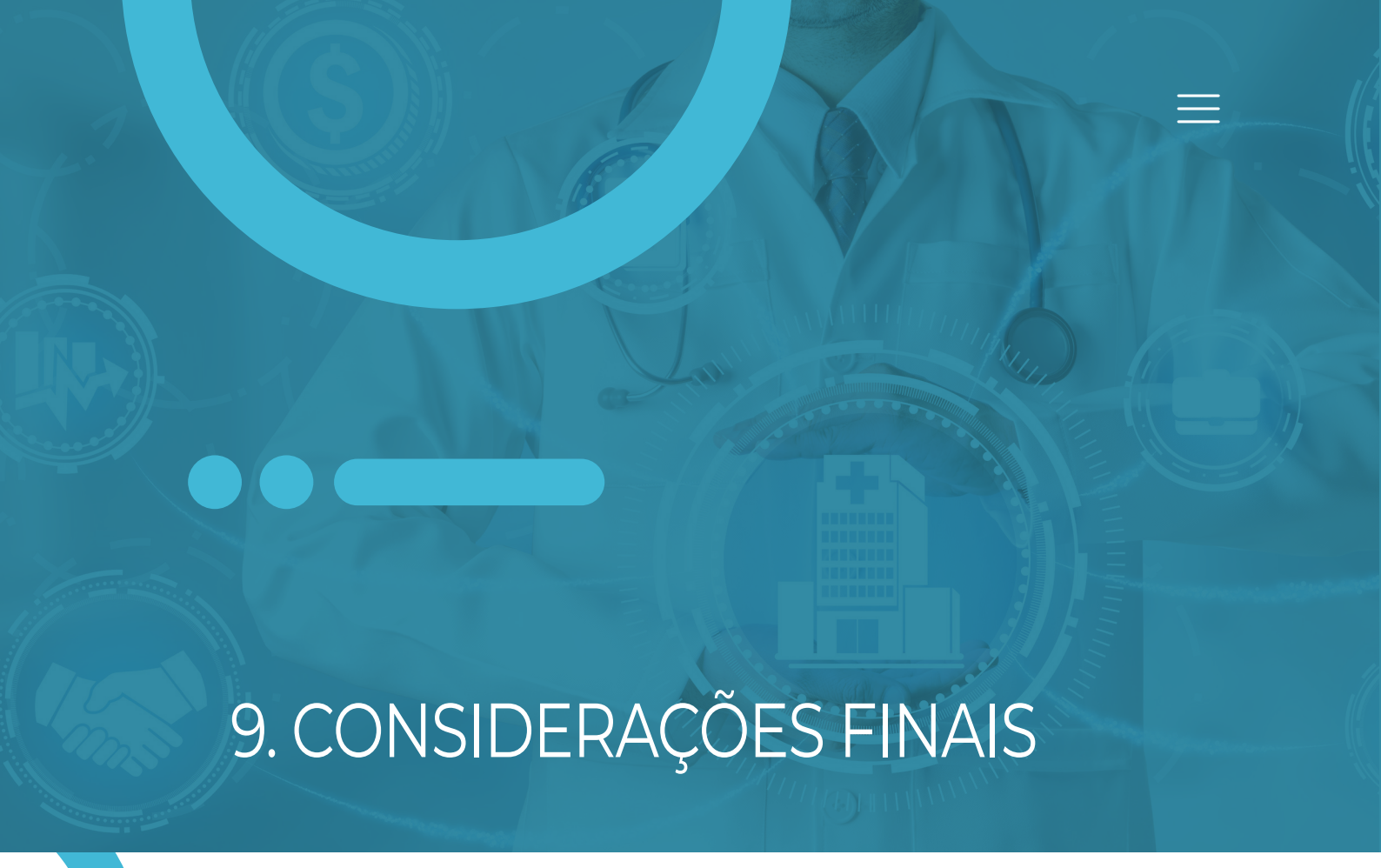
Os resultados do estimador *Synthetic Difference-in-Differences* para as variáveis em log natural (perman. e UTI/mês) mostram que, nas especificações analisadas, não há evidência estatisticamente significativa de efeito do tratamento sobre os desfechos considerados. A inferência foi baseada em um procedimento de placebo com 200 simulações. Esse procedimento gera uma distribuição empírica dos efeitos

esperados sob a hipótese nula, reatribuindo falsos tratamentos a unidades ou períodos que não deveriam ser tratados e recalculando o ATT em cada rodada. A partir dessa distribuição, calcula-se o erro-padrão como a dispersão dos ATT de placebo e os p-valores são obtidos como a proporção de simulações de placebo com efeito igual ou mais extremo que o observado. Essa abordagem tem a vantagem de capturar a estrutura real dos dados sem depender de pressupostos paramétricos fortes. No entanto, o uso de apenas 200 simulações introduz alguma incerteza de Monte Carlo nas estimativas dos erros-padrão e p-valores; recomenda-se aumentar para 500 ou 1000 simulações para testar a estabilidade.

Para a variável permanência com período de tratamento restrito a 2023, o ATT é 0,02284. Em log natural, isso corresponde a um aumento percentual aproximado de 2,31% (exp(0,02284)-1). O intervalo de confiança de 95% vai de -0,05355 a 0,09923, equivalente a uma redução de até 5,22% ou aumento de até 10,44%. O t-estatístico de 0,59 e o p-valor de 0,558 indicam falta de significância. Ao ampliar para 2020 e 2023, o ATT para perm sobe para 0,03765 (?3,84%), intervalo de -0,04945 a 0,12476 e p-valor de 0,397, ainda não significativo.

Para UTI/mês em 2023, o ATT é -0,00588 (?-0,59%), intervalo de -0,13058 a 0,11882 e p-valor de 0,926. Com 2020 e 2023, ATT=0,03388 (?3,45%), intervalo -0,08751 a 0,15527 e p-valor de 0,584, também não significativo. Os coeficientes pontuais são pequenos e os intervalos de confiança amplos, permitindo aumentos ou reduções que poderiam ser relevantes.

Em resumo, os coeficientes para permanência variam entre 2,31% e 3,84% e para UTI/mês entre -0,59% e 3,45%, dependendo do período, mas nenhum é estatisticamente significativo. A variância estimada por placebo com 200 simulações indica dispersão compatível com ruído sob a hipótese nula. Como mostrado aqui, nenhum resultado para o projeto SNM foi significante, indicando que o programa não gerou redução significante na permanência de pacientes e nem em uso de UTI nos estados do DF e ES.



9. CONSIDERAÇÕES FINAIS

Esse texto busca apresentar as principais ferramentas de avaliação de impacto para subsidiar as ações do Ministério da Saúde. Foram apresentados os principais métodos apontando as vantagens e as limitações de cada um deles. Na sequência foi realizada uma apresentação formal do método de diferenças em diferenças que é o mais utilizado em trabalhos do tipo. Na apresentação foi discutido que estimativas do tipo antes e depois e de comparação simples podem ser obtidas como passos intermediários do método de diferença em diferenças. A apresentação termina com o exemplo do efeito do aumento do salário-mínimo em Nova Jérsei discutido em Card e Krueger (1994).

Na sequência fizemos uma apresentação mais geral do estimador de diferenças em diferenças permitindo o uso de dados em painel. Foi feita uma discussão sobre grupos de controle e apresentação formal do método de controle sintético.

Finalmente foi discutido o método de diferenças em diferenças sintético que representa o estado da arte na literatura de inferência causal. A discussão envolveu uma apresentação formal do SDID, destacando a relação desse estimador com o SC e o DiD. Após a discussão formal, foi apresentado o caso do controle de tabaco na Califórnia para ilustrar os três métodos propostos.

Não é pretensão desse trabalho determinar um método definitivo para ser usado em toda e qualquer avaliação no âmbito do Ministério, em vez disso, o trabalho apresenta um cardápio de métodos que podem ser utilizados de acordo com as características de cada caso real em análise. Dificuldades na coleta de dados, restrição de tempo ou de pessoal, características especificas da política em análise podem fazer com que um método menos sofisticado, como “antes e depois” ou comparação simples, seja o mais adequado para uma avaliação específica. A ideia geral é que uma avaliação simples pode ser preferível a nenhuma avaliação.

10. REFERÊNCIAS

Abadie, Alberto e Javier Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, v.93, 2003.

Abadie, Alberto, Diamond, e Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, v. 105 (490), 2010.

Angrist, Joshua e Jorn-Steffen Pischke. *Mostly Harmless Econometrics*: An Empiricist's Companion. Princeton, Princeton University Press, 2009.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, e Stefan Wager. Synthetic Difference in Differences. *American Economic Review*, 2021.

Athey, Susan e Guido W. Imbens. The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Literature*, v.31 (2), 2017.

Athey, Susan e Guido W. Imbens. Design-based Analysis in Difference--in-Differences Settings with Staggered Adoption. *Journal of Econometrics*, 2021.

Athey, Susan e S. Stern. The impact of information technology on emergency health care outcomes. *The RAND Journal of Economics*, v.33 (3), 2002.

Bertrand, Marianne, Ester Duflo, e Sello Mullainathan. Should We Trust in Differences-in-Differences Estimates? *Quarterly Journal of Economics*, 2004.

Callaway, Brantly e Pedro H.C. Sant'Anna. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*, v.225 (2), 2021.

Card, David e Alan Krueger. Minimum wages and employement: a case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 1994.

Hastie, Trevor, Robert Tibshirani, e Jerome Friedman. *The Elements of Statistical Learning*: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2017.

Hayashi, Fumio. *Econometrics*. Princeton, Princeton University Press, 2000.

Imai, Kosuke e Nora Webb Williams. *Quantitative Social Science*: an introduction in tidyverse. Princeton University Press, 2022.

McCullough, Jeffrey S., Stephen T. Parente, e Robert Town. Health information technology and patient outcomes: the role of information and labor coordination. *RAND Journal of Economics*, v.47 (1), 2016.

Mundlak, Yair. On the Pooling of Cross Section and Time Series Data. *Econometrica*, v.46, 1978.

Romano, Joseph P. e Michael Wolf. Ressurecting Weighted Least Squares. *Journal of Econometrics*, v.197, 2017.

Roth, Jonathan e Pedro H.C. Sant'Anna. When Is Parallel Trends Sensitive to Functional Form? *Econometrica*, v.91 (2), 2023.

Tuma, Paula, et al. *A National Implementation Project to Prevent Healthcare-Associated Infections in Intensive Care Units*: A Collaborative Initiative Using the Breakthrough Series Model. Open Forum Infect Dis, Mar, 2023.

Wooldridge, Jeffrey. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MIT Press, 2010.

Wooldridge, Jeffrey. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimator. *Working Paper*, 2021.

ceag.unb.br @unb_oficial @ceag_unb ceag@unb.br

CEAG-UnB
Centro de Estudos Avançados de Governo e Administração Pública

UnB

BRASIL BEM CUIDADO
MAIS SAÚDE PARA QUEM MAIS PRECISA

SUS 35 ANOS

MINISTÉRIO DA SAÚDE

GOVERNO DO BRASIL
DO LADO DO POVO BRASILEIRO