BNDES O banco nacional do desenvolvimento

textos para discussão

130 | Outubro de 2018

Uma solução automatizada para avaliações de impacto em estudos de caso: o Modelo Automatizado em R para Verificação de Impacto (MARVIm) – Módulo de Controle Sintético

Ricardo Agostini Martini
Daniel da Silva Grimaldi
Marília de Figueiredo Jordão
João Paulo de Oliveira Pereira
Marcus Magno Fernandes Tortorelli

Presidente do BNDES
Dyogo Henrique de Oliveira

Diretoria de Transformação Estratégica e Digital
Ricardo Ramos

Área de Planejamento Estratégico
Mauricio dos Santos Neves

textos para discussão

130 | Outubro de 2018

Uma solução automatizada para avaliações de impacto em estudos de caso:

o Modelo Automatizado em R para Verificação de Impacto (MARVIm) – Módulo de Controle Sintético

Ricardo Agostini Martini
Daniel da Silva Grimaldi
Marília de Figueiredo Jordão
João Paulo de Oliveira Pereira
Marcus Magno Fernandes Tortorelli

Resumo

O objetivo do presente trabalho é apresentar o Modelo Automatizado em R para Verificação de Impacto (MARVIm) Módulo de Controle Sintético, uma metodologia automatizada de avaliação de efeitos de intervenções públicas. Para isso, combina uma série de *scripts* e funções em R que realizam avaliações baseadas em controle sintético. Esse método de avaliação foi orginalmente desenvolvido com o objetivo de detectar efeitos de uma política ou de um evento sobre algum indicador de interesse de uma unidade exposta à intervenção. Portanto, o objetivo desse módulo do MARVIm é servir como um modelo automatizado para avaliações de impacto em intervenções com poucas unidades tratadas, sejam elas firmas ou unidades geográficas. O presente trabalho inclui um estudo de caso com o objetivo de verificar possíveis impactos da construção de usinas eólicas nas economias municipais no período de 2007 a 2014. O método do controle sintético foi aplicado para cada caso, somando 37 avaliações individuais, que foram então compiladas e analisadas de maneira agregada. Os resultados obtidos se mostraram heterogêneos, mas com efeitos medianos positivos, concordando com a bibliografia levantada.

Palavras-chave: Avaliação de impacto. Controle sintético. *Machine learning*. Programação em R. Usinas eólicas.

Abstract

The objective of the present work is to present the Modelo Automatizado em R para Verificação de Impacto (MARVIm) Synthetic Control Module, a methodology for evaluating the effects of public interventions in an automated way. To do this, it combines a series of scripts and R-functions that carry out evaluations based on synthetic control. This method of evaluation was originally developed with the objective of detecting the effects of a policy or an event on some indicator of interest of a unit exposed to the intervention. Therefore, the objective of this module of MARVIm is to serve as an automated model for impact assessments in interventions with a few treated units, whether they are firms or geographic units. The present work includes a case study with the objective of verifying possible impacts of the construction of wind power plants in the municipal economies from 2007 to 2014. The synthetic control method was applied for each case, adding 37 individual evaluations, which were then compiled and analyzed in an aggregate manner. The results obtained were heterogeneous, but with median positive effects, agreeing with the bibliography raised.

Keywords: Impact evaluation. Synthetic control methods. Machine learning. R Programming. Wind power plants.

Sumário

Ricardo Agostini Martini é economista do BNDES e mestre em Economia pelo Centro de Desenvolvimento e Planejamento Regional da Universidade Federal de Minas Gerais. Daniel da Silva Grimaldi é economista do BNDES e mestre em Economia pela Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo. Marília de Figueiredo Jordão é estagiária do BNDES e graduanda em Estatística na Universidade Federal Fluminense. João Paulo de Oliveira Pereira é economista do BNDES e bacharel em Economia pela PUC-Rio. Marcus Magno Fernandes Tortorelli é analista de sistemas do BNDES e mestre em Pesquisa Operacional pela PUC-Rio.

1. Introdução

O objetivo do presente trabalho é apresentar o Modelo Automatizado em R para Verificação de Impacto (MARVIm) Módulo de Controle Sintético, uma metodologia automatizada de avaliação de efeitos de intervenções públicas. Para isso, combina uma série de *scripts* e funções em R que automatizam avaliações baseadas em controle sintético. Esse método de avaliação foi orginalmente desenvolvido para estudos de caso, isto é, com o objetivo de detectar efeitos de uma política ou de um evento sobre algum indicador de interesse de uma unidade exposta à intervenção,[1] ou seja, consiste em um método adequado para os casos em que o problema de micronumerosidade dificulta inferências estatísticas tomadas em exercícios de avaliação de impacto quando realizados pelos métodos tradicionais de pareamento e de diferenças em diferenças. Portanto, o objetivo desse módulo do MARVIm é servir como um modelo automatizado para avaliações de impacto em intervenções com poucas unidades tratadas, sejam elas firmas ou unidades geográficas. Seu maior desafio é precisamente sistematizar e automatizar esse método, de modo a permitir que seja aplicável a um grande conjunto de intervenções em unidades cujas características estão mensuradas em diferentes bases de dados.

O desenvolvimento do MARVIm segue a tendência de uma crescente preocupação com a efetividade das políticas públicas. Nos últimos anos, em um esforço por maior eficiência do gasto público, avaliações de impacto de intervenções públicas têm se tornado cada vez mais comuns. As avaliações também cumprem o fundamental papel de gerar informações para que políticas sejam aperfeiçoadas.

Para o BNDES, a tendência de preocupação com a efetividade das intervenções públicas representa uma oportunidade para expandir suas atividades de monitoramento e avaliação. No entanto, essa oportunidade acarreta a necessidade de se enfrentarem dois desafios. Em primeiro lugar, o desafio de capturar a causalidade de cada intervenção realizada, isto é, de verificar se a relação entre a política realizada e o comportamento dos indivíduos afetados por ela pode ser compreendida como uma relação de causa e efeito. Em segundo lugar, o desafio de dar escala às avaliações de efetividade, com o objetivo de permitir o processamento de um grande volume de dados em reduzido espaço de tempo e com a necessidade de se evitar a repetição de esforços.

Para que os desafios sejam superados, é necessário ter em consideração que a avaliação de políticas públicas não consiste apenas na mensuração de um determinado conjunto de indicadores de desempenho socioeconômico em si, mas também procura responder à seguinte pergunta: o que teria acontecido com os indivíduos (sejam pessoas, sejam empresas, sejam unidades geográficas) alvos da intervenção

[1] Ver, por exemplo, Abadie e Gardeazabal (2003) e Abadie, Diamond e Hainmueller (2010; 2011; 2015).

pública caso não tivessem sido beneficiados com o projeto? Em outras palavras, a avaliação busca verificar o impacto da política pública.

Essa atividade parte do desenho de um experimento. Assim, elegem-se dois grupos comparáveis de indivíduos: um "grupo tratado", apoiado pela intervenção, e um "grupo de controle", não apoiado. Esse desenho permite que se possa medir o efeito do tratamento sobre a população além do que provavelmente teria acontecido sem ele. Esse cenário alternativo é denominado de "contrafactual". Ao se considerar não apenas a evolução do grupo tratado no tempo, antes e depois da intervenção, mas também a tendência natural que esse grupo seguiria, evidenciada pela evolução do grupo de controle (o contrafactual), torna-se possível mensurar o impacto da intervenção.

Nesse sentido, o MARVIm é um modelo automatizado e programado na linguagem R, desenvolvido pelo Departamento de Avaliação de Promoção da Efetividade (DEAPE), da Área de Planejamento Estratégico (AP) do BNDES. Seu objetivo é consolidar uma série de métodos padronizados de avaliação de impacto de intervenções de modo a providenciar uma primeira resposta ao sistema de monitoramento e avaliação do Banco. Mais especificamente, o MARVIm procura ser uma ferramenta de levantamento de informações para uma análise preliminar de impacto, de modo a atender às crescentes demandas de avaliação pelo BNDES.

O primeiro módulo do MARVIm foi voltado para métodos baseados em pareamento e diferenças em diferenças.[2] Especificamente para esse modelo, foi construída uma base de dados no nível da firma que consolida informações sobre emprego, exportações e dados contábeis. Seu objetivo foi o de buscar inferir o impacto das políticas do BNDES controlando dois distintos processos de seleção que fazem com que o apoio do Banco não seja aleatoriamente distribuído entre as empresas, que poderiam enviesar estimativas empíricas realizadas por métodos tradicionais de regressão. Primeiro, há o processo de seleção das próprias empresas, cujas decisões de procurar ou não o apoio do BNDES dependem de suas próprias características. Segundo, há o processo de decisão do BNDES ou, em caso de operações indiretas, do repassador por ele autorizado, que escolhe fazer ou não uma operação com base nas características das firmas demandantes. Para controlar esses processos, os métodos incorporados no primeiro módulo do MARVIm buscam comparar as firmas apoiadas com firmas não apoiadas identificadas na base de dados de análise, mas com características semelhantes. O grau de semelhança entre as firmas dos dois grupos é estimado pelo cálculo de uma regressão logística que busca explicar a probabilidade de uma empresa ser apoiada pelo Banco com base em suas características observáveis. Com essa estimação, a avaliação de impacto

[2] Ver Albuquerque *et al*. (2017).

é realizada por meio do algoritmo do vizinho mais próximo, de modo que cada unidade tratada seja comparada com a unidade de controle com probabilidade estimada de tratamento mais próxima a ela.

Todavia, é necessário levar em consideração o fato de que, para que os resultados estimados no exercício empírico possam ser expandidos para uma análise da política efetuada, é necessário que as estimativas obtidas por essa técnica permitam inferências estatísticas confiáveis. Isto é, é preciso que se tenha confiança de que os resultados obtidos para a amostra refletem o efeito da intervenção para a população como um todo. Para isso, é fundamental que a amostra de unidades tratadas tenha um volume apropriado. Ou seja, não pode haver micronumerosidade de tratados. Esse ponto é sensível para muitas das intervenções realizadas pelo BNDES, como para o caso do impacto local, em nível geográfico, das grandes obras de infraestrutura.

Nesse sentido, o presente texto apresenta outra metodologia, adequada aos casos em que as intervenções atingem um número pequeno de unidades. Essa metodologia é a estimação por controle sintético, a qual procura comparar cada unidade tratada não com uma unidade de controle específica com características consideradas semelhantes com base em um modelo logístico, mas sim com uma unidade artificial de controle que consiste em uma média ponderada de todas as unidades elegíveis para comparação disponíveis na base de dados. Essa metodologia também foi automatizada em um modelo em linguagem R, e foi denominada de MARVIm Módulo de Controle Sintético.

Em linhas gerais, o MARVIm Módulo de Controle Sintético recebe do avaliador alguns parâmetros para a análise de impacto. Dentre eles, é necessária uma base de dados contendo observações das unidades que serão consideradas na avaliação, assim como os anos de tratamento e as variáveis de interesse. A partir desses insumos, o modelo pode ser resumido em uma série de etapas. Primeiro, são selecionadas as variáveis de controle que melhor preveem a variável de interesse, por meio do método adaLASSO (do inglês *adaptive lasso*, ou lasso adaptativo). Depois disso, as unidades potenciais de comparação são alocadas a *clusters*, constituídos pela maior similaridade das unidades dentro de seu grupo e maior diferença entre os grupos. Em seguida, a base de dados é reduzida apenas ao *cluster* que contém a unidade tratada, e assim é construído o controle sintético que é comparado à unidade tratada. A diferença entre o desempenho da unidade tratada e seu controle sintético após o tratamento é considerado o efeito da intervenção. Por fim, se o exercício envolver mais de uma unidade tratada, as avaliações individuais são compiladas e os resultados agregados são observados. O objetivo do MARVIm é possibilitar ao pesquisador um conjunto de informações de modo a permitir uma análise preliminar sobre possíveis efeitos de um tratamento, assim como sobre a magnitude desses efeitos.

O presente trabalho inclui um estudo de caso com o objetivo de verificar possíveis impactos da construção de usinas eólicas nas economias municipais. Mais especificamente, pretende-se aqui analisar a evolução do produto interno bruto (PIB) *per capita* dos municípios que receberam investimentos na construção de usinas eólicas em comparação a um grupo de outros municípios com características semelhantes, mas que não recebeu essa intervenção. O período coberto pela avaliação vai de 2007 a 2014, em razão da melhor disponibilidade de dados nesse período. O método do controle sintético foi aplicado para cada caso, somando 37 avaliações individuais, que foram então compiladas e analisadas de maneira agregada. Os resultados obtidos se mostraram heterogêneos, mas com efeitos medianos positivos, concordando com a bibliografia levantada. Os efeitos compilados são mais favoráveis em dois anos após o início das obras civis das estruturas das usinas e em obras de parques eólicos com investimentos mais elevados.

O estudo aqui realizado utilizou uma base de dados municipais construída como uma consolidação de 250 variáveis quantitativas provenientes de 12 fontes diferentes. Essas fontes são: o Instituto Brasileiro de Geografia e Estatística (IBGE); o Sistema de Informações sobre Mortalidade (SIM) e o Cadastro Nacional de Estabelecimentos de Saúde (Cnes), ambos do Departamento de Informática do Sistema Único de Saúde (Datasus); informações da Estatística Bancária por Município (Estban); dados da Secretaria de Comércio Exterior (Secex); órgão do Ministério do Desenvolvimento, Indústria e Comércio Exterior (Mdic); informações do Instituto Nacional de Pesquisas Espaciais; a Relação Anual de Informações Sociais (Rais), do Ministério do Trabalho e Emprego; informações financeiras dos municípios (Finbra); banco de dados criado pela Secretaria do Tesouro Nacional (STN), em convênio com a Caixa Econômica Federal (CEF); o Censo Escolar e o Censo do Ensino Superior, ambos do Instituto Nacional de Estudos e Pesquisas (Inep); o Sistema Nacional de Informações sobre Saneamento (Snis), além de uma série de variáveis derivadas construídas com o cruzamento dessas variáveis brutas. Por fim, o banco de dados da Agência Nacional de Energia Elétrica (Aneel) forneceu informações sobre os valores de potência outorgada para todas as usinas eólicas avaliadas, assim como o ano de construção e o município em que está localizado o parque eólico.

O texto está organizado da seguinte maneira: esta introdução é seguida pela descrição da base de dados municipais consolidada, assim como as principais informações de suas fontes. A metodologia é apresentada a seguir, na qual são expostos os modelos teóricos da estimação por controle sintético e do método adaLASSO de seleção de variáveis de controle. Em seguida, é feita a descrição passo a passo de uma estimação do novo módulo do MARVIm. Depois, expõe-se o estudo de caso, que inclui uma descrição sobre o panorama da energia eólica

no Brasil, com base em dados da Aneel, uma revisão da bibliografia sobre seus impactos em nível local, estatísticas descritivas do exercício empírico proposto, uma avaliação individual e os resultados das estimações individuais compiladas. Por fim, serão apresentadas as considerações finais do trabalho e possíveis desdobramentos futuros dessa linha de pesquisa.

2. Base de dados

Para cumprir o desafio de entregar um extenso conjunto de avaliações de impacto em reduzidos períodos de tempo, é fundamental que se tenha à disposição bases de dados capazes de cobrir um grande número de unidades. Nesse sentido, a metodologia de avaliação de impacto que será desenvolvida no presente trabalho, denominada Modelo Automatizado em R para Verificação de Impacto (MARVIm Módulo de Controle Sintético), é compatível com uma ampla gama de bases de dados. Os pré-requisitos exigidos para a possibilidade de uma avaliação são a presença de um indicador que identifique as unidades individuais, um indicador que identifique o período de tempo e uma variável de interesse com alguma variabilidade (isto é, de desvio-padrão não nulo). Além disso, devem existir dados para, no mínimo, cinco anos antes e um ano depois do tratamento. Como o cálculo do controle sintético é mais confiável quanto melhor for o ajuste das curvas pré-tratamento, esse ajuste depende da disponibilidade dos dados ao longo do tempo e da presença de covariadas que colaborem para explicar a variável de interesse.

Nesse sentido, tendo em vista a necessidade de realizar avaliações em âmbito local de algumas políticas do BNDES, o Departamento de Avaliação e Promoção da Efetividade (DEAPE) construiu uma base de dados dos municípios brasileiros. Essa base é uma consolidação de informações de diversas fontes, que serão descritas a seguir. A base municipal conta com um total de 269 variáveis, incluindo um identificador de ano, um identificador de código do município no IBGE, um indicador do nome do município, 16 variáveis de tipificação e 250 variáveis numéricas que podem servir como indicadores de interesse ou covariadas para as estimações de controle sintético.

As variáveis de tipificação têm a finalidade de servirem como possíveis filtros para determinadas avaliações. Por exemplo, em alguns exercícios sobre desmatamento, é desejável calcular o impacto de uma intervenção sobre um município comparando seu desempenho com o de outro município da região amazônica. As tipificações disponíveis na base de dados incluem a unidade da federação e seu código no IBGE, as regiões imediatas, intermediárias e ampliadas de articulação urbana tal como definidas pelo documento Divisão Urbano-Regional (IBGE, 2013),

a situação urbana ou rural do município segundo os Censos de 2000 e de 2010, o bioma e a situação urbana ou rural do município de acordo com o Instituto Interamericano de Cooperação para a Agricultura (IICA), o nível de centralidade municipal segundo o Projeto Regiões de Influência das Cidades (IBGE, 2007) e duas medidas da área do município, uma adotada pelo IBGE, outra adotada pelo Instituto Nacional de Pesquisas Espaciais (Inpe), especificamente para a região amazônica.

As 250 variáveis numéricas são provenientes de 12 fontes diferentes de informações municipais, as quais incluem:

- O Sistema de Informações sobre Mortalidade (SIM), do Departamento de Informática do Sistema Único de Saúde (Datasus), com dados de causas de mortes.

- O Cadastro Nacional de Estabelecimentos de Saúde (Cnes), também do Datasus, com informações sobre número de estabelecimentos e equipamentos de saúde.

- Informações de economia e população do IBGE, incluindo o PIB do município e sua desagregação nos setores de agropecuária, indústria, serviços e administração pública, tanto em proporção quanto em valor agregado.

- Informações da Estatística Bancária por Município (Estban), incluindo o número de agências bancárias e o montante de empréstimos, financiamentos e outras operações bancárias, além do ativo e passivo total do setor.

- Informações sobre importações e exportações da Secretaria de Comércio Exterior (Secex), órgão do Mdic.

- Informações sobre desmatamento e cobertura florestal na Amazônia Legal, do Inpe.

- A Relação Anual de Informações Sociais (Rais), do Ministério do Trabalho e Emprego, com informações sobre número de empregados, vínculos empregatícios e de estabelecimentos, salário médio e média de idade dos empregados, com informações consolidadas em âmbito municipal.

- Informações financeiras dos municípios (Finbra), banco de dados criado pela Secretaria do Tesouro Nacional (STN), em convênio com a Caixa Econômica Federal (CEF). Essa base inclui informações sobre receitas e despesas públicas desagregadas por função. Inclui também ativos, passivos, patrimônio líquido e resultados acumulados da administração pública.

- Dados do Censo Escolar, do Inep, com informações sobre número de escolas, turmas e alunos matriculados, assim como sua média de idade.

- Dados do Censo do Ensino Superior, também do Inep, com informações sobre o número de estabelecimentos e de alunos, distribuídos por turno e por condição (entrantes ou concluintes), assim como sua média de idade.

- Dados do Sistema Nacional de Informações sobre Saneamento (Snis) sobre os sistemas de água, esgoto e coleta de resíduos e o acesso a eles.

- Além disso, a base municipal conta com variáveis derivadas, construídas a partir do cruzamento de dois ou mais indicadores de diferentes fontes. A maior parte dessas variáveis indica valores integrais, como proporção da população, das receitas do governo ou da área do município. Por exemplo, tem-se o PIB *per capita*, as despesas com habitação como proporção da receita corrente líquida da prefeitura, a área desmatada sobre a área total e a densidade demográfica do município.

As principais informações sobre essas fontes de dados estão registradas na Tabela 1. Nessa tabela estão representados, por fonte de dados, o número de variáveis, o menor ano inicial de cobertura, o maior ano final e as médias de preenchimento, tanto das variáveis em si como a proporção dos municípios cobertos com pelo menos uma observação.

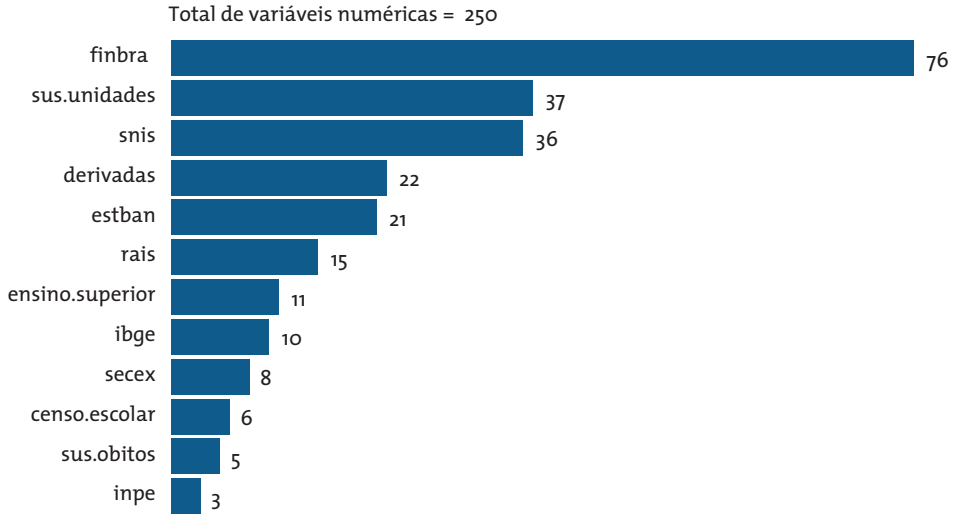
Tabela 1. Informações gerais sobre as fontes das variáveis municipais

Fonte	Variáveis	Ano inicial	Ano final	Preenchimento (%)	Cobertura nos municípios (%)
Censo Escolar	6	2007	2015	99,9	100,0
Derivadas	22	1999	2016	71,0	77,8
Ensino superior	11	2009	2014	85,5	86,0
Estban	21	1999	2015	61,7	69,5
Finbra	76	2002	2015	83,7	96,3
IBGE	10	1999	2016	99,5	99,6
Inpe	3	2000	2015	13,6	13,6
Rais	15	2002	2014	99,8	100,0
Secex	8	2000	2016	56,3	67,5
Snis	36	1999	2014	49,8	80,5
Óbitos	5	1999	2014	93,9	100,0
Unidades SUS	37	2005	2015	56,4	56,4

Fonte: Elaboração própria.

Do total das 250 variáveis numéricas presentes na base de dados, o destaque são as contas públicas municipais, com um total de 76 indicadores, seguido do registro de unidades do SUS, com 37 indicadores.

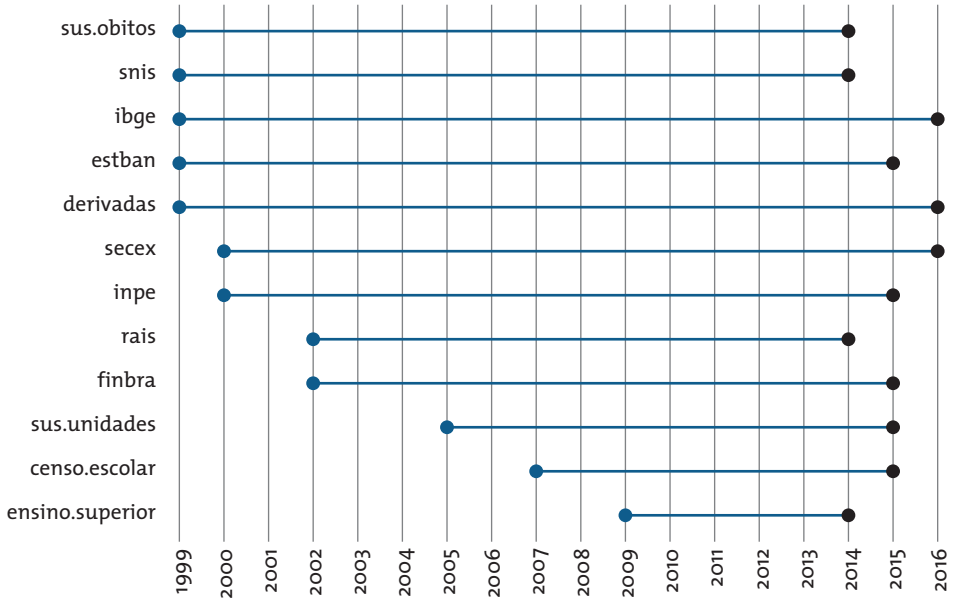
Gráfico 1. Número de variáveis por fonte de dados



Fonte: Elaboração própria.

Quanto à cobertura temporal, a base de dados municipais consolidada abrange o período total de 1999 a 2016. O destaque são as bases do IBGE e as variáveis derivadas, que cobrem o período inteiro. Por outro lado, os censos das escolas e do ensino superior são relativamente mais restritos. De forma geral, a base é mais completa no período de 2002 a 2014.

Gráfico 2. Cobertura anual máxima das fontes de dados

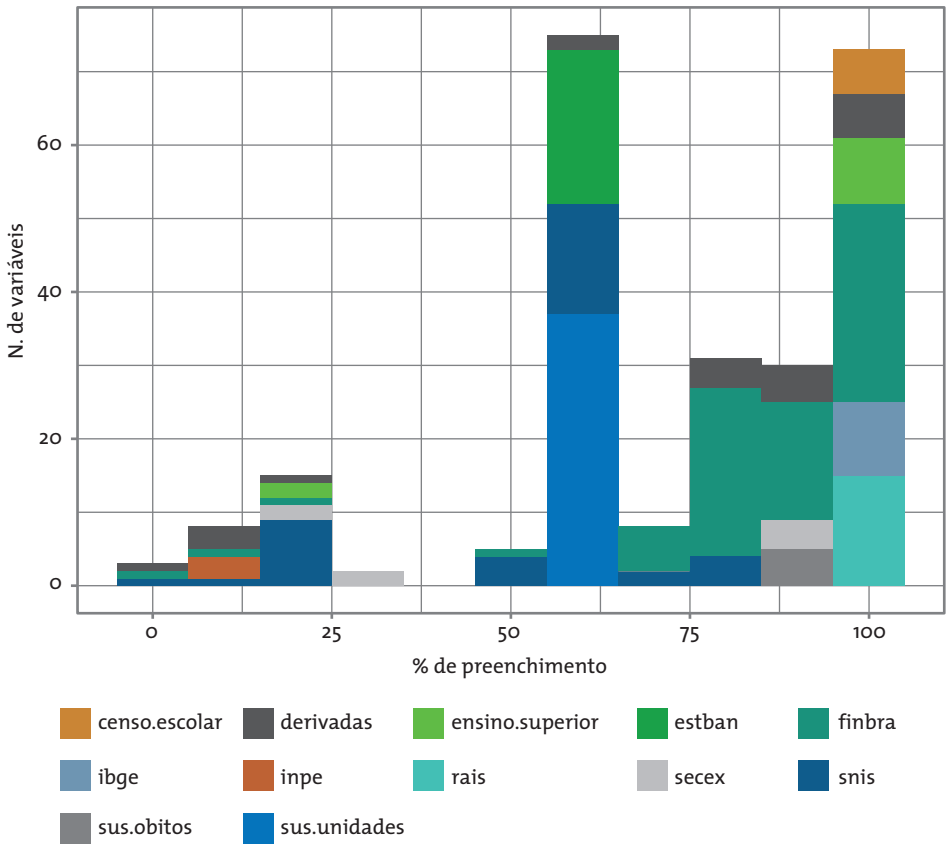


Fonte: Elaboração própria.

Uma avaliação de impacto exige informações completas sobre os indicadores de interesse, assim como de suas covariadas. Essa necessidade é ainda mais importante em estimativas realizadas por controle sintético, as quais são mais

consistentes quanto mais extenso for o período de ajuste. Dessa maneira, variáveis com pouco preenchimento, isto é, com muitas informações incompletas, tendem a ser menos relevantes nos modelos estimados, assim como são descartadas quando for utilizado o método automatizado de seleção de covariadas. Quanto ao preenchimento de variáveis, o Gráfico 3 mostra a distribuição de variáveis por fonte e por percentual de informações disponíveis para cada município brasileiro e ano. O Censo Escolar, a Rais, o Sistema de Informações sobre Mortalidade e o IBGE apresentam as informações mais completas. Já o Inpe parece apresentar as informações menos completas. Todavia, isso se deve ao fato de que suas variáveis são observadas apenas para os municípios da Amazônia Legal. É importante destacar que taxas de preenchimento de 70% ou mais tendem a favorecer as variáveis a serem escolhidas como covariadas dos indicadores de interesse pelo método de seleção automatizado.

Gráfico 3. Histograma de preenchimento (%) das variáveis por fonte

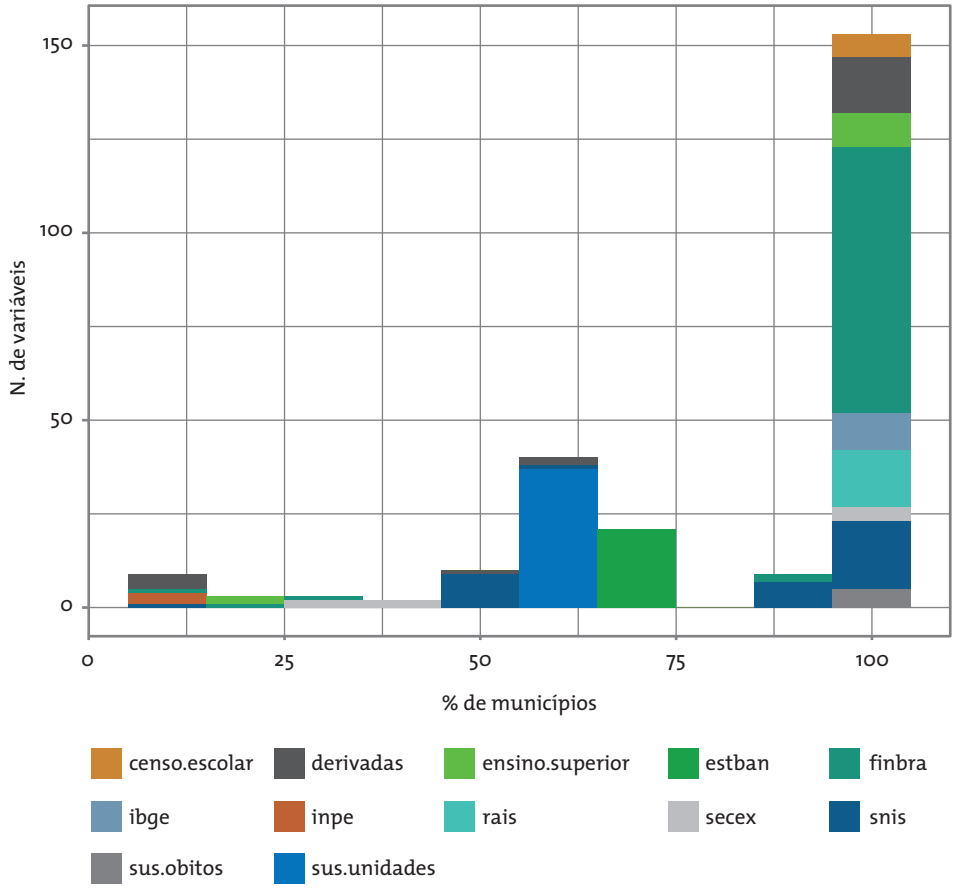


Fonte: Elaboração própria.

Quanto à cobertura por municípios, o percentual de preenchimento foi determinado como a proporção de municípios brasileiros com pelo menos uma informação ao longo do período de cobertura, para cada variável. O histograma está representado no Gráfico 4. Novamente, o Censo Escolar, a Rais, o Sistema de Informações sobre Mortalidade e o IBGE apresentam as informações mais com-

pletas. De forma geral, observa-se que os indicadores médios de preenchimento e de cobertura por município são semelhantes em cada fonte. Isso sugere que os problemas de falta de dados em algumas séries devem-se mais à inexistência de dados para alguns municípios – sobretudo os de menor porte – do que à descontinuidade das séries. A exceção a essa tendência é a base do Snis, que está presente em mais de 80% dos municípios brasileiros, mas com uma taxa de preenchimento total de menos de 50%.

Gráfico 4. Histograma de preenchimento (%) das variáveis nos municípios por fonte



Fonte: Elaboração própria.

3. O Modelo Automatizado em R para Verificação de Impacto (MARVIm) – Módulo de Controle Sintético

3.1 O Método de Avaliação de Impacto por Controle Sintético

A análise de impacto de intervenções públicas em âmbito local, como da construção de usinas eólicas sobre o desempenho econômico dos municípios contemplados, traz uma série de desafios de ordem empírica (ASSUNÇÃO; COSTA; SZERMAN, 2016). Em primeiro lugar, é preciso que alguma técnica seja utilizada para construção de um cenário contrafactual. Ou seja, é necessário que se leve em

consideração a mensuração da variável de interesse em um cenário hipotético no qual esses municípios não tenham sofrido intervenção. Dessa maneira, a medida do impacto será a diferença entre a variável de interesse – no caso, o PIB *per capita* – observada nos municípios beneficiados (tratados) e o PIB *per capita* do contrafactual após o tratamento.

Em segundo lugar, a escolha de construir obras de infraestrutura em localidades específicas pode estar baseada em fatores técnicos, sociais, políticos, econômicos e ambientais. Isto é, como a escolha dos municípios contemplados por essas obras não é aleatória, espera-se que as localidades tratadas tenham características diferentes das não tratadas. Dessa maneira, métodos empíricos baseados em simples comparações de médias entre o grupo de unidades afetadas pela intervenção e o grupo das não afetadas levariam a estimativas viesadas sobre supostos impactos.

Em terceiro lugar, o número de unidades afetadas por esse tipo de intervenção, geralmente, é pequeno. Dessa maneira, os resultados estimados por meio de métodos de análise de impacto baseados em pareamento e em diferenças em diferenças podem ter problemas de inferência estatística.

Nesse sentido, a metodologia de controle sintético foi originalmente concebida para estudos de caso, sendo apropriada para quando se tem poucos tratados, como é o caso em análise. Parte-se do pressuposto que uma combinação de unidades não tratadas compõe melhor contrafactual para a tratada do que qualquer outra individualmente. Para definir as ponderações aplicadas a cada combinação, o método se vale de características mensuráveis de todas as unidades disponíveis para a análise. O trabalho da metodologia de controle sintético é procurar simular, a partir de uma combinação de informações observadas de unidades não tratadas, a mesma trajetória prévia ao tratamento que se observou na unidade beneficiária, conforme foi desenvolvido por Abadie e Gardeazabal (2003) e Abadie, Diamond e Hainmueller (2010; 2015).

Supõe-se uma amostra de $J + 1$ unidades (por exemplo, municípios) indexadas por j. A unidade $j = 1$ é a unidade de interesse, isto é, a unidade tratada por uma intervenção. As demais unidades, de $j = 2$ a $j = J + 1$, constituem o chamado *donor pool*, isto é, o conjunto de unidades não tratadas pela intervenção e que servem como potenciais unidades de comparação com a unidade tratada.

Pressupõe-se que a amostra é um painel balanceado para a variável de interesse, isto é, que é composta por dados longitudinais e que todas as suas unidades são observadas no período de $t = 1,...,T$. A amostra inclui um número positivo de períodos pré-intervenção T_0, assim como de períodos pós-intervenção T_1, de modo que $T = T_0 + T_1$. A intervenção que será avaliada consiste na exposição da unidade $j = 1$ ao tratamento durante os períodos $t = T_0 + 1,...,T$, considerando que essa intervenção não tenha efeitos durante o período pré-tratamento $t = 1,...,T_0$.

Dessa maneira, o objetivo da análise de impacto nessa amostra é medir o efeito da intervenção sobre a unidade tratada em um indicador de interesse para o período pós-tratamento.

Por hipótese, considera-se que as características pré-tratamento da unidade de interesse são mais bem aproximadas por uma combinação das unidades não tratadas do que por qualquer uma dessas unidades não tratadas isoladamente. Dessa maneira, o controle sintético pode ser entendido como uma média ponderada das unidades do *donor pool* que será comparado com a unidade tratada. O controle sintético é representado por um vetor (J X 1) de pesos $W = (w_2,...,w_{j+1})$, tal que $0 \leq w_j \leq 1$ para $j = 2, ... ,J$ e $w_2 + ... + w_{j+1} = 1$ Dessa maneira, a escolha de qualquer valor particular de W é equivalente à escolha de um controle sintético.

Seja X_1 um vetor (K X 1) contendo as características pré-tratamento da unidade tratada, as quais se pretende aproximar o máximo possível. X_0, por sua vez, é uma matriz (K X J) contendo os valores das mesmas variáveis para o *donor pool*. Observa-se que K equivale ao número de variáveis disponíveis para mensurar as características das unidades no período pré-tratamento, sendo preditoras da variável de interesse e não sendo afetadas pela intervenção nesse período. Nesse conjunto de variáveis, pode-se incluir os valores da própria variável de interesse antes do tratamento.

A diferença entre as características da unidade tratada e do controle sintético é dada pelo vetor $X_1 - X_0W$, sendo que o objetivo da metodologia aqui aplicada é escolher o vetor de pesos W^* que minimiza essa distância. Esse valor é obtido da seguinte maneira: para $m = 1,...,K$, seja X_{1m} o valor da variável m para a unidade tratada e X_{0m} um vetor (1 X J) que contém os valores da variável m para as unidades do *donor pool*, deve-se escolher o W^* que minimiza:

$$min_{w \in W} \sum_{m=1}^{K} v_m \left(X_{1m} - X_{0m} W \right)^2$$

Nessa equação, v_m é um peso que reflete a importância relativa atribuída à variável m quando se mede a discrepância entre X_1 e X_0W.

Seja Y_{jt} a variável de interesse da unidade j no tempo t. Y_1 é um vetor (T_1 X 1) dos valores pós-intervenção da variável de interesse para a unidade tratada, de modo que $Y_1 = (Y_{1T_0+1},...,Y_{1T})'$. Y_0 é uma matriz (T_1 X J) em que a coluna j contém os valores pós-intervenção da variável de interesse para a unidade j + 1. Dessa maneira, a variável de interesse do controle sintético é $Y_1^* = Y_0W^*$.

O estimador de controle sintético do impacto do tratamento é dado pela comparação entre os valores da variável de interesse para a unidade tratada e para a unidade de controle sintético no período pós-tratamento:

$$\delta = Y_1 - Y_1^*$$

$$\delta = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

A consistência do estimador de controle sintético será maior quanto maior for o número de períodos pré-tratamento que estiver presente na base de dados (ABADIE; DIAMOND; HAINMUELLER, 2010). Isso ocorre porque esse fator contribui por reduzir o papel de variáveis não observadas na determinação da trajetória pré-tratamento da variável de interesse.

Ao sistematizar o processo de estimação de contrafactuais, o método do controle sintético permite a realização de uma série de exercícios de falsificação, denominados de testes de placebo. Esses exercícios permitem tentativas de testes de inferência estatística, tanto qualitativa como quantitativa. Por exemplo, pode-se aplicar o controle sintético para verificar o efeito da intervenção de interesse sobre as unidades integrantes do *donor pool*, não tratadas, para assim comparar o comportamento da variável de interesse na comparação da trajetória referente à unidade tratada e aos placebos. Se houver algum efeito que possa ser atribuído ao tratamento, espera-se que a unidade tratada tenha sido mais intensamente afetada que os placebos.

A maneira mais comum de mensurar o efeito de um tratamento por meio do método do controle sintético é pela razão do erro quadrático médio pós-tratamento e pré-tratamento, ou Root Mean Squared Prediction Error *ratio* (RMSPE *ratio*). Esse indicador equivale à razão entre os desvios quadrados da trajetória da variável de interesse entre a unidade de referência (tratada ou placebo) e seu correspondente controle sintético para cada período de tempo, depois e antes do ponto de tratamento. Quanto maior for esse valor, maior efeito pode ser associado à intervenção.

$$\text{RMSPE} = \left(\frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_1 t - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{1/2}$$

Um problema relacionado ao RMSPE *ratio* ocorre em casos de ajustes perfeitos no período pré-tratamento. Nesses casos, o denominador da razão tende a zero, e o indicador tende ao infinito independentemente de seu valor após o tratamento, mascarando o verdadeiro efeito da intervenção. Uma alternativa para esse problema é o cálculo da razão das áreas sobre as curvas depois e antes do tratamento (*Area under Curve ratio*, ou AUC *ratio*). Com esse índice, o ajuste perfeito pré-tratamento faz com que o denominador da razão convirja para um e o efeito após o tratamento seja destacado.

Em caso de intervenções com mais de uma unidade tratada, o controle sintético deve ser aplicado para cada caso, individualmente. Após isso, para se obterem medidas do efeito da intervenção como um todo, as estimativas individuais devem ser compiladas (ASSUNÇÃO; COSTA; SZERMAN, 2016). Ou seja, em vez de se considerar uma única unidade tratada $j = 1$, considera-se um conjunto de G unidades tratadas por uma intervenção, as quais são indexadas por $g = 1,2,...,G$. Seja T_{0g} o ano em que houve o tratamento em cada unidade tratada. Para melhor comparar o tratamento em diferentes anos, esses são normalizados em $\tau = t - T_{0g}$, em que $\tau = 0$ é o ano do tratamento de cada unidade em G.

Após a normalização temporal, seja $\hat{\delta}_{j\tau g} = y_{j\tau g} - \hat{y}^{*}_{j\tau g}$ o efeito estimado da intervenção na unidade $g \in G$, pertencente ao conjunto J de unidades tratadas e não tratadas no período τ. Por exemplo, pode referir-se ao impacto da construção da usina eólica no município g, integrante do total J de municípios brasileiros, no ano τ. Dessa maneira, os resultados das unidades tratadas são compilados para cada ponto no tempo, de modo a se obter uma distribuição empírica. Portanto, o impacto médio dos G tratamentos em cada unidade g será:

$$\bar{\delta}_t = \frac{\sum_{g=1}^{G} \delta^{*}_{g1\tau}}{G} = \frac{\sum_{g=1}^{G} \left(y_{g1\tau} - y^{*}_{g1\tau} \right)}{G}$$

Em caso de elevada heterogeneidade de efeitos entre as unidades tratadas, pode ser mais vantajoso utilizar a mediana em vez da média para compilar os resultados individuais, assim como os percentis 0,25 e 0,75 para cada caso:

$$Px\left(\delta_{\tau}\right) = Px\left(\delta^{*}_{g1\tau}\right)$$

Nessa fórmula, Px equivale ao percentil escolhido do efeito para cada τ de tratamento.

Observa-se que, nos casos em que a intervenção não foi aleatoriamente atribuída entre as unidades, dois problemas podem ameaçar as conclusões quanto à inferência a respeito dos efeitos individuais compilados. Em primeiro lugar, pode haver viés, isto é, a trajetória da variável de interesse de uma unidade tratada pode estar separada de seu controle sintético desde antes da intervenção. Em segundo lugar, pode haver heterocedasticidade de choques idiossincráticos. Esse problema acontece quando uma unidade tratada recebe choques aleatórios sobre sua variável de interesse com uma variância diferente de seu controle sintético. Se os choques atingirem a unidade tratada com maior variância do que o controle sintético, poderá haver superestimação do efeito do tratamento. Se os choques atingirem o controle sintético com maior variância do que a unidade tratada, pode haver subestimação do tratamento.

Os dois problemas – viés e heterocedasticidade de choques – são detectados no exercício empírico por meio de um mau ajuste pré-tratamento da variável de interesse, e isso pode ser testado sob a forma de um elevado RMSPE pré-tratamento. Por isso, uma forma simples de controlar a influência desses casos na estimação do efeito compilado das intervenções é a eliminação das unidades tratadas com um nível de RMSPE pré-tratamento acima de um patamar escolhido.

3.2 Passo a passo de uma avaliação do MARVIm – Módulo de Controle Sintético

O Departamento de Avaliação e Promoção da Efetividade (DEAPE) da Área de Planejamento (AP) do BNDES desenvolveu uma metodologia, denominada Modelo Automatizado em R para Verificação de Impacto (MARVIm) – Módulo de Controle Sintético, que combina uma série de *scripts* e funções em R que automatizam avaliações baseadas em controle sintético. Esse método de avaliação foi orginalmente desenvolvido para estudos de caso, isto é, com o objetivo de detectar efeitos de uma política ou de um evento sobre algum indicador de interesse de uma unidade exposta à intervenção.[3] Ou seja, é um método adequado para os casos em que o problema de micronumerosidade torna inconsistente a avaliação de impacto quando realizada pelos métodos tradicionais de pareamento e de diferenças em diferenças. Portanto, o objetivo do segundo módulo do MARVIm é servir como um modelo automatizado para avaliações de impacto em intervenções com poucas unidades tratadas, sejam elas firmas ou unidades geográficas. Seu maior desafio é precisamente sistematizar e automatizar esse método, de modo a permitir que seja aplicável a um grande conjunto de intervenções em unidades cujas características estão mensuradas em diferentes bases de dados.

Em linhas gerais, o MARVIm Módulo de Controle Sintético recebe do avaliador os *inputs* básicos para a análise de impacto, o que inclui uma base de dados contendo observações de municípios ou firmas que serão considerados na avaliação, assim como os anos de tratamento e as variáveis de interesse. A partir desses insumos, o modelo percorre os seguintes passos: (i) seleciona, em um exercício de *machine learning*, entre todas as variáveis da base de dados, aquelas que melhor preveem a variável de interesse; (ii) de posse das selecionadas, aloca as unidades a *clusters*, constituídos pela maior similaridade (em variáveis observadas) de unidades *intra-clusters* e maior diferença entre *clusters*; (iii) por fim, com a base de dados reduzida apenas ao *cluster* que contém a unidade tratada, constrói o controle sintético e compara-o à unidade tratada. O avaliador utiliza-se dessa comparação para formar um parecer sobre a capacidade ou não de atribuir efeitos ao tratamento e, em caso afirmativo, qual a magnitude desse efeito. Para isso, o MARVIm é ajustado por uma série de parâmetros que são escolhidos pelo pesquisador, os quais são:

[3] Ver, por exemplo, Abadie, Diamond e Hainmueller (2010; 2011; 2015).

a) Parâmetros básicos de avaliação

Carrega uma base de dados de análise e identifica as variáveis de identificação de indivíduo e de tempo. Define as unidades tratadas e as variáveis de interesse para a avaliação. Define também as unidades que serão removidas da análise a critério do pesquisador (por exemplo, unidades que foram tratadas em períodos anteriores à cobertura da base de dados).

b) Filtros para o *donor pool*

Para reduzir o número de potenciais controles, o pesquisador pode ligar e desligar filtros que identificam os controles com alguma característica da unidade tratada. Por exemplo, se a unidade tratada for um município, o pesquisador pode restringir os potenciais controles apenas aos municípios localizados no mesmo estado.

c) AdaLASSO

Esse procedimento busca reduzir a dimensionalidade da base de dados, de forma a tornar a estimação computacionalmente mais rápida e estável. Por meio de técnicas de regressão, procura identificar as covariadas que melhor expliquem cada variável de interesse e excluir as variáveis menos importantes da base de análise. Além de escolher se aplica esse procedimento, o pesquisador pode escolher o número máximo de covariadas associadas a cada variável de interesse.

d) Análise de *outlier*

Permite que o MARVIm identifique e exclua da lista de potenciais doadores unidades cujo comportamento das variáveis de interesse seja considerado não comparável com as unidades tratadas.

e) Teste de placebo

Para verificar a robustez das conclusões tomadas a partir da avaliação por controle sintético, o pesquisador pode realizar um número definido de testes de placebo, isto é, repetições do exercício de avaliação para unidades não tratadas, de modo a comparar os seus resultados com aqueles obtidos pela avaliação da unidade tratada.

O primeiro passo de uma avaliação de impacto realizada pelo MARVIm – Módulo de Controle Sintético é o processo de carregar e validar a base de dados. Nesse passo, ocorrem a identificação das unidades tratadas e a disponibilização de informações sobre todas as unidades individuais presentes, assim como a identificação da variável de tempo. Nesse passo, também ocorre o teste dos períodos de disponibilidade de informações para cada um dos indivíduos tratados e suas variáveis de interesse, de modo a avaliar a possibilidade da estimação de um controle sintético. Para a avaliação individual ser validada, o MARVIm procura detectar

uma série de condições relativas aos dados de cada unidade tratada presentes na base de análise. As condições incluem: a presença de alguma informação sobre a variável de interesse informada; informações com alguma variabilidade (isto é, o seu desvio-padrão deve ser não nulo); e a disponibilidade das informações por, no mínimo, cinco anos antes e um ano após o período de tratamento informado.

Geralmente, para que exercícios de avaliação de impacto possam obter estimativas confiáveis, é necessário que utilizem bases de dados extensas, tanto em relação a unidades observadas como em relação a variáveis. Os modelos de alta dimensionalidade estão cada vez mais presentes na literatura, já que a inclusão de um grande número de variáveis pode contribuir para ganhos de capacidade preditiva dos modelos (KONZEN, 2014). Porém, quando a dimensionalidade do modelo é grande em relação ao tamanho da amostra, os métodos tradicionais de regressão podem apresentar problemas. Primeiro, porque aumenta a dificuldade de tornar os modelos interpretáveis. Segundo, porque os modelos perdem robustez. Terceiro, porque há comprometimento da eficiência computacional. Quarto, em função da perda de eficiência quanto à inferência estatística. Quinto, porque há problemas com correlação espúria entre as covariadas do modelo, a qual pode ser elevada mesmo quando elas forem independentes e identicamente distribuídas.

No caso de modelos de controle sintético, esses problemas tendem a ser mais graves, já que o método tem sua eficiência computacional muito sensível à extensão das bases de dados. Não obstante, nesses modelos a estimação do contrafactual é dependente do conjunto de covariadas presente na base de dados de análise. Por isso, há a necessidade de filtrar as variáveis mais importantes para explicar a trajetória da variável de interesse sobre a qual será calculado o efeito da intervenção realizada.

Uma solução para problemas referentes à alta dimensionalidade dos modelos é a suposição de esparsidade do vetor de parâmetros. Isto é, a suposição de que muitos de seus componentes são iguais a zero. Essa hipótese pode produzir estimativas viesadas, mas contribui com a identificação das covariadas mais importantes, assim como com a obtenção de um modelo mais parcimonioso. Da mesma maneira, reduz a complexidade do modelo e seu custo em termos computacionais.

Por essas razões, o segundo passo de uma avaliação pelo MARVIm é a seleção de covariadas relevantes para cada indicador de interesse escolhido. Para isso, é utilizada uma metodologia automatizada de seleção de variáveis, de modo a reduzir a massa de dados, o que, além de propiciar ganhos computacionais, dá maior previsibilidade ao modelo e maior facilidade para a interpretação dos resultados obtidos. Essa metodologia é chamada de adaLASSO, ou LASSO Adaptativo. O nome LASSO deriva de Least Absolute Shrinkage and Selection Operator, ou operador de menor contração e seleção absolutos. O método consiste em uma regressão linear em que o indicador de interesse é escolhido como variável dependente em

função de todas as demais variáveis da base de dados. Essa regressão, no entanto, conta como uma função de punição, que força a soma dos valores absolutos dos coeficientes estimados a ser menor do que um determinado valor.[4] Particularmente no caso do adaLASSO, a função de punição apresenta pesos adaptativos para punir diferentes coeficientes estimados pelo modelo. Isso assegura que o método de seleção de covariadas apresente consistência na seleção de variáveis e a normalidade assintótica.

O LASSO é um método de encolhimento do conjunto de coeficientes estimados de um modelo desenvolvido por Tibshirani (1996). Esse método consiste, simplificadamente, na introdução de uma punição ao conjunto de norma L_1 dos coeficientes, isto é, uma punição na soma dos valores absolutos dos coeficientes. Seu objetivo é permitir a estimação de um modelo que produza previsões com pequena variância e que determine um conjunto de preditores que melhor explicam a variável de interesse. A punição introduzida tende a zerar alguns dos coeficientes estimados, o que não apenas reduz a dimensionalidade do espaço paramétrico, mas também seleciona as covariadas mais relevantes.

O LASSO pode ser entendido como uma técnica de regularização. Isto é, considera-se uma função de erro do tipo E = *medida de erro* + λ* *complexidade do modelo*. No caso de uma regressão, a medida de erro equivale à soma dos quadrados dos resíduos estimados. O segundo termo representa a punição dos modelos com maior complexidade e variância dos estimadores, sendo que λ representa a severidade dessa penalidade. Quanto maior for λ, mais simplificado será o modelo estimado, ainda que isso leve a um maior viés. Em termos formais, as estimativas LASSO são obtidas por meio da minimização dos quadrados dos resíduos sujeita a uma punição de norma L_1 dos coeficientes:

$$\hat{\beta}^{LASSO} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji} \right)^2 \; sujeito\, a \sum_{j=1}^{k}\left|\beta_j\right| \leq t$$

Nessa equação, o parâmetro de ajuste $t \geq 0$ controla a punição aplicada ao conjunto de coeficientes. Para todo t, tem-se que $\hat{\beta}_0 = \bar{y}$. Assumindo que $\bar{y} = 0$, pode-se omitir β_0. Sendo $\left\{\hat{\beta}_j^0\right\}_{1<j<k}$ o conjunto de coeficientes estimados por mínimos quadrados ordinários (MQO) e $t_0 = \Sigma\left|\hat{\beta}_j^0\right|$, observa-se que valores $t < t_0$ levarão a um encolhimento dos coeficientes em direção a zero, ao passo que valores $t \geq t_0$ aproximarão os coeficientes das estimativas LASSO das estimativas de MQO. A equação pode ser desenvolvida utilizando-se o Lagrangiano:

$$\hat{\beta}^{LASSO} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k}\left|\beta_j\right| \right\}$$

[4] Mais detalhes em Zou (2006).

Aqui, $\lambda \geq 0$ é uma função do parâmetro de ajuste *t*. Quanto maior o valor de λ, maior será a penalidade imposta ao somatório dos coeficientes. Por outro lado, se $\lambda = 0$, as estimativas LASSO serão iguais às estimativas de MQO. Esse parâmetro é escolhido por um procedimento de *K-fold cross-validation*, no qual a amostra é particionada aleatoriamente em *K* subamostras de tamanhos iguais, e o modelo é iterativamente estimado eliminando-se uma subamostra. Assim, fazem-se previsões com base em cada estimação e compara-se com a subamostra removida, calculando-se o erro quadrático médio de previsão (EQMP) para avaliar a qualidade do ajuste naquela subamostra. Portanto, o λ ótimo deve minimizar o EQMP médio nas *K* subamostras.

O LASSO apresenta menor variabilidade entre outras opções de modelos de redução de dimensionalidade (KONZEN, 2014). Além disso, por encolher alguns coeficientes para zero, destaca as covariadas mais relevantes para explicar uma variável de interesse. Por fim, é capaz de realizar a escolha das variáveis e a estimação dos coeficientes simultaneamente. Contudo, deve-se observar que nem sempre o LASSO é consistente na escolha de variáveis (ZHAO; YU, 2006). Isso significa que a solução esparsa – isto é, de dimensionalidade reduzida, com alguns coeficientes reduzidos a zero – pode não representar o modelo verdadeiro quando o tamanho da amostra tende ao infinito.

Nesse sentido, o LASSO Adaptativo, ou adaLASSO, é um método que pretende dar consistência às estimativas LASSO por meio da atribuição de diferentes pesos para diferentes coeficientes (ZOU, 2006). Ou seja:

$$\hat{\beta}^{adaLASSO} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} u_j \left| \beta_j \right| \right\}$$

$$u_j = \left| \hat{\beta}_j^{ridge} \right|^{-\tau}, \tau > 0$$

Ou seja, u_j é um vetor de pesos individuais para auxiliar a seleção das variáveis relevantes para o modelo. Esse vetor é função de $\hat{\beta}_j^{ridge}$, isto é, dos coeficientes estimados por meio de uma regressão que controla possíveis efeitos de multicolinearidade (regressão *ridge*). Uma variável relevante x_j terá um valor $\hat{\beta}_j^{ridge}$ comparativamente elevado, o que reduz o peso u_j atribuído a seu coeficiente. Por outro lado, uma variável irrelevante x_j terá um valor $\hat{\beta}_j^{ridge}$ comparativamente baixo, o que aumenta o peso u_j atribuído a seu coeficiente. Assim, o modelo atribuirá maior punição dos coeficientes das variáveis que aparentam ser pouco importantes para explicar a variável de interesse.

Sob um conjunto de pesos u_j apropriados, o adaLASSO tem propriedades Oracle, isto é, é consistente na seleção de variáveis e apresenta normalidade assintótica. Ou seja, o método seleciona corretamente as variáveis relevantes quando o modelo aumenta. Além disso, as estimativas dos coeficientes não nulos seguem

assintoticamente a mesma distribuição dos estimadores de MQO quando este for estimado apenas com as variáveis relevantes.

Em resumo, o método de seleção automatizada de variáveis apresenta três propriedades importantes, que colaboram com a avaliação de impacto pelo MARVIm de modo geral (FAN; LI, 2001). Primeiro, o não viés, de modo que os estimadores de parâmetros de valor elevado têm maior consistência que os demais. Segundo, a esparsidade, de acordo com a qual a complexidade do modelo é reduzida com a determinação dos menores coeficientes em zero. Terceiro, a continuidade, em que o estimador é contínuo nos dados para evitar a instabilidade na capacidade preditiva do modelo.

Para um melhor ajuste das estimativas do adaLASSO, alguns procedimentos são sistematicamente realizados. Primeiro, são excluídas as unidades sem informação para a variável de interesse. Após isso, são descartadas as variáveis com menos de 70% de preenchimento ou com percentual de preenchimento inferior ao da variável de interesse (o que for menor). Além disso, as variáveis derivadas não são regredidas em função das variáveis brutas usadas em sua construção. Por exemplo, PIB *per capita* não é regredido sobre PIB e população. Por fim, para que as estimativas do adaLASSO não sejam direcionadas em favor das variáveis com menor magnitude, elas são normalizadas durante esse procedimento. O MARVIm retorna a lista das variáveis mais relevantes para explicar a variável de interesse selecionada.

Para a execução do controle sintético, é necessária a definição dos potenciais controles, entre todas as unidades não tratadas presentes na base de dados. Essa definição tem o objetivo de priorizar as unidades semelhantes à avaliada, de modo a reduzir a massa de dados, assim como de descartar possíveis casos *outliers*. Para realizar essa seleção, as variáveis selecionadas na primeira etapa são utilizadas para classificar os indivíduos em diferentes *clusters*. Para isso, foi utilizado um algoritmo de clusterização paramétrico multidimensional. Os parâmetros são estimados por meio de um algoritmo EM (*expectation-maximization*) iniciado por uma clusterização baseada em um modelo hierárquico.[5] As características de cada *cluster* são determinadas pela matriz de covariâncias. Apenas os indivíduos classificados no mesmo *cluster* do indivíduo tratado continuam na base de dados de análise após essa etapa.

Um problema identificado na literatura são indivíduos presentes como potenciais controles que sofrem algum choque exógeno no período pós-tratamento e, portanto, podem viesar o controle sintético. A fim de identificar esses casos, é efetuada uma análise de *outliers* no grupo selecionado para *donor pool* na clusterização inicial. Essa análise é feita em duas etapas.

A primeira etapa consiste em identificar indivíduos com trajetória semelhante no período pré-tratamento. Para tal, aplica-se o procedimento em três passos proposto

[5] Mais detalhes sobre o algoritmo podem ser encontrados em Sundberg (1974; 1976) e Dempster, Laird e Rubin (1977).

por Leffondree *et al.* (2004) para identificação de padrões longitudinais. Assim, são calculadas 24 estatísticas descritivas das características das trajetórias. Após isso, é utilizada uma análise de fatores para selecionar as estatísticas com maior poder descritivo. Por fim, utilizam-se as estatísticas selecionadas como dimensões para clusterização dos indivíduos.

Na segunda etapa, as trajetórias dos indivíduos no período pós-tratamento, para a variável de interesse, são comparadas em cada um dos *clusters*. Para tal, é calculada uma distância em relação à curva média para cada um dos indivíduos. Utilizando essa medida de distância, o critério MAD (*median absolute deviation*) é utilizado para classificar os *outliers*. Os indivíduos identificados como *outliers* são retirados da base de indivíduos controles potenciais (filtrada anteriormente na etapa de clusterização).

Esse procedimento em duas etapas foi adotado para garantir a comparabilidade das trajetórias da variável de interesse em cada um dos indivíduos no período pós- e pré-tratamento. Apenas os indivíduos que apresentem uma trajetória semelhante à de outro grupo de indivíduos no período pré-tratamento e uma trajetória bem distinta no período pós-tratamento são considerados *outliers*.

Definidos os potenciais controles e as variáveis correlacionadas à variável de interesse a ser utilizada no teste de impacto, temos as informações necessárias para a realização da comparação via controle sintético, para cada caso solicitado pelo avaliador. Aqui, há a devida execução dos procedimentos de otimização e estimação para cada uma das variáveis de interesses listadas em cada uma das unidades tratadas. Em caso de erro, este é armazenado pelo MARVIm, e o modelo passa automaticamente para a próxima avaliação – isto é, o próximo par "unidade tratada" e "variável de interesse" – solicitada.

Fundamentalmente, o método do controle sintético é um procedimento baseado em dados para a construção e a escolha de unidades de controle para fins de comparação com unidades expostas a um tratamento. O método parte do pressuposto de que uma combinação de unidades de controle pode ser uma melhor base de comparação para a unidade tratada do que qualquer outra unidade isolada. Assim, o MARVIm procura construir uma unidade de controle artificial a partir de dados das unidades não tratadas reais. Essa unidade construída, denominada de unidade de controle sintético, é uma média ponderada de todas as unidades presentes na base de dados de análise que não foram tratadas. O conjunto de pesos que ponderam cada unidade não tratada no controle sintético é calculado por meio de um algoritmo que, ao mesmo tempo, minimiza a diferença prévia da trajetória da variável de interesse da unidade tratada e os potenciais controles antes do tratamento, e outro conjunto de pesos que ponderam a importância de cada uma de uma série de outras variáveis que explicam a trajetória da própria variável de interesse. Dessa maneira, o método é capaz de explicitar três importantes propriedades de cada avaliação: (i) os pesos de cada covariada na cons-

trução do controle sintético; (ii) a contribuição de cada potencial unidade de controle para o contrafactual construído; (iii) as similaridades entre a unidade exposta à intervenção e seu contrafactual, isto é, o desempenho do indicador de interesse de ambos antes e depois do tratamento, tendo demais características levadas em consideração.

O objetivo desse procedimento é encontrar a trajetória aproximada do indicador de interesse que a unidade tratada provavelmente teria seguido caso não tivesse sofrido a intervenção. Para isso, calcula-se a média ponderada das unidades não tratadas cujos pesos minimizam a distância do comportamento da unidade tratada no período anterior ao tratamento. Em seguida, a trajetória do controle sintético é projetada no período pós-tratamento e comparada com a trajetória da unidade tratada. A diferença entre essas duas trajetórias é entendida como o efeito do tratamento.

Para saber se a diferença entre as trajetórias foi realmente influenciada pela intervenção sofrida pela unidade tratada, é realizado um conjunto de testes de placebo com a mesma base de dados. A ideia fundamental desses testes é a inferência de permutação, isto é, parte da noção de que a distribuição de um teste estatístico é computada a partir de permutações aleatórias entre a categorização das unidades nos grupos de tratamento e de controle.

Em uma avaliação realizada pelo método do controle sintético, o teste de placebo consiste na realização do mesmo exercício para um número escolhido de unidades não tratadas presentes na base de dados, considerando-se o mesmo indicador de interesse e o mesmo ano de suposto tratamento. Isso permite verificar se o efeito calculado da intervenção, isto é, a diferença entre a trajetória do tratado e de seu controle sintético, é relevante para a unidade tratada em relação ao efeito calculado sobre uma unidade aleatória. Em outras palavras, o exercício examina se o efeito estimado da intervenção real é elevado em relação à distribuição dos efeitos estimados para unidades não expostas à intervenção. Se a intervenção não tiver efeito, seu efeito estimado não deve ser destoante em relação à distribuição dos efeitos placebo. Para evidenciar que o tratamento tenha tido efetividade, é esperado que ambos os indicadores sejam mais elevados para as unidades tratadas do que para os testes de placebo. Essa probabilidade é computada como um pseudo *p-valor*.

Após a realização de todos os possíveis placebos, ocorre o cálculo de duas estatísticas de ajuste da estimação para cada caso. A primeira estatística é a razão entre a raiz do erro quadrático médio entre cada unidade e seu controle sintético em cada ano, depois e antes do tratamento. A segunda é a razão entre a área sob a curva da diferença entre o tratado e o controle sintético depois e antes do tratamento. Observa-se que ambos os indicadores são razões, de modo que eles punem os testes de placebo com mau ajustamento no período pré-tratamento. A razão entre a raiz do erro quadrático médio depois e antes do tratamento (RMSPE *ratio*) é o procedimento mais comum para verificar o descolamento da trajetória da variável de interesse no período após o tratamento. Contudo, é sensível a ajustes perfeitos

no período pré-tratamento, de modo que seu denominador, nesses casos, tende a zero, e seu valor final tende ao infinito mesmo que a trajetória pós-tratamento da unidade tratada não se separe do controle sintético. Já a razão entre as áreas sob a curva depois e antes do tratamento é menos sensível a valores relacionados ao ajuste perfeito, que nesse caso tendem a 1, não a 0. Contudo, essa razão apresenta menor variabilidade de valores do que o indicador anterior.

Em casos de avaliações com mais de uma unidade tratada, o MARVIm aplica o método de controle sintético para cada caso individual. Depois disso, as estimativas são compiladas, de modo que os anos de tratamento de cada caso são normalizados em um índice de referência t.[6] Assim, $t - n$ corresponderá ao número de anos pré-tratamento, ao passo que $t + n$ corresponderá ao número de anos pós-tratamento. Em seguida, os resultados são calculados para cada ano de referência segundo a escala normalizada. Os resultados correspondem às estatísticas sobre os efeitos das intervenções individuais para cada ano de referência, como a média ou a mediana; nesse caso, relativamente mais útil nas avaliações em que há demasiada heterogeneidade de resultados individuais.

Deve-se levar em consideração que, antes dos procedimentos de normalização e compilação, é necessário observar os casos em que houve problemas na estimação do controle sintético individual. Dois padrões de casos são especialmente relevantes, quais sejam: os casos em que não houve ajuste pré-tratamento, isto é, a unidade tratada não era comparável com seu controle sintético; e aqueles em que a unidade tratada recebeu forte choque antes do tratamento sobre sua variável de interesse, e não foi acompanhada por seu controle sintético. Desse modo, a unidade já chegou no período de tratamento separada de seu controle, de modo que qualquer efeito calculado pós-tratamento não pode ser necessariamente atribuído a ele. Ambos os casos serão facilmente detectados pelo MARVIm, já que terão computados elevados índices de RMSPE no período pré-tratamento. As unidades assim afetadas podem ser eliminadas caso seja escolhido um nível máximo de RMSPE pré-tratamento permitido para a amostra de unidades tratadas. Esse procedimento é recomendável para que as estimativas compiladas sejam consistentes, isto é, que o resultado final da avaliação não seja afetado por unidades de comportamento anômalo.

4. Estudo de caso: análise de impacto da construção de usinas eólicas nos municípios beneficiados

4.1 Panorama do setor de energia eólica no Brasil

Uma vez apresentada a metodologia de avaliação de impacto por controle sintético, assim como a iniciativa por parte do DEAPE de automatizá-la por meio do

[6] Essas técnicas de compilação de resultados individuais baseadas em controle sintético foram inspiradas em Assunção, Costa e Szerman (2016).

MARVIm – Módulo de Controle Sintético, o presente trabalho apresentará uma aplicação. Esse estudo de caso é a análise do impacto da construção de usinas eólicas sobre as economias dos municípios beneficiados. A presente seção inclui uma breve descrição sobre o panorama do setor de energia eólica no Brasil, a partir dos dados da Aneel, uma revisão bibliográfica de estudos anteriores sobre o tema, uma avaliação pelo MARVIm para um caso individual e a avaliação compilada para todas as unidades tratadas.

Conceitualmente, pode-se entender um parque eólico ou usina eólica como um conjunto de moinhos de vento ou turbinas que são usados para gerar energia elétrica por meio de seus aerogeradores, os quais são empurrados pelo vento. As turbinas são destinadas a transformar energia cinética do vento em elétrica. A principal vantagem da energia eólica em relação a outras fontes é que se trata de uma fonte de energia renovável e limpa, pois não emite os gases de efeito estufa que contribuem para o aquecimento global. Além disso, não produz resíduos ao gerar eletricidade. Também se deve destacar que a fonte é considerada inesgotável e não há custos associados à obtenção de uma matéria-prima combustível, diferentemente do que ocorre com combustíveis fósseis, assim como há baixos riscos ambientais, que são mais comuns em usinas hidrelétrica e nucleares, por exemplo (COSTA; CASOTTI; AZEVEDO, 2009).

Atualmente, a maior fonte energética do Brasil é de matriz hidrelétrica. Apesar de sua produção não poluente e barata, essa dependência pode acarretar danos. Isso porque, diante de períodos de seca, os reservatórios de água podem esvaziar e surgir a necessidade de colocar em funcionamento usinas termelétricas, caras e poluentes. Para que o gargalo de energia seja solucionado, investimentos em energia alternativa tornam-se de suma importância. O Brasil apresenta um dos maiores volumes de ventos do mundo, assim como baixa probabilidade de ocorrência de fenômenos climáticos extremos. Dessa forma, o país tem possibilidades concretas de ampliar seu uso de energia eólica. Segundo o *Atlas do potencial eólico* (AMARANTE; ZACK; SÁ, 2001), o território nacional apresenta ventos com potencial que proporcionariam o equivalente a 272 terawatt-hora por ano (TWh), o que representa aproximadamente 64% do consumo nacional de energia elétrica, que gira em torno de 424 TWh. Atualmente, considera-se que o potencial eólico brasileiro seja superior ao estimado pelo atlas (TOLMASQUIM, 2016).

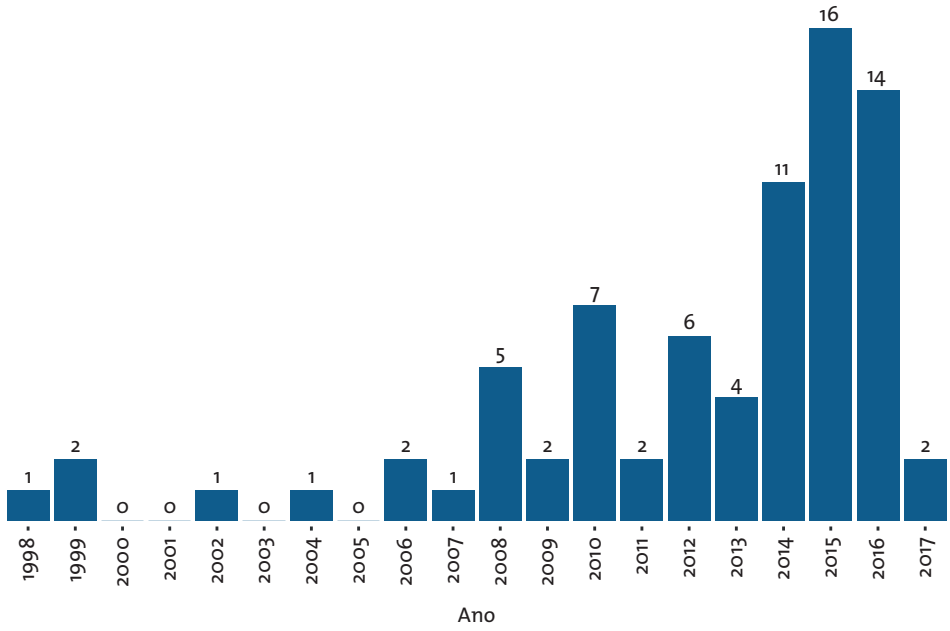
O Brasil, nos últimos anos, vem acelerando a implantação de parques eólicos. Até 2014, a quantidade de energia eólica produzida era de 4 terawatt-hora por ano. O país iniciou o mês de fevereiro de 2018 com capacidade instalada de 12,95 gigawatts (GW), equivalente a 8,3% do total na matriz elétrica nacional (Aneel e Abeeólica).[7] Nesse mês, havia ainda 4,8 GW de capacidade em construção.

[7] Ver *sites* da Abeeólica – Associação Brasileira de Energia Eólica – www.portalabeeolica.org.br/ e Aneel – Agência Nacional de Energia Elétrica – www.aneel.gov.br.

Os incentivos no setor de energia eólica no Brasil iniciaram-se no ano de 2002, a partir do Programa de Incentivo às Fontes Alternativas de Energia Elétrica (Proinfa). Os investimentos ocorrem por meio de leilões especificamente para a geração de eólicas, e também para outras fontes de energias alternativas (LAGE; PROCESSI, 2013). O BNDES apoia o setor de energia eólica com linhas especiais de financiamento para empresas do segmento de geração, bem como para as cadeias produtivas de máquinas e equipamentos. Para os proprietários dos parques eólicos, o Banco disponibiliza dois produtos: o Finem, que visa apoiar investimentos em aumento da capacidade e construção de novas plantas; e o Finame, que visa financiar a venda de máquinas e equipamentos já negociados com as respectivas compradoras (COSTA; CASOTTI; AZEVEDO, 2009).

O Gráfico 5 apresenta o número de municípios brasileiros que receberam suas primeiras usinas eólicas por ano. Vale ressaltar, de acordo com esse gráfico, que o período de maior inauguração de eólicas por município foi entre os anos de 2014 e 2017. Dessa maneira, a maior parte dos parques eólicos não pôde ser avaliada neste trabalho por falta de dados nas variáveis de comparação referentes aos anos subsequentes a 2015 no banco de dados do IBGE.

Gráfico 5. Número de municípios com a primeira usina construída (por ano)



Fonte: Elaboração própria, com base em dados da Aneel (www.aneel.gov.br).

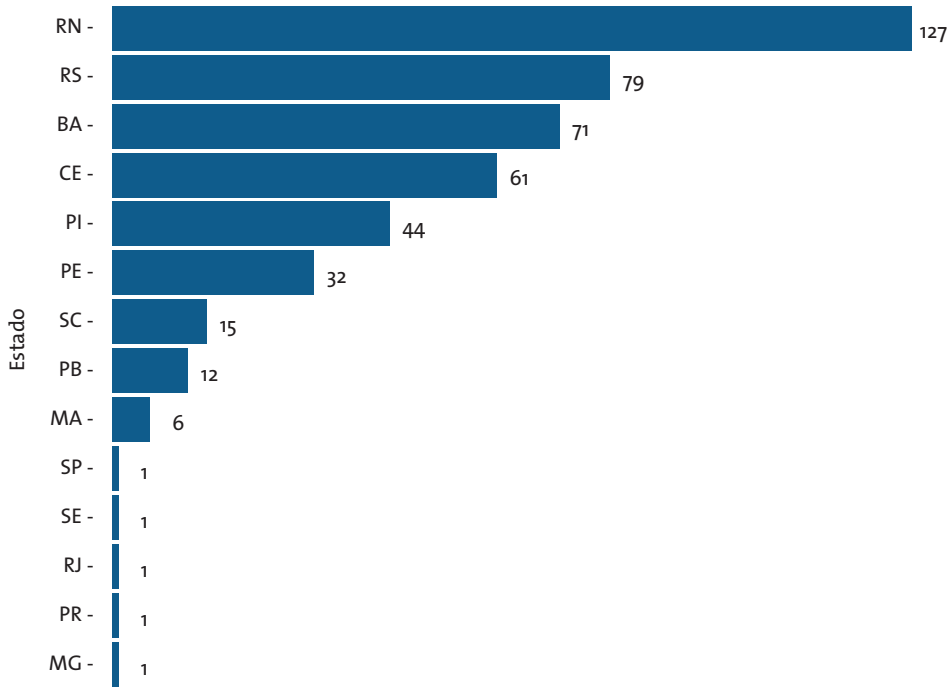
O Mapa 1, a seguir, indica como as usinas eólicas estão distribuídas no Brasil, agregando os parques eólicos por período de construção. Os períodos foram divididos em três grupos, isto é, antes (1998-2006), durante (2007-2014) e depois (2015-2017) do período de análise do presente trabalho, definido pela disponibilidade de dados. O mapa evidencia uma concentração de usinas no período mais recente e em municípios localizados nas regiões Nordeste e Sul:

Mapa 1. Municípios com parques eólicos pelo ano da primeira usina construída

Fonte: Elaboração própria, com base em dados da Aneel (www.aneel.gov.br).

O Gráfico 6 representa a distribuição de usinas eólicas nos estados brasileiros, em todo o período coberto pelos dados da Aneel (1998-2017). Atualmente, 14 estados produzem energia elétrica em usinas eólicas. A região Nordeste apresenta a maioria das usinas eólicas no país, com 78% do total, distribuídas em oito estados. O estado que apresenta a maior quantidade de usinas é o Rio Grande do Norte, concentrando 28% do total nacional. A região Sul vem atrás, com 21% das usinas distribuídas em três estados, sendo que só o Rio Grande do Sul representa 17% de eólicas no Brasil. A região Sudeste conta com menos de 1% da fonte energética eólica, e as usinas estão presentes em três estados.

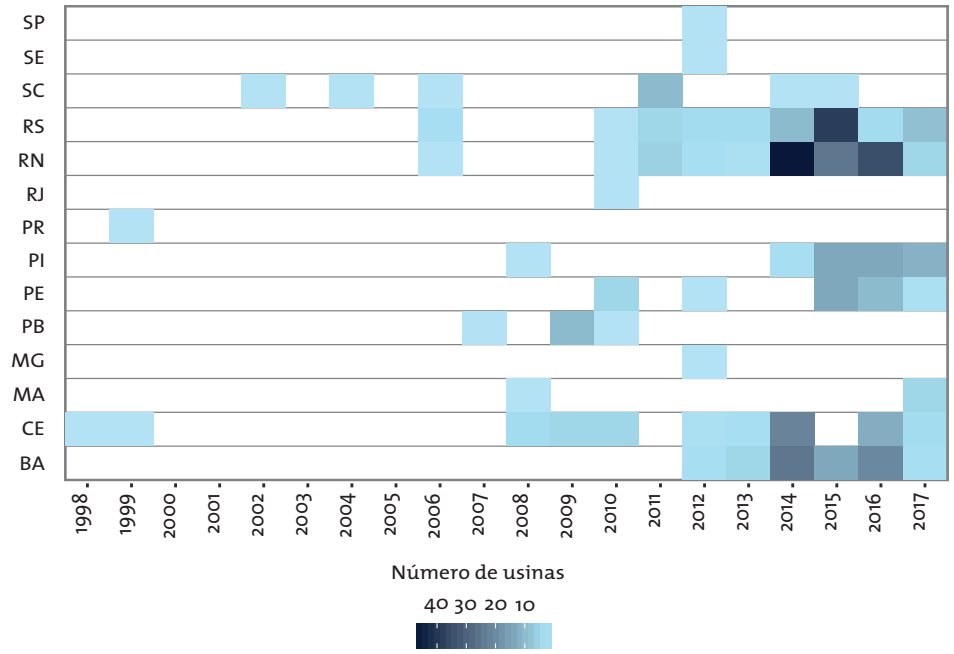
Gráfico 6. Número total de usinas por estado (1998-2017)



Fonte: Elaboração própria, com base em dados da Aneel (www.aneel.gov.br).

O Gráfico 7 representa a distribuição de usinas eólicas nos estados brasileiros e o ano de sua construção. Mais uma vez, é evidenciada a concentração das usinas no período a partir de 2014 e nos estados da região Nordeste, sobretudo no Rio Grande do Norte. A concentração das novas usinas nessa região está se acentuando desde 2013.

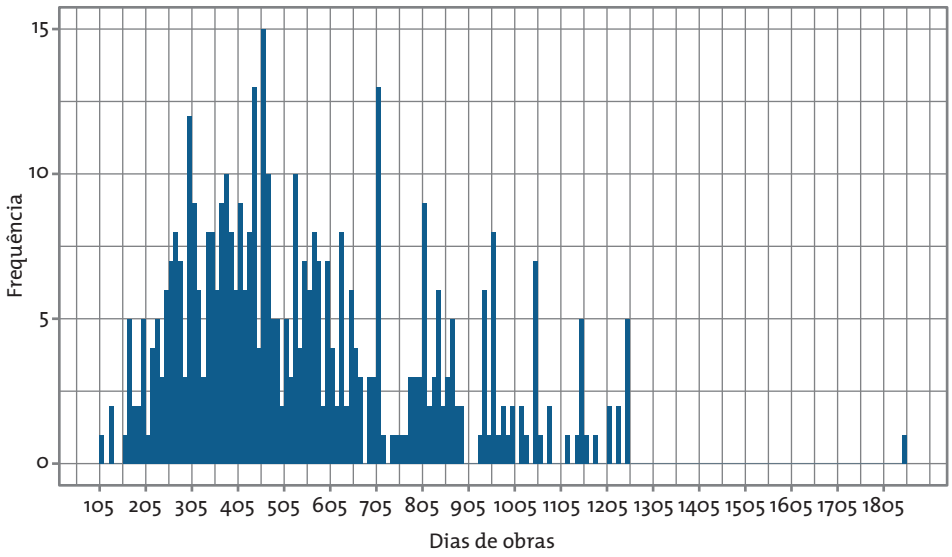
Gráfico 7. Número de usinas por estado e ano de construção



Fonte: Elaboração própria, com base em dados da Aneel (www.aneel.gov.br).

O tempo de construção de uma usina eólica foi definido como o número de dias entre o início das obras civis de suas estruturas e o início de sua operação comercial. No período entre 2005 e 2017, o tempo de construção foi, em média, de 557 dias, ou cerca de 1 ano e meio, conforme mostra o histograma a seguir. Destaca-se que não foram encontradas informações sobre o cronograma de obras das usinas que iniciaram operações antes de 2005. Todavia, essas unidades somam apenas cinco casos de um total de 452 usinas eólicas identificadas pelo presente trabalho.

Gráfico 8. Histograma do tempo de construção das usinas eólicas (2005-2017)



Fonte: Elaboração própria, com base em dados da Aneel (www.aneel.gov.br).

4.2 Efeitos locais da construção de usinas eólicas no Brasil: evidências anteriores

Há estudos que já procuraram analisar a importância da introdução de energias renováveis em âmbito local, tanto do ponto de vista setorial como por meio de análises quantitativas de impacto. As obras de construção dos parques eólicos envolvem a mobilização de investimentos em capital físico e mão de obra, o que dinamizaria as economias locais. Além disso, pelo fato de poder coexistir com outras atividades ligadas ao uso da terra, os parques eólicos podem ajudar o desenvolvimento econômico de regiões agrícolas, com capacidade de, dessa forma, melhorar esse setor, aumentando a qualidade de vida e reduzindo desigualdades sociais (AIE, 2002; COSTA; PRATES, 2005). Mais especificamente, o investimento em energia eólica poderia estar associado ao aumento de renda dos pequenos proprietários de terras em regiões ermas do Brasil. Isso se deve aos arrendamentos de suas terras para a instalação de torres eólicas (Abeeólica).[8]

[8] Ver *site* da Abeeólica – Associação Brasileira de Energia Eólica – www.portalabeeolica.org.br.

Além desses estudos setoriais, na literatura, há uma crescente série de estudos empíricos que procuraram medir o impacto de usinas eólicas em âmbito local no Brasil. Simas (2012) procurou verificar a contribuição da energia eólica para a geração de empregos no Brasil, tomando como base de dados primários uma série de entrevistas com 18 parques eólicos e empresas de componentes para o setor e utilizando uma análise baseada em matriz insumo-produto. O trabalho chegou à conclusão de que o setor de energia eólica tem o potencial de geração de 330 mil empregos até 2020, principalmente no setor de construção.

Estudando especificamente os efeitos em municípios, Resende (2015) procurou verificar o impacto da construção de usinas eólicas sobre os preços dos aluguéis nos entornos. O trabalho utilizou dados do Censo Demográfico (IBGE) de 1991 e 2010 para cobrir 18 municípios que receberam alguma usina eólica entre esses anos. Por meio da metodologia do controle sintético, o estudo concluiu que os aluguéis ficaram 8,3% mais caros nos municípios que receberam usinas eólicas do que o que foi previsto pelo grupo de controle.

Especificamente para os municípios da região Nordeste, Rodrigues, Gonçalves e Chagas (2016) procuraram analisar o impacto dos parques eólicos no nível de emprego, na massa salarial e no número de firmas em cada unidade. A análise partiu do cruzamento entre dados da Aneel com dados de mercado de trabalho da Rais, do Ministério do Trabalho. Por meio de um pareamento por escore de propensão, o estudo verificou que os municípios com usinas eólicas apresentaram maior massa salarial e número de firmas que as demais, ao passo que o efeito sobre o emprego não foi significante.

Já para o caso da região Sul, Rintzel, Alves e Massuquetti (2017) cruzaram dados de mercado de trabalho da Rais com dados de valor adicionado desagregado por grandes setores do IBGE e informações de receitas de impostos do Sistema de Coleta de Dados Contábeis e Fiscais dos Entes da Federação (SISTN) em 1999, 2006 e 2013. O estudo procurou verificar efeitos da instalação de parques eólicos nos municípios por meio do método estrutural diferencial, o qual é baseado na comparação do desempenho dos municípios tratados com a média da região para cada variável. O trabalho verificou resultados positivos para todas as variáveis, especialmente para o nível de emprego.

Ou seja, os estudos empíricos levantados destacaram potenciais efeitos positivos da construção de usinas eólicas em suas localidades. Todavia, esses estudos observaram casos específicos, e não controlaram o problema da micronumerosidade de casos tratados, o que compromete inferências estatísticas sobre a avaliação. O único estudo que tratou esse problema pelo método do controle sintético trabalhou com apenas um ponto no período pré-tratamento, de modo que suas previsões podem não ter consistência. Dessa maneira, a contribuição do presente trabalho é exata-

mente utilizar uma metodologia apropriada para avaliação com poucos tratados, valendo-se de ampla disponibilidade de dados para melhor verificar o impacto da construção de usinas eólicas nos municípios brasileiros.

Nesse levantamento da literatura empírica sobre os impactos locais da construção de usinas eólicas, é importante destacar o trabalho de Assunção, Costa e Szerman (2016), o qual, ainda que tenha se focado em usinas hidrelétricas, foi a principal inspiração metodológica do presente artigo. Esse trabalho procurou avaliar os efeitos da construção de usinas hidrelétricas em 82 municípios brasileiros entre 2002 e 2011 em uma série de indicadores socioeconômicos municipais, tais como PIB *per capita*, taxa de crescimento do PIB, número de empregos formais, número de empresas formais e tamanho da população. Os autores utilizaram a metodologia do controle sintético, a qual, por se basear na construção de 82 estudos de caso comparáveis, permite a estimação dos efeitos dinâmicos desde o início da construção das usinas, assim como o cálculo do efeito mediano da construção e a distribuição desses efeitos por ano, no curto e no médio prazo. O estudo observou que o impacto da construção das usinas hidrelétricas é estimulado apenas no curto prazo, tendendo a zero no quinto ano após o início das obras. O único impacto de médio prazo foi observado no nível de emprego formal no município. Por fim, os impactos nos municípios são muito heterogêneos, em todos os indicadores estimados.

4.3 Efeitos locais da construção de usinas eólicas no Brasil: estatísticas descritivas

O presente estudo de caso pretende calcular o impacto da construção de usinas eólicas sobre o PIB *per capita* dos municípios. Em virtude da disponibilidade de dados, a avaliação considera como unidades tratadas os 37 municípios beneficiados de 2007 a 2014 por entrada em operação de sua primeira usina eólica, expostos na Tabela 2. Já os potenciais controles (isto é, o *donor pool*) inclui um total de 5.490 municípios que não têm usinas eólicas. Esse conjunto não inclui o município de Gravatá (PE), em que, apesar de sua primeira usina eólica ter entrado em construção em 2010, suas obras se iniciaram apenas em 2005, de modo que faltam dados para a análise. Outros 42 municípios que receberam sua primeira usina eólica antes de 2007 ou após 2014 foram considerados contaminados e excluídos da amostra. O exercício realizado procurou construir um controle sintético para cada unidade tratada, a partir de combinações de municípios não tratados de todo o Brasil. Considerando que os maiores efeitos sobre as economias locais são provenientes das obras de implantação dos parques, para melhor capturar os efeitos da construção das usinas eólicas, o controle sintético foi aplicado para cada caso com base no ano de início das obras civis das estruturas, de acordo com o Acompanhamento das Centrais Geradoras Eólicas, da Aneel. Para evitar distorções, alguns ajustes foram realizados. Para as obras que duraram mais de um ano, com seu mês de início entre julho e dezembro, o ano de tratamento foi deslocado para o ano imediatamente

posterior, uma vez que a maior parte das suas obras acontecera nesse ano. Para as obras que duraram até um ano, considerou-se o ano de tratamento aquele em que ocorrera o maior número de dias de obras.

Tabela 2. Lista de municípios avaliados, ano e potência de sua primeira usina eólica

Município	Tratamento	Potência outorgada (MW)	Anos de obras
Acaraú – CE	2008	70.800	2
Amontada – CE	2009	54.600	0
Aracati – CE	2007	10.500	1
Areia Branca – RN	2012	20.000	1
Barra dos Coqueiros – SE	2012	34.500	0
Beberibe – CE	2008	25.600	0
Boituva – SP	2012	2,24	0
Brotas de Macaúbas – BA	2011	95.190	1
Cabo de Santo Agostinho – PE	2012	2.000	0
Caetité – BA	2011	296.820	3
Camocim – CE	2008	105.000	1
Cururupu – MA	2008	22,5	0
Galinhos – RN	2012	118.570	2
Guamaré – RN	2010	51.000	0
Guanambi – BA	2011	167.840	3
Igaporã – BA	2011	143.840	3
Itarema – CE	2012	30.000	2
Iturama – MG	2012	156	0
João Câmara – RN	2011	39.600	1
Macaparana – PE	2010	4.950	0
Mataraca – PB	2007	10.200	0
Palmares do Sul – RS	2010	9.200	0
Paracuru – CE	2007	25.200	1
Parazinho – RN	2011	466.000	3
Parnaíba – PI	2008	18.000	0
Pedra Grande – RN	2012	118.400	2
Pelotas – RS	2014	1,98	0
Pombos – PE	2009	4.950	1
Sant'Ana do Livramento – RS	2011	60.000	0
São Francisco de Itabapoana – RJ	2010	28.050	0
São Miguel do Gostoso – RN	2013	51.200	1
Sento Sé – BA	2012	90.000	1
Sobradinho – BA	2012	48.000	1
Trairi – CE	2012	55.392	1
Tramandaí – RS	2010	70.000	1
Tubarão – SC	2014	2.099,5	0
Xangri-lá – RS	2014	27.675	0

Fonte: Elaboração própria, com base em dados da Aneel.

A maior parte dos municípios tratados foi afetada por obras de até dois anos, conforme mostra o Gráfico 9. As obras mais duradouras, por outro lado, são geralmente associadas a parques eólicos maiores, com maior potência outorgada.

Gráfico 9. Usinas eólicas por anos de obras (municípios tratados)

Fonte: Elaboração própria, com base em dados da Aneel.

A Tabela 3, a seguir, apresenta estatísticas descritivas sobre os municípios tratados em relação ao total de municípios elegíveis como unidades de comparação (*donor pool*), considerando-se o ano imediatamente anterior aos primeiros tratamentos (2006). Apresentam-se a média e o desvio-padrão de 16 indicadores com o objetivo de representar o perfil dos dois grupos de municípios. Conforme mostra a tabela, os tratados estavam, em 2006, com uma renda anual média por habitante acima da média nacional. Além disso, apresentaram maior porte do que os não tratados em população, número de estabelecimentos, receitas totais, despesas com saúde e educação e extensão da rede de água. Já os potenciais controles apresentaram vantagem quanto ao PIB, receitas tributárias, menores homicídios *per capita* e extensão da rede de esgotos. Pelos dados apresentados, não é possível inferir diferenciais de níveis de desenvolvimento entre os dois grupos.

Tabela 3. Resultados comparados: perfil das unidades (2006)

Variável	*Donor pool*		Tratados	
	Média	Desvio-padrão	Média	Desvio-padrão
PIB (R$)	432.458.035,18	4.687.964.973,25	402.910.918,92	592.410.986,73
PIB *per capita* (R$)	8.052,04	9.605,62	9.167,03	9.715,34
Adm. Pública/PIB (%)	33,37	17,00	30,94	16,97
Agropecuária/PIB (%)	22,27	14,72	12,07	9,72
Indústria/PIB (%)	13,77	14,92	25,01	24,25
Serviços/PIB (%)	30,59	12,57	31,99	16,00
População	33.017,24	197.831,82	49.427,19	62.365,64
N. estabelecimentos	445,85	3.596,20	520,84	1.042,37
N. vínculos de emprego	6.250,85	66.662,33	6.087,70	10.492,52
Massa salarial (R$)	7.343.984,40	109.022.685,90	5.094.955,44	10.576.407,88
Receita total (R$)	33.571.586,64	280.819.705,38	42.600.486,33	49.834.522,83
Receita tributária (R$)	6.426.819,96	119.969.415,13	4.137.954,52	6.393.477,11
Desp. educação (R$)	8.170.058,99	58.382.637,37	11.505.349,30	11.511.485,81
Desp. saúde (R$)	7.280.861,06	51.270.657,61	9.035.692,83	12.589.765,74
Óbitos por homicídios por cem mil habitantes	12,56	16,93	14,23	16,61
Extensão da rede de água (km)	89,45	143,03	59,55	94,68
Extensão da rede de esgotos (km)	96,59	425,65	110,50	162,49
Total de municípios	5.490		37	

Fonte: Elaboração própria, com base em dados de IBGE, Rais, Finbra, SIM-Datasus e Snis.

4.4 Efeitos locais da construção de usinas eólicas no Brasil: resultados do MARVIm para avaliação individual em Tramandaí (RS)

Conforme mencionado anteriormente, este trabalho busca fazer uma análise de impacto compilada de 37 municípios que receberam investimentos em parques eólicos. Para esclarecer a presente seção, que descreve uma avaliação realizada pelo MARVIm – Módulo de Controle Sintético, serão apresentados os procedimentos referentes a um caso específico. Esse caso corresponde à avaliação do impacto da construção de uma usina eólica sobre o PIB *per capita* do município de Tramandaí (RS), iniciada no ano de 2010. A Tabela 4, a seguir, apresenta algumas informações elementares desse município.

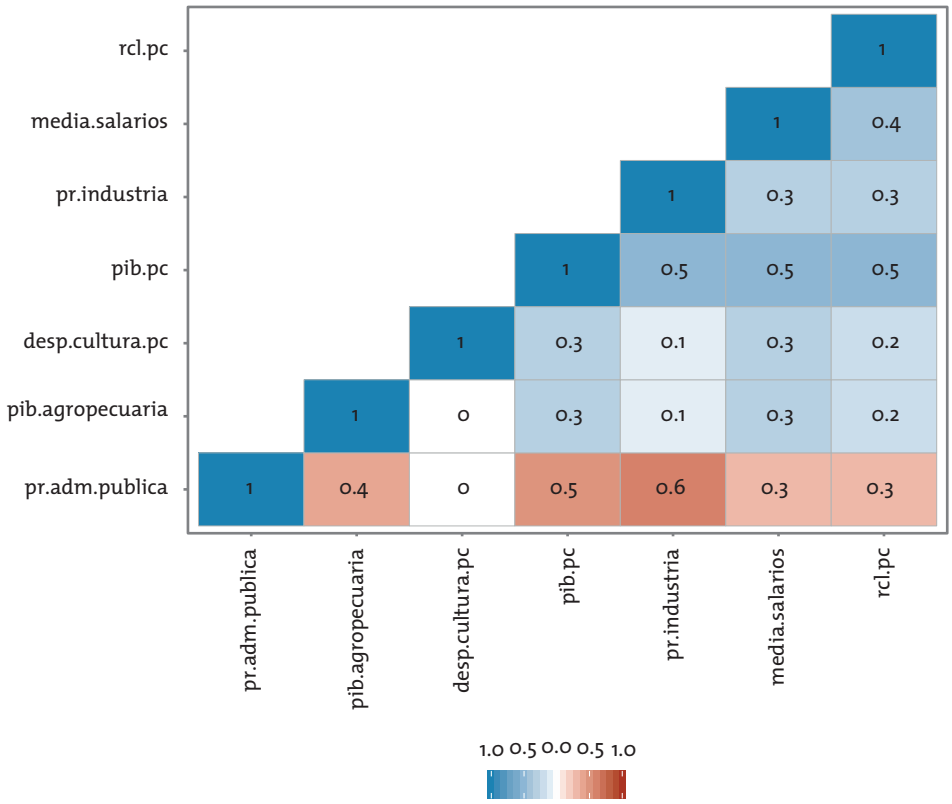
Tabela 4. Informações de Tramandaí (RS)

Indicador	Valor	Último dado disponível
Código IBGE	4321600	2016
Nome	Tramandaí – RS	2016
Bioma (IICA)	Linha de Costa	2016
Nível de regiões de influência das cidades (Regic)	4B	2016
Nome Regic	Centro de Zona B	2016
Área (km²)	144.408	2016
População	46.962	2016
PIB (R$)	876.973.000	2015
PIB *per capita* (R$)	18.912,92	2015
Agropecuária/PIB (%)	1,62	2015
Indústria/PIB (%)	11,21	2015
Adm. pública/PIB (%)	24,64	2015
Serviços/PIB (%)	62,52	2015
Índice de Gini (Rais)	0,31	2013
Densidade demográfica (hab./km²)	325,2	2016

Fontes: IBGE, Rais e Instituto Interamericano de Cooperação para a Agricultura (IICA).

Uma vez que a base de dados e os *inputs* básicos da avaliação estejam definidos, o primeiro passo de uma avaliação do MARVIm é a escolha das covariadas que melhor explicam a variável de interesse escolhida. O objetivo desse procedimento é reduzir a dimensionalidade do modelo empírico. Esse processo é realizado por meio do adaLASSO, uma metodologia automatizada justamente para esse fim. Após aplicada essa metodologia, o MARVIm retorna a lista das variáveis mais relevantes para explicar a variável de interesse selecionada na forma de um gráfico de matriz de correlações, denominado correlograma (Gráfico 10). No caso apresentado, a variável de interesse é o PIB *per capita* municipal, e as estimativas do adaLASSO consideraram boas variáveis preditivas a receita corrente líquida *per capita*, a média salarial, a participação do setor industrial no PIB, a despesa do governo municipal com cultura *per capita*, o PIB da agropecuária e a participação da administração pública no PIB.

Gráfico 10. Correlograma das variáveis selecionadas pelo adaLASSO



Fonte: Elaboração própria.

Uma vez obtidas as covariadas mais relevantes para explicar o indicador de interesse, o MARVIm pode fazer a estimação do controle sintético. Conforme explicado anteriormente, esse controle é uma unidade artificial baseada na ponderação de todas as unidades elegíveis para comparação (*donor pool*), já filtradas pelo processo de clusterização, que melhor se ajuste à trajetória do PIB *per capita* de Tramandaí antes de 2010, que é o ano de tratamento. Os resultados gerados pelo MARVIm consistem em três tabelas, as quais gerarão todos os gráficos e demais figuras presentes no relatório de avaliação. A primeira tabela (Tabela 5) descreve as médias de cada covariada selecionada para a unidade tratada, para o controle sintético e para todas as potenciais unidades de controle, assim como indica os pesos de cada uma na construção do controle sintético. A Tabela 6 apresenta os pesos das unidades de controle para o mesmo caso. A Tabela 7, por sua vez, contém os resultados da estimação do controle sintético, em que são representados os valores da variável de interesse para a unidade tratada, a unidade de controle sintético e a diferença entre elas. Os resultados observados para o caso de Tramandaí (RS) são os seguintes:

Tabela 5. Médias e peso das covariadas na construção do controle sintético

Variável	Tratado	Controle sintético	*Donor pool*	Peso da variável (%)
PIB *per capita* (R$)	7.524,00	7.512,03	8.201,91	64,56
Despesa com cultura *per capita* (R$)	3,83	4,53	8,07	22,15
Valor agregado da agropecuária (R$)	3.453.040,11	4.883.471,57	11.753.222,06	9,99
Receita corrente líquida *per capita* (R$)	528,88	501,11	202,81	3,23
Indústria/PIB (%)	9,94	7,52	6,48	0,02
Administração Pública/PIB (%)	24,88	27,87	25,91	0,02
Média de salários (R$)	760,38	704,3	702,58	0,02

Fonte: Elaboração própria.

Tabela 6. Peso das unidades de controle na construção do controle sintético

Código do controle	Nome do controle	Peso do controle (%)
4305454	Cidreira – RS	55,76
4201950	Balneário Arroio do Silva – SC	33,72
3304300	Rio Bonito – RJ	9,10
5204656	Campinaçu – GO	1,42

Fonte: Elaboração própria.

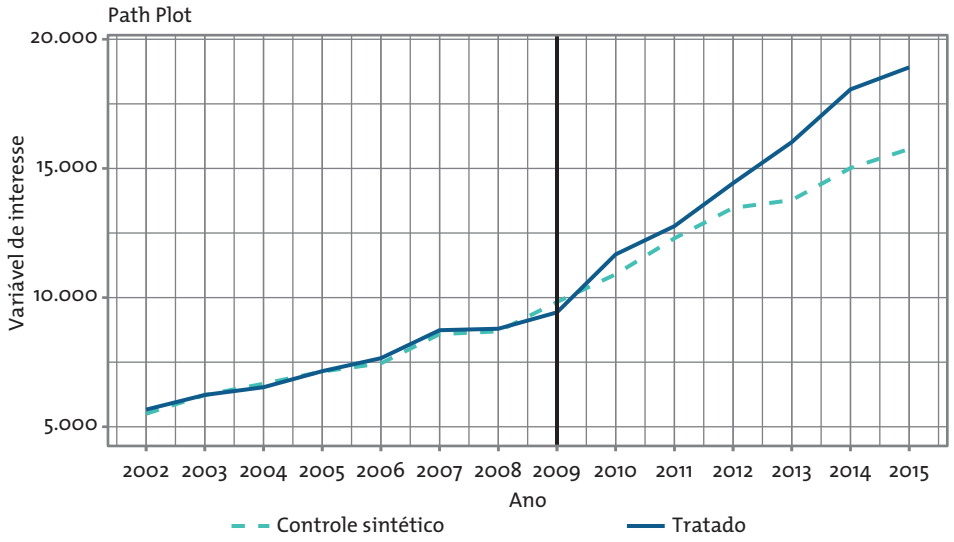
Tabela 7. Resultados da estimação do controle sintético

Ano	Tratado	Controle sintético	*Donor pool*	Diferença	Diferença (%)
2002	5.663,76	5.503,73	5.558,15	160,02	2,91
2003	6.231,50	6.228,73	7.230,28	2,78	0,04
2004	6.530,37	6.664,30	7.222,09	(133,94)	(2,01)
2005	7.151,79	7.136,99	6.873,83	14,80	0,21
2006	7.653,92	7.453,60	8.020,34	200,32	2,69
2007	8.737,71	8.586,39	9.231,00	151,32	1,76
2008	8.794,02	8.694,05	10.412,39	99,97	1,15
2009	9.428,95	9.828,46	11.067,19	(399,51)	(4,06)
2010	11.673,15	10.901,86	12.751,32	771,30	7,07
2011	12.765,06	12.295,67	14.608,72	469,39	3,82
2012	14.430,20	13.472,97	15.279,47	957,22	7,10
2013	16.018,37	13.774,76	18.859,73	2.243,61	16,29
2014	18.059,92	15.014,26	20.124,59	3.045,66	20,29
2015	18.912,92	15.752,29	21.448,77	3.160,63	20,06

Fonte: Elaboração própria.

O objetivo desse procedimento é encontrar a trajetória aproximada do indicador de interesse que a unidade tratada provavelmente teria seguido caso não tivesse sofrido a intervenção. Para isso, calcula-se a média ponderada das unidades não tratadas cujos pesos minimizam a distância do comportamento da unidade tratada no período anterior ao tratamento. Em seguida, a trajetória do controle sintético é projetada no período pós-tratamento e comparada com a trajetória da unidade tratada. A diferença entre essas duas trajetórias é entendida como o efeito do tratamento. O Gráfico 11, denominado Path Plot, ilustra esse procedimento, tendo em vista o caso já mencionado de Tramandaí (RS).

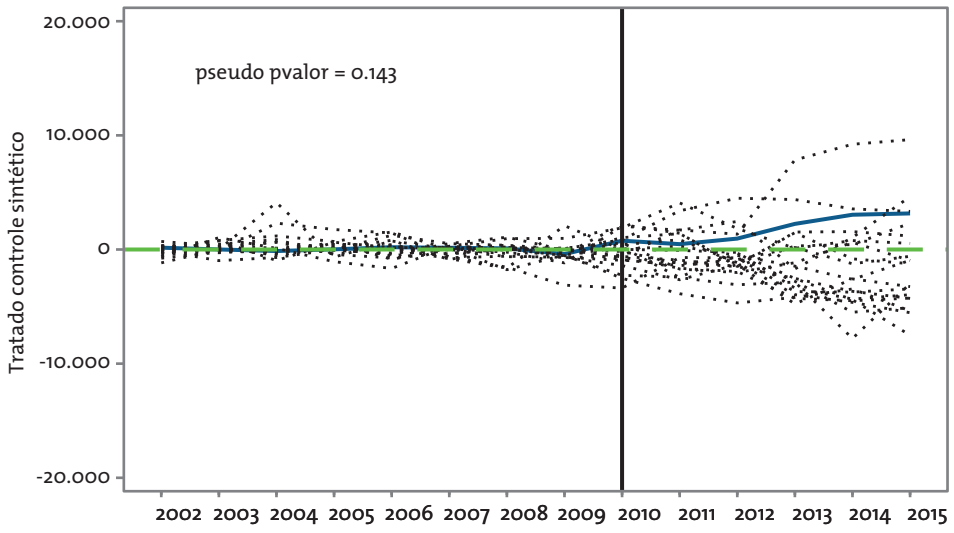
Gráfico 11. Trajetória do PIB *per capita* de Tramandaí (RS) e de seu controle sintético



Fonte: Elaboração própria.

O passo seguinte da avaliação individual por meio do MARVIm é a realização do teste de placebo. Esse teste consiste na realização do mesmo exercício para um número escolhido de unidades não tratadas presentes na base de dados, considerando-se o mesmo indicador de interesse e o mesmo ano de suposto tratamento. Para evidenciar que o tratamento tenha tido efetividade, a trajetória da variável de interesse (o PIB *per capita*) deve ser mais elevada para a unidade tratada do que para os testes-placebo. A probabilidade de isso ocorrer, dada a amostra disponível de testes-placebo, é computada como um pseudo *p-valor*. O Gráfico 12 ilustra essa noção, para o caso da avaliação de Tramandaí (RS). Esse gráfico é denominado Gaps Plot, e é comum na bibliografia sobre controle sintético. Aqui, a unidade

Gráfico 12. Diferença entre cada unidade e seu controle sintético

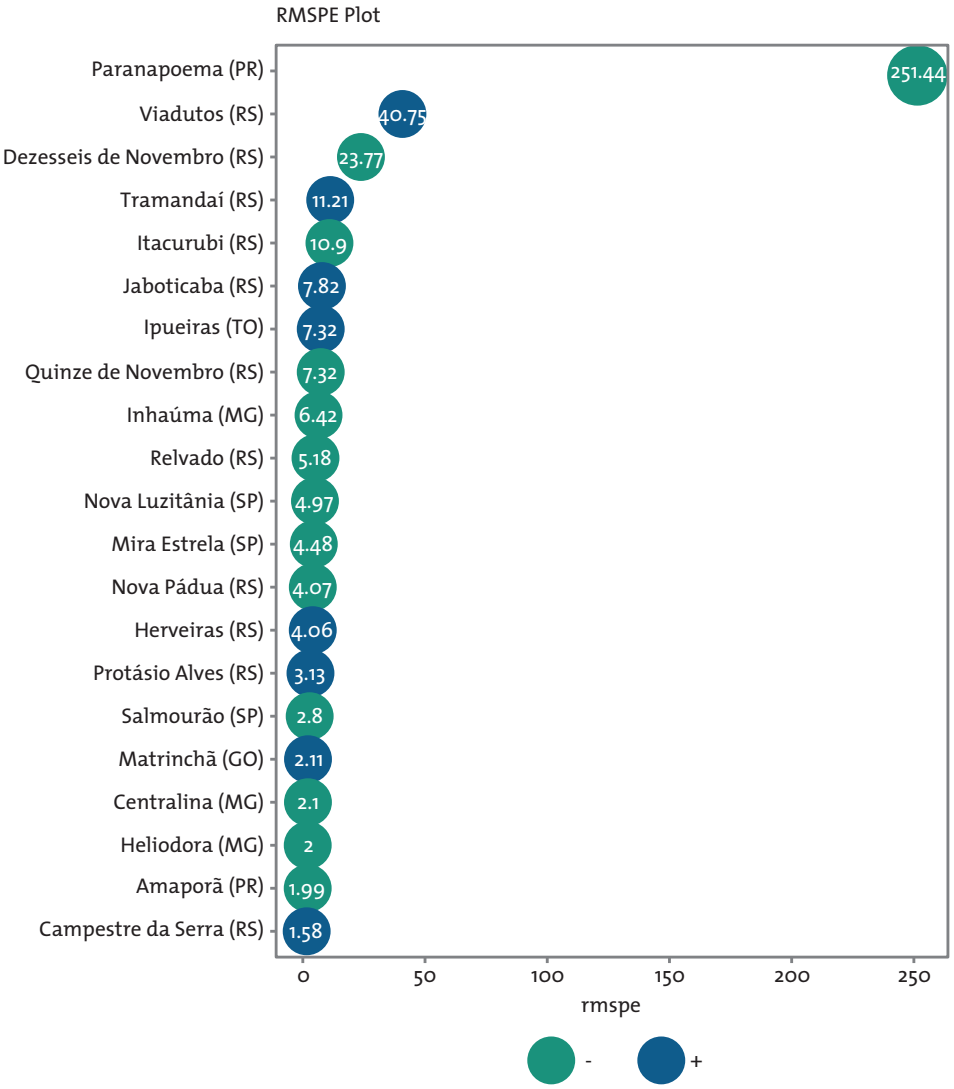


Fonte: Elaboração própria.

tratada está representada pela curva azul, e cada teste-placebo está representado por uma curva pontilhada.

Após a realização dos testes-placebo, é feito o cálculo das duas estatísticas de ajuste da estimação para cada caso. Conforme descrito na seção anterior, a primeira estatística é a razão entre a raiz do erro quadrático médio entre cada unidade e seu controle sintético em cada ano, depois e antes do tratamento (RMSPE *ratio*). A segunda estatística é a razão entre a área sob a curva da diferença entre o tratado e o controle sintético depois e antes do tratamento (AUC *ratio*). Os casos relacionados à avaliação de Tramandaí (RS) estão a seguir. Os gráficos estão sinalizados de modo que, se na média após o tratamento a trajetória unidade se manteve acima de seu controle sintético, seu sinal é positivo; caso contrário, é negativo.

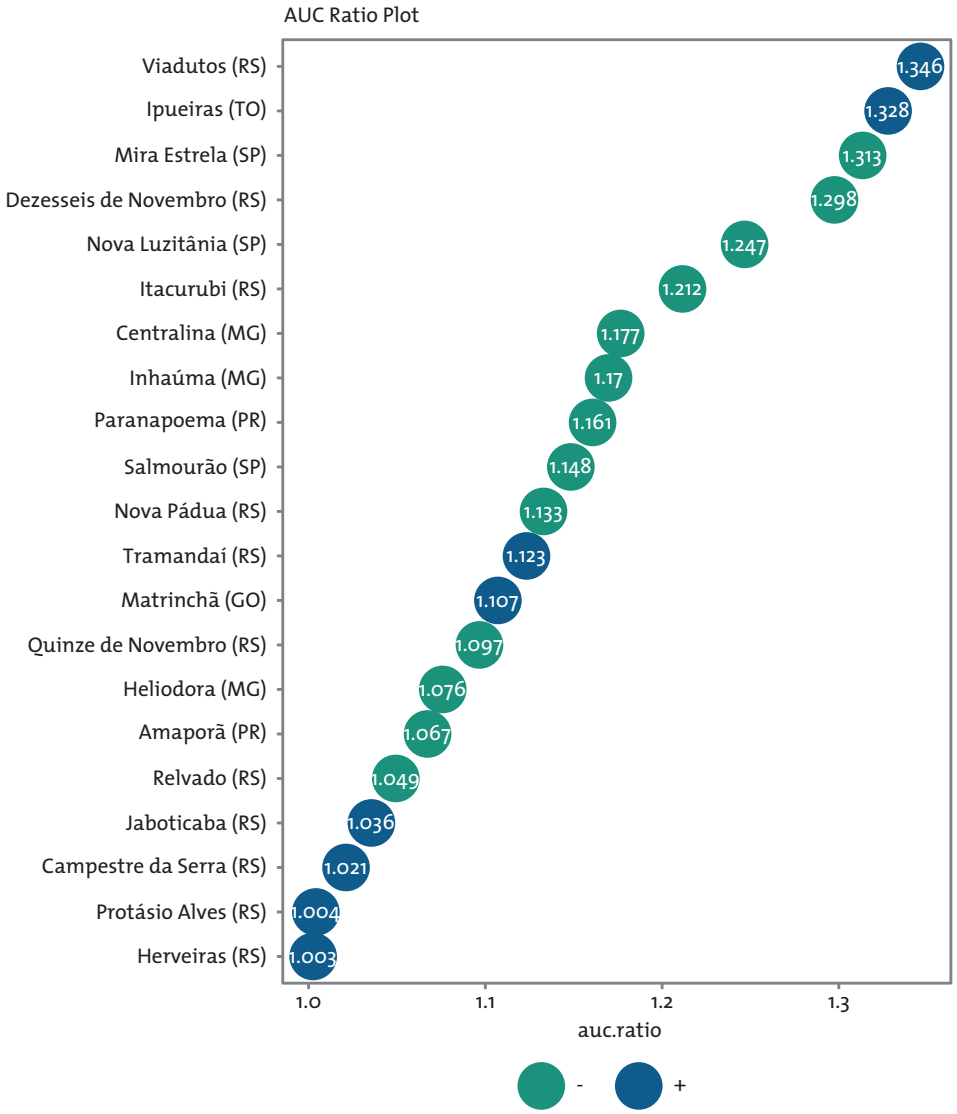
Gráfico 13. Razão do erro quadrático médio depois e antes do tratamento



Fonte: Elaboração própria.

Obs.: Não estão representados os casos em que ocorreram ajustes perfeitos.

Gráfico 14. Razão do erro quadrático médio depois e antes do tratamento



Fonte: Elaboração própria.

Para o caso de Tramandaí, pode-se observar que a análise do MARVIm demonstrou impacto positivo para a intervenção realizada (a construção da usina eólica), o qual foi mais intenso a partir de 2013, isto é, três anos após o início das obras. Todavia, a magnitude desse impacto se mostrou modesta em comparação com os testes de placebo, os quais se mostraram, respectivamente, o segundo e o terceiro município com maior variação positiva pelo RMSPE *ratio* e pelo AUC *ratio*. O melhor desempenho computado pela primeira estatística decorre do bom ajuste pré-tratamento do controle sintético ao caso do município de Tramandaí.

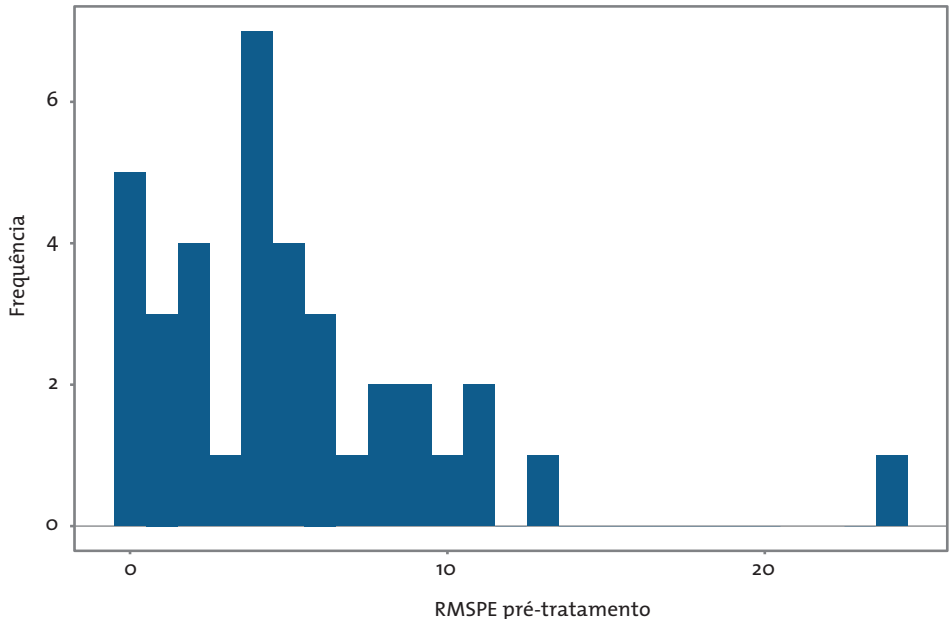
4.5 Efeitos locais da construção de usinas eólicas no Brasil: resultados do MARVIm para avaliação compilada

Nessa seção, o método do controle sintético foi aplicado individualmente para comparar cada município tratado com seu contrafactual. Foram realizadas 37 avaliações individuais. O intuito foi fazer uma análise agregada de todos os municípios brasilei-

ros que tiveram usinas eólicas entrando em operação de 2007 a 2014, a partir de cada avaliação individual. Para melhor capturar o efeito da construção das usinas eólicas, o controle sintético foi aplicado considerando-se como o momento do tratamento o início das obras civis das estruturas da primeira usina de cada município, incluindo os ajustes explicados na seção anterior. Conforme anteriormente relatado, para construir a trajetória do PIB *per capita* em cada cenário contrafactual, o MARVIm, por meio do método adaLASSO de seleção de covariadas, considerou relevantes seis variáveis em âmbito municipal que a explicam: a receita corrente líquida *per capita* (variável derivada a partir de dados do Finbra e do IBGE), o salário médio (Rais), a proporção da indústria no PIB (IBGE), a despesa *per capita* em cultura (variável derivada a partir de dados do Finbra e do IBGE), o PIB da agropecuária (IBGE) e a proporção da administração pública no PIB (IBGE). Os resultados das avaliações individuais encontram-se em apêndice ao presente trabalho.

Para fins de compilação dos resultados individuais, foram desconsiderados do trabalho três municípios: Guamaré (RN), Sobradinho (BA) e Galinhos (RN). Os dois primeiros apresentaram resultados discrepantes dos outros por razões alheias à produção de energia eólica. Guamaré sofreu sua a maior crise no ano de 2011 por conta da queda dos preços do petróleo na região, setor do qual sua economia é muito dependente. Sobradinho teve uma de suas maiores secas em 2013, a qual esvaziou seu reservatório hídrico e comprometeu o fornecimento de energia advinda de sua usina hidrelétrica. Para o caso do município de Galinhos, houve problemas com o ajuste no período pré-tratamento no controle sintético, de modo que não foi encontrado um contrafactual consistente para a verificação de impacto. Seu RMSPE pré-tratamento foi igual a 23,9, muito acima da média da amostra (5,1), conforme mostra o Gráfico 15. Restaram 34 casos para a avaliação compilada.

Gráfico 15. Histograma da distribuição de RMSPE pré-tratamento nos 37 casos individuais avaliados



Fonte: Elaboração própria.

A primeira análise feita com base nos resultados consistiu em uma comparação da proporção de casos em que o PIB *per capita* dos tratados ficou acima dos valores observados para seus respectivos controles. Intuitivamente, assumindo que o método não é viesado para a definição dos controles, deveríamos observar que, se a construção das usinas eólicas não tiver nenhum impacto sobre as localidades, essa proporção deveria circular no entorno de 50% ao longo do tempo – tal como ocorre com a proporção de "caras" e "coroas" após *n* lançamentos de uma moeda não viciada.

A Tabela 8, a seguir, mostra as estatísticas descritivas dos municípios tratados e a proporção de casos positivos por ano de referência, em que *t* é o ano de tratamento para cada caso.[9] Para julgar se esses valores são estatisticamente significantes, um intervalo de confiança foi construído com base em uma distribuição de Bernoulli, assumindo um parâmetro *p* de 50%. Dessa maneira, a hipótese nula assume que em cada momento do tempo há 50% de chance de um município tratado estar melhor que seu controle. Caso a proporção observada fique acima do limite superior dos intervalos de confiança, há evidência de que a proporção de comparações positivas será estatisticamente superior à de comparações negativas.

Tabela 8. Proporção de efeitos positivos por ano de referência

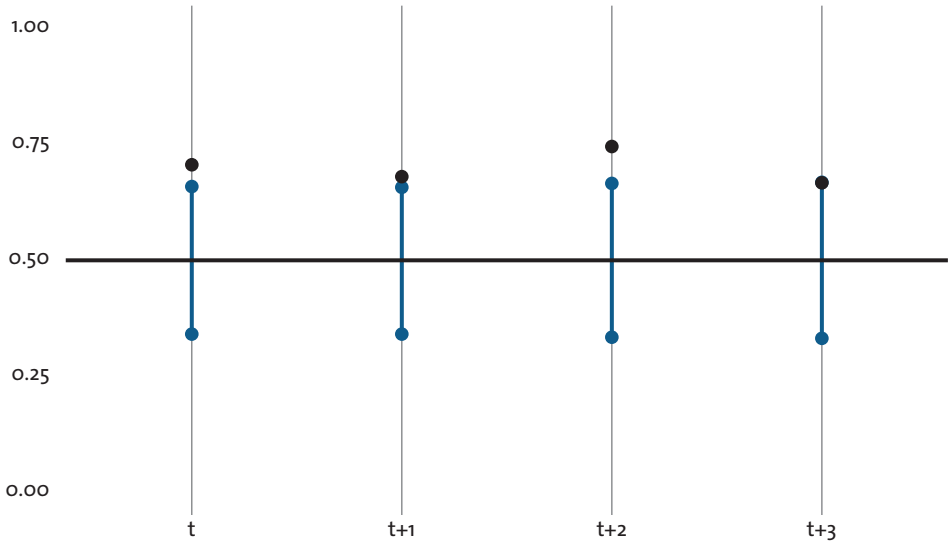
Ano	N. de casos positivos	N. de casos negativos	Proporção de casos positivos	Limite inferior	Limite superior	*p*
t	24	10	0,71	0,34	0,66	0,5
t + 1	23	11	0,68	0,34	0,66	0,5
t + 2	23	8	0,74	0,33	0,67	0,5
t + 3	20	10	0,67	0,33	0,67	0,5

Fonte: Elaboração própria.

O Gráfico 16 representa um gráfico de dispersão centrado em 0,50, com pontos marcando os valores das proporções de efeitos positivos. A linha azul vertical representa o intervalo de confiança estimado por uma aproximação qui-quadrado. De acordo com o gráfico, é possível observar o descolamento da proporção de efeitos positivos em relação ao intervalo de confiança até dois anos após o tratamento, nova evidência favorável ao impacto positivo das usinas sobre as economias locais. No terceiro ano, a proporção de efeitos positivos continua maior do que a de efeitos negativos, mas toca o limite superior do intervalo de confiança. Isso pode ser uma indicação do arrefecimento dos efeitos das obras com menor tempo de duração.

[9] O ano *t* equivale ao ano de construção do parque eólico, sendo *t* + 1 o ano em que ele entrou em operação.

Gráfico 16. Intervalo de confiança e proporção de efeitos positivos para cada ano de referência



Fonte: Elaboração própria.

Outra análise levou em consideração a magnitude dos impactos estimados, representados pela diferença percentual entre a trajetória de cada unidade tratada e seu controle sintético em cada ano. Os resultados individuais foram normalizados, de modo que o ano de tratamento para todos os casos foi alinhado em *t*, e os demais anos de análise seguiram sua referência. Para cada ano de análise, os resultados individuais foram compilados, sendo calculadas algumas estatísticas. Em virtude da heterogeneidade dos resultados individuais, observados pelos elevados desvios-padrão, optou-se por observar primeiramente as medianas, assim como os percentis 0,25 e 0,75 da distribuição. Esse procedimento já havia sido adotado por Assunção, Costa e Szerman (2016), os quais também se depararam com resultados individuais heterogêneos sobre os efeitos municipais de obras de usinas hidrelétricas. Os resultados encontrados estão representados na Tabela 9.

Tabela 9. Resultados compilados por ano de referência (%)

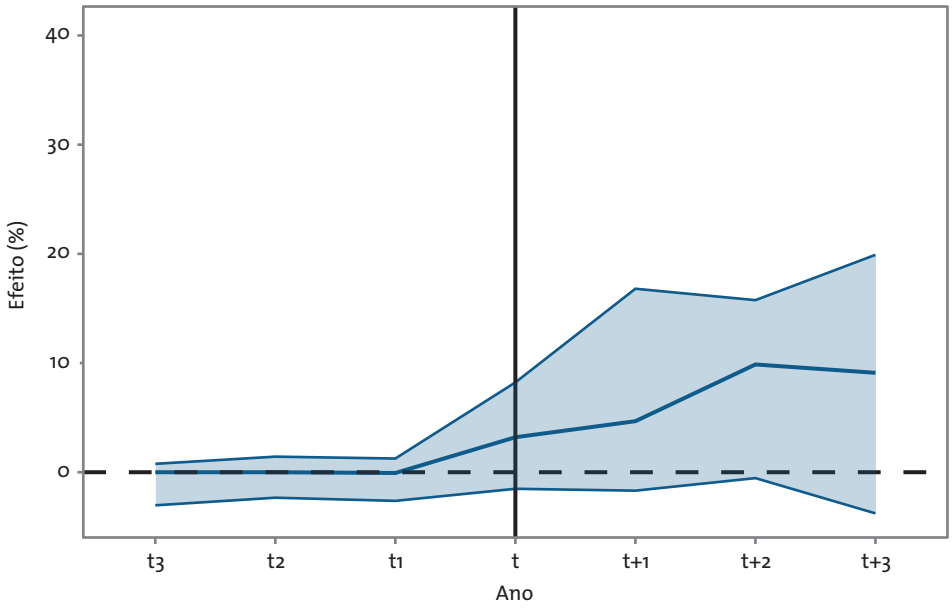
Ano	Média	DP	P. 0,25	Mediana	P. 0,75	Mínimo	Máximo
t - 3	(1,32)	5,01	(3,03)	0,00	0,77	(18,71)	9,7
T - 2	(0,61)	5,01	(2,33)	0,00	1,43	(11,08)	12,6
t - 1	(1,08)	3,90	(2,62)	(0,07)	1,26	(15,61)	6,21
t	10,45	47,56	(1,52)	3,21	8,23	(24,47)	273,48
***t* + 1**	36,16	179,27	(1,69)	4,68	16,81	(37,05)	1046,45
***t* + 2**	27,37	89,99	(0,54)	9,87	15,76	(48,25)	484,52
***t* + 3**	40,44	109,23	(3,77)	9,11	19,91	(46,65)	475,68

Fonte: Elaboração própria.

O Gráfico 17 desenha o comportamento do efeito percentual das diferenças da variável PIB *per capita* entre tratados e controles ao longo do tempo. A linha

azul central mostra o efeito mediano entre todos os municípios tratados. As linhas secundárias abaixo e acima da mediana indicam os efeitos no primeiro e terceiro quartis, respectivamente. Observa-se que, em geral, há efeito mediano positivo no período pós-tratamento dos municípios sob análise, de 4,7% no primeiro ano, 9,9% no segundo ano e 9,1% no terceiro ano. Também se observa um aumento da dispersão das linhas no último ano. Novamente, isso pode estar associado ao arrefecimento dos efeitos das obras mais curtas sobre as economias locais.

Gráfico 17. Mediana das diferenças percentuais de cada unidade tratada em relação a seu controle sintético



Fonte: Elaboração própria.

Outra questão importante a ser estudada diz respeito aos componentes do impacto econômico das usinas eólicas nos municípios. Conforme levantado na bibliografia, os efeitos podem ocorrer tanto em virtude da mobilização de insumos para as obras de construção dos parques eólicos como em razão das receitas oriundas do arrendamento de terras após sua implantação. Esse ponto foi abordado pelo presente trabalho a partir de um exercício baseado em regressões. Nessas regressões, o efeito percentual calculado sobre cada município de t até $t + 3$ foi modelado em função de um indicador de potência referente à obra em curso e da potência já instalada no município. Além disso, a fim de obter a avaliação de acordo com o potencial econômico do município tratado em relação ao investimento recebido, foi criada uma variável de dose do tratamento. Essa variável equivale à razão entre o total de potência outorgada do parque eólico, em megawatts (MW), e o PIB *per capita* do município. A dose foi calculada com o valor de PIB *per capita* fixado no período em que as primeiras usinas estavam em construção. Todos os indicadores utilizados nas regressões foram normalizados pelas suas médias, de modo a permitir uma medida mais intuitiva de comparação.

Para cada par de variáveis independentes foi estimado um modelo baseado em mínimos quadrados ordinários empilhados (MQOE) e outro modelo baseado em efeitos fixos (EF). Intuitivamente, o modelo de mínimos quadrados trata cada par indivíduo-tempo como uma unidade independente na amostra. O modelo de efeitos fixos, por sua vez, controla o efeito de possíveis heterogeneidades individuais ao normalizar o valor de cada variável pela média do indivíduo ao qual está associado. A Tabela 10 mostra os resultados das regressões estimadas. Em nenhuma delas as medidas referentes à estrutura instalada no município foram estatisticamente significantes. A potência da obra foi significante nos dois modelos, sendo que o coeficiente estimado pelo MQOE teve magnitude superior ao estimado por EF. O indicador de dose da obra só foi estatisticamente significante no modelo de MQOE, o que provavelmente decorre do fato de que o modelo de EF controla o efeito do PIB *per capita* fixado no primeiro ano da obra.

Tabela 10. Resultados das estimações dos determinantes dos efeitos das usinas eólicas sobre as economias municipais

	Variável dependente: impacto no PIB *per capita* (%)			
	EF	MQOE	EF	MQOE
Potência obra	19,00**	50,00***		
	(8,80)	(7,60)		
Potência instalada	-6,40	-18,00		
	(8,20)	(11,00)		
Dose obra			10,00	25,00**
			(11,00)	(12,00)
Dose instalada			4,80	-6,20
			(8,00)	(12,00)
Intercepto		8,20		28,00***
		(9,20)		(10,00)
Observações	129	129	129	129
R2	0,05	0,28	0,02	0,04
R2 ajustado	0,03	0,27	0,01	0,04
Estatística F	2,40 (*df* = 2; 93)	24,00*** (*df* = 2; 126)	0,86 (*df* = 2; 93)	2,40* (*df* = 2; 126)

Fonte: Elaboração própria.

Nota: *p<0,1; **p<0,05; ***p<0,01

Com base nos 34 casos aqui avaliados e compilados, foi possível observar o impacto positivo dos parques eólicos sobre o PIB *per capita* dos municípios afetados, ainda que haja expressiva variabilidade. Os exercícios de avaliação concordaram em relação ao crescimento do efeito e aos anos em que se apresentou maior impacto positivo. Os efeitos, com mediana estimada entre 4,7% e 9,9%, foram mais claros entre dois e três anos após o início da construção, que estão relacionados aos maiores parques instalados. Também se verificou maior dispersão de efeitos no terceiro ano, sinalizando possível esgotamento dos efeitos das obras de menor duração. A análise por regressões baseada em modelos de efeitos fixos

e de mínimos quadrados empilhados destacou que o principal determinante dos efeitos é a magnitude da obra realizada, e não a estrutura instalada no município. Portanto, as evidências empíricas apontam para a hipótese de que o impacto das usinas sobre as economias locais vem da mobilização de recursos para as obras de instalação, e não do arrendamento das terras.

A metodologia aqui adotada é inovadora em relação às avaliações dos impactos locais das obras de usinas eólicas no Brasil. A estimação de controle sintético não apenas é mais consistente para avaliações com micronumerosidade de unidades tratadas, como também, ao contrário da literatura levantada, permitiu observar que os efeitos são diferentes ao longo do tempo após a construção e em função da dose de tratamento. Os resultados verificados, tanto em relação aos efeitos como em relação à heterogeneidade entre as unidades avaliadas, são semelhantes àqueles observados por Assunção, Costa e Szerman (2016), em um estudo sobre efeitos locais da construção de usinas eólicas. Esse trabalho também observou que os efeitos locais das obras, mensurados em indicadores de produção econômica e de mercado de trabalho, tendem a dissipar-se no quinto ano depois da construção. Essa constatação não pôde ser observada no presente trabalho em função da indisponibilidade de dados, principalmente pelo fato de que as usinas eólicas são de construção relativamente mais recente, a partir de 2014. Todavia, as evidências empíricas aqui obtidas indicam que as obras têm efeitos econômicos superiores ao do arrendamento de terras provenientes da potência instalada no município.

É importante destacar que, de acordo com os dados da Aneel, o período de maior crescimento de usinas eólicas entrando em operação no Brasil foi a partir de 2014. O presente estudo investiga parques eólicos instalados até 2014, ou seja, conta com uma base de dados pequena e um curto período para análise. Não obstante, há municípios que só foram analisados apenas um ano após o tratamento por falta de dados na base municipal.

Portanto, há uma potencial agenda de estudos futuros sobre a avaliação das eólicas sobre os municípios beneficiados. Espera-se que essa agenda seja cada vez mais consistente e mostre maiores evidências de impacto, por haver maior dosagem de investimento nos municípios e também por conter um banco de dados com maior quantidade de tratados.

5. Conclusão

O presente trabalho representa mais um esforço para proporcionar um aumento da escala das atividades de monitoramento e avaliação do BNDES. O Modelo Automatizado em R para Verificação de Impacto (MARVIm) é uma ferramenta que permite a construção sistemática de informações padronizadas sobre a efetividade

das políticas do BNDES. Essas informações são baseadas em avaliações causais, a partir de bases de dados de indivíduos apoiados e não apoiados pelo Banco. Mais especificamente, o MARVIm – Módulo de Controle Sintético faz uso do método de controle sintético, uma técnica adequada para a sistematização de estudos de caso. Isto é, trata-se de uma ferramenta para exercícios com poucas unidades tratadas, de modo que os métodos estatísticos mais comuns para a avaliação de impacto, como o pareamento e a diferenças em diferenças, são prejudicados pela micronumerosidade.

A metodologia aqui desenvolvida é particularmente útil para verificar efeitos de políticas que envolvem unidades geográficas. Nesse sentido, foi construída uma base de dados consolidada sobre os municípios brasileiros. Essa base concentra 250 variáveis quantitativas de 12 fontes para o período total de 1999 a 2015, de modo que permite a realização de avaliações sobre múltiplas dimensões da realidade local. O MARVIm – Módulo de Controle Sintético inclui consigo uma metodologia de seleção automatizada para a seleção de covariadas, denominada adaLASSO. Esse procedimento visa identificar e selecionar as variáveis mais importantes para explicar cada indicador de interesse escolhido, reduzindo, dessa maneira, a dimensionalidade da base de dados. O adaLASSO tem a vantagem de permitir a estimação de um modelo consistente na seleção das variáveis mais importantes para cada caso, além de apresentar propriedades estatísticas desejáveis, como a normalidade assintótica.

Para demonstrar o potencial da capacidade da ferramenta em verificar a efetividade de políticas públicas, realizou-se aqui um estudo de caso, com foco na análise de impacto da construção de usinas eólicas sobre o PIB *per capita* dos municípios beneficiados. Para isso, foi realizada uma avaliação individual baseada em controle sintético para os municípios que receberam sua primeira usina eólica no período de 2007 a 2014. Após isso, os resultados individuais foram compilados. De forma geral, observaram-se efeitos positivos dos parques eólicos sobre as economias municipais, ainda que com muita variabilidade entre os casos individuais. Esses resultados são compatíveis com a bibliografia levantada na análise. A mediana dos efeitos estimados oscilou entre 4,7% e 9,9% do PIB *per capita* a mais para as unidades tratadas em relação a seus respectivos controles sintéticos. Os efeitos foram maiores nos municípios que receberam investimentos em obras de parques eólicos maiores e entre dois e três anos após o início da construção. Por fim, o efeito tende a se tornar mais heterogêneo no terceiro ano após o início das obras, o que pode indicar que o efeito delas pode estar se dissipando.

Portanto, o MARVIm – Módulo de Controle Sintético se mostrou uma metodologia inovadora e eficaz para realizar análise de impacto de intervenções com micronumerosidade de unidades tratadas. Ele avalia o impacto das intervenções ao longo do tempo após o tratamento, isto é, observa uma trajetória de impacto,

e não a mera comparação entre dois pontos fixos no tempo. Portanto, é capaz de diferenciar efeitos permanentes de efeitos transitórios de intervenções públicas. Além disso, é capaz de lidar com bases de dados extensas, uma vez que está programado para selecionar automaticamente as covariadas mais relevantes para cada variável de interesse escolhida pelo pesquisador. Por fim, o exercício de aplicação do MARVIm no presente trabalho foi feito para unidades municipais, mas pode ser estendido para quaisquer outros tipos de unidades, desde que sejam respeitadas as exigências do modelo quanto a insumos básicos.

Uma agenda futura para a continuidade do desenvolvimento dessa ferramenta de análise passa pela questão da inferência estatística das estimativas de controle sintético. Isto é, é preciso aprimorar as medidas de quanto os resultados encontrados são confiáveis, e como elas podem ser utilizadas para testes de hipóteses. Nesse sentido, os testes de placebo são um bom passo inicial, mas novos testes merecem ser desenvolvidos para assegurar a captura da causalidade dos efeitos do tratamento sobre as unidades a ele expostas.

Referências

ABADIE, A.; GARDEAZABAL, J. The economic costs of conflict: a case study of the Basque Country. *The American Economic Review*, [S.l.], v. 93, n. 1, p. 113-132, 2003.

ABADIE, A., DIAMOND, A.; HAINMUELLER, J. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, [S.l.], v. 105, n. 490, p. 493-505, 2010.

_____. Synth: an R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, [S.l.], v. 42, n. 13, jun. 2011.

_____. Comparative politics and the synthetic control method. *American Journal of Political Science*, [S.l.], v. 59, n. 2, p. 495-510, 2015.

AIE – AGÊNCIA INTERNACIONAL DE ENERGIA. *Renewable energy working party: renewable energy into the mainstream,*. Sittard, Oct., 2002.

ALBUQUERQUE, B. *et al*. *Uma solução automatizada para avaliações quantitativas de impacto:* primeiros resultados do MARVIm. 2017. Disponível em: <https://www.bndes.gov.br/wps/portal/site/home/transparencia/resultados-para-a-sociedade/Estudos-sobre-a-efetividade-do-BNDES>.
Acesso em: 20 ago. 2018.

AMARANTE, O.; ZACK, M.; SÁ, A. *Atlas do potencial eólico brasileiro*. Rio de Janeiro: Centro de Referência para Energia Solar e Eólica Sérgio de Salvo Brito (Cresesb); Centro de Pesquisas de Energia Elétrica (Cepel), 2001.

ASSUNÇÃO, J.; COSTA, F.; SZERMAN, D. *Efeitos locais de hidrelétricas no Brasil* – Climate policy initiative. Dezembro de 2016. Disponível em: <https://www.inputbrasil.org/wp-content/uploads/2017/01/CPI__Estudo_Efeitos-Locais-de-hidreletricas_no_Brasil.pdf>.Acesso em:

COSTA, R. C.; PRATES, C. P. T. O papel das fontes renováveis de energia no desenvolvimento do setor energético e barreiras à sua penetração no mercado. *BNDES Setorial*, Rio de Janeiro, n. 21, p. 5-30, mar. 2005.

COSTA, R. A.; CASOTTI, P. C.; AZEVEDO, R. L. S. – Um panorama da indústria de bens de capital relacionados à energia eólica. *BNDES Setorial*, Rio de Janeiro, n. 29, p. 229-278, mar. 2009.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, [S.l.] v. 39, n. 1, pp. 1-38. 1977.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Regiões de influência das cidades*. Rio de Janeiro, 2007.

_____. *Divisão urbano regional*. Rio de Janeiro, 2013.

KONZEN, E. *Penalizações tipo Lasso na seleção de covariáveis em séries temporais*. Dissertação (Mestrado em Economia) – Faculdade de Ciências Econômicas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

LAGE, E. S.; PROCESSI, L. D. Panorama do setor de energia eólica. *Revista do BNDES*, Rio de Janeiro, v. 39, jun. 2013.

LEFFONDREE, K. *et al*. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *Journal of Clinical Epidemiology*, [S.l.], v. 57, p. 1.049-1.062, 2004.

RESENDE, B. C. M. *O Efeito da implantação de usinas e*ólicas sobre o *preço dos aluguéis*. Trabalho de conclusão de curso (Graduação em Economia) – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2015.

RINTZEL, L. T.; ALVES, T. W.; MASSUQUETTI, A. Análise dos impactos econômicos decorrentes da instalação dos parques eólicos nos municípios da região Sul do Brasil. In: XX ENCONTRO DE ECONOMIA DA REGIÃO SUL, Porto Alegre, 2017. *Anais...,* 2017.

RODRIGUES, T. P.; GONÇALVES, S. L.; CHAGAS, A. L. S. *Usinas e*ólicas e o *mercado de trabalho nos municípios do nordeste brasileiro*. XLIV Encontro Nacional de Economia da Anpec – Associação Nacional dos Centros de Pós-Graduação em Economia, Foz do Iguaçu. *Anais*... 2016.

SIMAS, M. S. *Energia e*ólica e *desenvolvimento sustentável no Brasil:* estimativa da geração de empregos por meio de uma matriz insumo-produto ampliada. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Energia da Universidade de São Paulo, Universidade de São Paulo, São Paulo, 2012.

SUNDBERG, R. Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics*, [S.l.], v. 1, n. 2, pp. 49-58. 1974.

_____. An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in statistics – simulation and computation*. [S.l.], v. 5, n. 1, p. 55–64. 1976.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), London, v. 58, p. 267-288, 1996.

TOLMASQUIM, M. T. *Energia renovável:* hidráulica, biomassa, eólica, solar, oceânica. Rio de Janeiro: Empresa de Pesquisa Energética (EPE), 2016.

ZHAO, P.; YU, B. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, [S.l.], v. 7, p. 2.541-2.563, 2016.

ZOU, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, [S.l.], v. 101, p. 1.418-1.429, 2006.

Apêndice: Resultados das estimações individuais de controle sintético

Tabela 1. Resultados das estimações individuais de controle sintético (%)

Município	t-3	t-2	t-1	t	t+1	t+2	t+3
Acaraú – CE	2,47	2,56	(4,56)	(11,64)	(26,41)	(21,66)	(16,17)
Amontada – CE	3,31	(8,08)	1,06	3,22	4,91	12,97	(1,00)
Aracati – CE	3,13	0,85	(0,46)	(14,29)	8,47	1,28	7,27
Areia Branca – RN	(18,71)	(6,49)	(1,25)	3,33	(8,94)	(15,01)	(46,65)
Barra dos Coqueiros – SE	(1,00)	(11,08)	(9,52)	(7,93)	(18,65)	(9,92)	(23,45)
Beberibe – CE	0,16	0,04	0,08	16,33	13,40	11,78	19,99
Boituva - SP	0,00	0,00	0,00	8,97	7,05	15,67	(0,36)
Brotas de Macaúbas – BA	(4,11)	9,12	(4,50)	(1,39)	6,52	(6,88)	(8,31)
Cabo de Santo Agostinho – PE	0,27	(0,18)	0,14	6,12	13,02	11,22	14,47
Caetité – BA	0,12	0,03	0,13	11,22	0,69	1,53	28,10
Camocim – CE	1,11	(10,03)	(15,61)	(5,54)	1,60	4,72	1,39
Cururupu – MA	(2,76)	0,14	(0,74)	(1,56)	0,66	15,85	15,32
Galinhos – RN	20,62	14,57	(26,94)	(41,14)	(45,41)	(40,33)	(33,31)
Guamaré – RN	5,50	2,64	(5,53)	6,33	(61,82)	(102,59)	(99,49)
Guanambi – BA	0,49	(5,76)	(3,56)	4,77	4,44	9,87	10,08
Igaporã – BA	(6,09)	3,69	1,38	5,45	14,63	11,19	129,35
Itarema – CE	(4,68)	(3,16)	(0,68)	(3,61)	19,45	8,67	8,13
Iturama – MG	3,36	(8,16)	6,21	(5,83)	(10,39)	(16,09)	(24,89)
João Câmara – RN	(3,12)	2,38	1,70	2,94	23,60	62,32	62,45
Macaparana – PE	0,19	(1,19)	0,50	2,02	(4,14)	(4,17)	(5,64)
Mataraca – PB	(5,91)	(6,37)	(3,63)	(24,47)	(37,05)	(48,25)	(22,90)
Palmares do Sul – RS	(7,38)	(0,14)	(1,16)	13,04	(1,32)	8,31	15,06
Paracuru – CE	(0,57)	0,32	(0,14)	0,54	18,35	4,94	16,64
Parazinho – RN	2,71	0,81	(5,40)	273,48	1046,45	484,52	475,68
Parnaíba – PI	0,72	12,60	(1,59)	(11,11)	(7,04)	(2,36)	(4,69)
Pedra Grande – RN	(1,19)	4,34	1,52	32,71	36,85	16,81	344,73
Pelotas – RS	9,70	6,93	(3,14)	0,46	(0,46)	#N/D	#N/D
Pombos – PE	0,00	0,00	0,00	5,29	16,52	13,60	5,44
Sant'Ana do Livramento – RS	3,89	(2,07)	1,58	3,19	1,22	14,44	13,06
São Francisco de Itabapoana – RJ	(8,95)	(2,03)	4,83	14,62	29,70	105,47	145,20
São Miguel do Gostoso – RN	(0,78)	(2,42)	2,03	2,48	16,90	93,33	#N/D
Sento Sé – BA	0,77	1,46	(2,78)	5,55	(6,55)	19,26	19,67
Sobradinho – BA	(7,18)	(3,48)	(14,66)	(2,03)	(48,41)	(55,49)	(44,22)
Trairi – CE	(10,33)	(1,85)	1,59	8,43	41,49	42,88	32,04
Tramandaí – RS	0,77	1,33	(2,15)	0,68	1,21	2,05	3,23
Tubarão – SC	0,00	0,00	0,00	7,62	(1,81)	#N/D	#N/D
Xangri-lá – RS	(2,38)	1,52	1,32	10,29	25,02	#N/D	#N/D

Fonte: Elaboração própria.

Coordenação Editorial

Gerência de Editoração e Memória do BNDES

Projeto Gráfico

Fernanda Costa e Silva

Produção Editorial

Expressão Editorial

Editoração Eletrônica

Expressão Editorial

Editado pelo
Departamento de Comunicação
Outubro de 2018

BNDES

www.bndes.gov.br