IMPACTO DE UNIDADES INTERLIGADAS NO SUB-REGISTRO CIVIL DE NASCIMENTO - UM ESTUDO PRELIMINAR

BRASÍLIA, 2023

Apoio:

FLACSO BRASIL

PNUD Empoderando vidas. Fortalecendo nações.

Realização:

MINISTÉRIO DOS DIREITOS HUMANOS E DA CIDADANIA

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

IMPACTO DE UNIDADES INTERLIGADAS NO SUB-REGISTRO CIVIL DE NASCIMENTO – UM ESTUDO PRELIMINAR

(Brasília, 2023)

Apoio:

Faculdade Latino-Americana de Ciências Sociais

Programa das Nações Unidas para o Desenvolvimento

Realização:

Coordenação-Geral de Promoção de Registro Civil de Nascimento

Diretoria de Promoção dos Direitos Humanos

Secretaria Nacional de Promoção e Defesa dos Direitos Humanos

Ministério dos Direitos Humanos e da Cidadania

Presidente da República
Luiz Inácio Lula da Silva

Vice-Presidente
Geraldo José Rodrigues Alckmin Filho

Ministério dos Direitos Humanos e da Cidadania
Silvio Luiz de Almeida

Secretaria Nacional de Promoção e Defesa dos Direitos Humanos
Isadora Brandão Araujo da Silva

Diretoria de Promoção dos Direitos Humanos
Alex André Vargem

Coordenação-Geral de Promoção do Registro Civil do Nascimento
Tula Vieira Brasileiro

IMPACTO DE UNIDADES INTERLIGADAS NO SUB-REGISTRO CIVIL DE NASCIMENTO – UM ESTUDO PRELIMINAR

Esta publicação foi organizada pela Faculdade Latino-Americana de Ciências Sociais – FLACSO Brasil. A edição desta obra foi viabilizada por meio do projeto "Apoio Técnico às Ações do Compromisso Nacional pela Erradicação do Sub-Registro Civil de Nascimento e Ampliação da Documentação Básica", realizado no âmbito da parceria estabelecida entre Flacso Brasil, Ministério dos Direitos Humanos e da Cidadania - MDHC e Programa das Nações Unidas para o Desenvolvimento – PNUD. Sua distribuição eletrônica ou impressa é gratuita.

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Garcez, Lucas Nogueira

Impacto de unidades interligadas no sub-registro civil de nascimento [livro eletrônico] : um estudo preliminar / Lucas Nogueira Garcez. -- 1. ed. -- Brasília, DF : Faculdade Latino-Americana de Ciências Sociais, 2022. -- (Coleção políticas de registro civil de nascimento e documentação básica) PDF.

Bibliografia.
ISBN 978-65-87718-38-5

1. Acesso à informação 2. Cartórios - Administração 3. Direito notarial e registral – Brasil 4. Direito notarial – Leis e legislação – Brasil 5. Maternidade 6. Registro Civil das Pessoas Naturais I. Título II. Série.

22-138364 CDD-360

Índices para catálogo sistemático:

1. Registro Civil de Nascimento : Serviço Social 360

Aline Graziele Benitez – Bibliotecária – CRB-1/3129

Faculdade Latino-Americana de Ciências Sociais – Flacso Brasil

Diretora
Rita Gomes do Nascimento

Coordenadora do Programa Cidadania, Participação Social e Políticas Públicas
Kathia Dudyk

Projeto "Apoio Técnico às Ações do Compromisso Nacional pela Erradicação do Sub-Registro Civil de Nascimento e Ampliação da Documentação Básica"

Coordenadora-Geral
Kathia Dudyk

Coordenação Executiva
Carolina Albuquerque Silva

Equipe
Aline Quintão de Araujo, Bárbara Alves Nonato, Fábio Merladet, Juliana Nascimento Lima, Márcia de Câmera Campos

Ficha Técnica

Autor
Lucas Nogueira Garcez

Edição
Carolina Albuquerque Silva

Projeto Gráfico e Diagramação
Vitor Reis Soares

SUMÁRIO

1. INTRODUÇÃO

Esta pesquisa foi desenvolvida por consultoria técnica especializada, ao longo do segundo semestre de 2020, no âmbito do Projeto "Apoio técnico às ações do compromisso nacional pela erradicação do sub-registro civil de nascimento e ampliação da documentação básica", realizado pela Sede Acadêmica Brasil da Faculdade Latino-Americana de Ciência Sociais (FLACSO), em parceria com a Secretaria Nacional de Promoção e Defesa dos Direitos Humanos (SNDH) do Ministério dos Direitos Humanos e da Cidadania (MDHC), e o Programa das Nações Unidas para o Desenvolvimento (PNUD).

A pesquisa analisou a relação entre unidades interligadas e sub-registro civil de nascimento. O termo "sub-registro civil de nascimento" pode compreender tanto (1) nascimentos que não são de qualquer maneira documentados, i.e., nem pelo sistema de saúde nem por cartórios[1], quanto (2) nascimentos que geram registro no sistema de saúde quando do nascimento, por meio da Declaração de Nascido Vivo (DNV), não havendo posterior emissão de uma certidão em cartório. Para os fins desse estudo adotamos como definição do sub-registro civil de nascimento o segundo caso. Ademais, definimos como unidade interligada um posto dedicado ao registro civil de nascimento dentro de hospital, maternidade ou estabelecimento de saúde que realiza parto, definido como unidade interligada pelo Estado. Como discutiremos adiante, há outras políticas semelhantes não classificadas como unidades interligadas. Essa imprecisão da definição foi incorporada nas análises.

O presente estudo tem por objetivos: (1) a partir do pareamento das bases do Sistema Nacional de Informações de Registro Civil (SIRC) e do Sistema de Informação sobre Nascidos Vivos (SINASC), disponibilizadas pelo MDHC, produzir medidas do sub-registro civil de nascimento e (2) usando dados de diversas fontes, avaliar o impacto da política de criação de unidades interligadas por meio de inferência causal. Trata-se de estudo preliminar, que necessita ser aprofundado a fim de fornecer nexos causais mais robustos entre a existência de unidades interligadas e a diminuição de sub-

[1] Drumond, Machado & França, *Underreporting of Live Births: Measurement Procedures Using the Hospital Information System, in* **Sistemas de Estatísticas Vitais: Avanços, Perspectivas e Desafios**, Revista de Saúde Pública, 2008.

registros. Em que pesem suas limitações, os resultados encontrados por meio da presente pesquisa corroboram esta hipótese, como será apresentado mais adiante.

Um dos objetivos da pesquisa é superar o viés de seleção na análise. Isto é, regiões que recebem unidades interligadas não são escolhidas aleatoriamente, mas por diversas características, inclusive algumas não observadas nos dados. Dessa forma, diferenças de sub-registro entre essas regiões podem refletir tais características, e não o real impacto das unidades interligadas. Dessa forma, o objetivo é separar o efeito causal, evitando correlações espúrias.

Um aspecto central desse processo é a escolha de variáveis de controle adequadas. A hipótese teórica chave que norteia tal escolha é a de que desigualdades sociais, econômicas e regionais, bem como fragilidade institucional e ausência do Estado, sejam o principal mecanismo causador do sub-registro civil de nascimento. Como descreveremos em mais detalhes, nosso foco na desigualdade justifica-se pelo fato de que ela afeta os custos e os benefícios do registro. Por exemplo, ela afeta a distância e o tempo de transporte que separam a residência da criança e o cartório, um custo inerente à escolha de registrar. Além disso, o benefício do registro é, por exemplo, claro para pais que pretendem matricular seus filhos numa escola, dado que isso não é possível sem a certidão de nascimento. Para uma família em local isolado sem qualquer perspectiva de acesso a saúde ou educação públicas e que deseja que as crianças tão somente ajudem no trabalho agrícola, o registro implica menos benefícios. A desigualdade também afeta o acesso à educação e informação, o que por sua vez influencia o conhecimento da necessidade e dos benefícios do registro.

Por fim, subjacente ao trabalho está a ideia de que o acesso ao sub-registro civil de nascimento e a identificação das políticas mais eficazes para tanto é essencial do ponto de vista social. Sem o registro, uma criança não existe perante o sistema jurídico, sendo privada de proteção e do acesso a serviços públicos. Além disso, sem o registro, o Estado não tem informações precisas sobre a população, o que dificulta formular e dimensionar políticas públicas adequadas e proteger os direitos fundamentais de todos.

2. LITERATURA APLICÁVEL

A estimativa de sub-registro na América Latina é de 10%[2]. A literatura[3] aponta uma relação inversa entre a distância até o cartório e a probabilidade de registro de um recém-nascido, de 4 a 12 pontos percentuais para cada 25km. Esse tipo de estudo se vale de dados individualizados do censo e a localização dos cartórios, que pode ser acessada via *Google Maps* e bases oficiais. Além dos custos da distância, custos e taxas administrativas também podem ser um fator importante. A literatura aponta que, no caso da Indonésia[4], o sub-registro civil de nascimento é impulsionado por custos impostos pelos cartórios, bem como a necessidade de apresentação da certidão de casamento dos pais, o que exclui do sistema crianças cujos pais não se casaram.

O problema do sub-registro é comum em países em desenvolvimento, de forma que estudos sobre o assunto na Europa e nos EUA são escassos. Ocorre, contudo, que um fenômeno análogo acontece em virtude da imigração ilegal. Por seu status, imigrantes ilegais não podem se apresentar as autoridades para registrar seus filhos. Nesse caso, os efeitos dessa modalidade de sub-registro são os mesmos observados nos países em desenvolvimento[5].

Como descrevemos acima, o registro é essencial para que o Estado possa garantir direitos fundamentais. O uso de trabalho infantil, por exemplo, é muito mais difícil de ser detectado se as vítimas não existem perante o ordenamento jurídico. As certidões de nascimento também são provas concretas da idade da criança. Dessa forma, o registro permite ao Estado tanto detectar se a criança não está matriculada na escola e, portanto, possivelmente trabalhando, bem como provar sua idade em um eventual processo. Esse efeito foi constatado pela literatura usando dados dos EUA na primeira metade do século XX[6]. Isso porque nessa época foram criadas as leis de registro civil dos Estados Unidos. As mesmas conclusões foram alcançadas usando dados

[2] UNICEF, **The State of the World's Children 2010: Children Survival**, 2010.
[3] Corbacho & Rivas, *Travelling the Distance: A GPS-Based Study of the Access to Birth Registration Services in Latin America and the Caribbean* in **Inter-American Development Bank Working Papers**, 2012.
[4] Duff, Kusumaningrum & Stark, *Barriers to Birth Registration in Indonesia* in **The Lancet**, 2016.
[5] Benuto, Casas, Gonzales & Newlands, *Being an Undocumented Child Immigrant* in **Children and Youth Services**, 2018.
[6] Fagernas, *Protection through Proof of Age: Birth Registration and Child Labor in Early 20th Century USA* in **University of Sussex Economics Department Working Paper Series**, 2011.

contemporâneos da Índia[7]. No caso particular da Índia, dada a grande população do país, os principais fatores que causam o sub-registro foram ausência do Estado ou órgãos de registro em determinados locais, bem como a falta de acesso à informação sobre a importância do registro.

O Banco Interamericano de Desenvolvimento realizou estudos comparativos da América Latina que incluíam o Brasil na análise[8]. Segundo o estudo, as variáveis que explicam o sub-registro são o sexo do recém-nascido, a educação e a renda da mãe, ausência de consultas de pré-natal e a educação do pai. O estudo aponta impactos geracionais da ausência do registro na América Latina, indicado que a chance de sub-registro aumenta quando o pai ou a mãe também não são registrados[9]. Outra consequência observada pelo estudo foi a ausência de vacinação em crianças que não foram registradas[10].

Essa literatura oferece uma base interessante, embora nem todas as conclusões sejam aplicáveis. A questão dos custos administrativos, por exemplo, não se aplica totalmente ao Brasil, que garantiu com a Lei n° 9.534/97 o registro gratuito. Além disso, esses estudos não analisaram o impacto das políticas introduzidas pelas **Leis n° 13.257/16 e n° 12.662/12, que introduziram as unidades interligada**s.

[7] George & Jaymon, *Constraints in Birth Registration: A Case Study in Andhra Pradesh* in **Civil Registration and Vital Statistics**, United Nations, 2015.
[8] Duryea, Olgiati & Stone, *The Under-Registration of Births in Latin America* in **Inter-American Development Bank Working Papers**, 2006.
[9] Rud & Castro, *Medición Cuantitativa Del Subregistro de Nacimientos e Indocumentación, Costos Socioeconómicos en Perú y República Dominicana* in **Inter-American Development Bank Working Papers**, 2011.
[10] Brito, Corbacho, Osorio, *Does Birth Under-Registration Reduce Childhood Immunization? Evidence from the Dominican Republic* in **Health Economics Review**, 2017.

3. MODELOS DE ANÁLISE ESTATÍSTICA

Os estudos que citamos acima variam em termos das fontes de dados. Contudo, todos usam modelos de escolha discreta ou modelos de dados em painel com efeitos fixos. Nessa seção revisitaremos os modelos propostos ao longo da pesquisa, discutiremos quais foram os modelos escolhidos como potencialmente viáveis e quais pudemos efetivamente implementar.

Discutiremos, em primeiro lugar, os modelos de escolha discreta. Nesse contexto a escolha discreta é também binária, i.e., registrar ou não registrar. A alternativa mais simples para modelar esse tipo de situação seria um modelo linear. Nesse caso, a probabilidade de uma escolha, por exemplo a de registrar uma criança, seria uma combinação linear das características da pessoa. Nesse caso os parâmetros podem ser interpretados como pesos ou como indicativos da forma pela qual cada característica influencia a probabilidade de escolha. Nessa estrutura de regressão linear teríamos, então, o Modelo Linear de Probabilidade, LPM, onde:

$$p(y = 1) = x_1\beta_1 + x_2\beta_2 + \cdots x_n\beta_n + \varepsilon \rightarrow \frac{\partial p}{\partial x_1} = \beta_1$$

Ocorre que em modelos lineares não podemos garantir que as probabilidades estarão entre zero e um. Dessa forma, para modelar escolhas discretas, usamos modelos não lineares. Esses modelos partem da probabilidade da escolha "y_i" de um indivíduo ocorrer:

$$p(y_i = 1|x_i) = p(x_i) = F(x_i) \rightarrow p(y_i = 0|x_i) = \big(1 - F(x_i)\big)$$

Assim, a probabilidade da escolha de um indivíduo "i" acontecer para indivíduos com as mesmas característica é, em geral:

$$p(x_i) = [F(x_i)]^{y_i}[1 - F(x_i)]^{1-y_i}$$

Assim, podemos encontrar uma função de verossimilhança para uma amostra. Essa função nos diz, basicamente, a probabilidade de encontrarmos o mesmo resultado em outra amostra com as mesmas características. A ideia é buscarmos o modelo sob o qual a probabilidade das escolhas observadas ocorrerem é máxima. A função é descrita por:

$$L = \prod_{i=1}^{n} [F(x_i)]^{y_i} [F(x_i)]^{1-y_i}$$

Aplicando logaritmos, tornamos a função mais simples e, como se trata de uma transformação monotônica, não alteramos a natureza da otimização:

$$log(L) = l = \sum_{i=1}^{n} y_i \times log\big(F(x_i)\big) + (1 - y_i) \times log\big(1 - F(x_i)\big)$$

Logo, podemos encontrar o vetor de parâmetros **"$\hat{\beta}$",** isto é, os "pesos" que refletem como cada característica individual reflete na escolha, tal que[11]:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i}{F(x_i)} f(x_i) x_i - \frac{(1 - y_i)}{1 - F(x_i)} f(x_i) x_i$$

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - F(x_i)}{F(x_i)\big(1 - F(x_i)\big)} f(x_i) x_i = 0$$

Onde "$F(.)$" é a distribuição logística, "$\Lambda(.)$", no caso do modelo Logit e a distribuição normal, "$\Phi(.)$" no caso do modelo Probit. Como aqui não se trata de modelos lineares, não se pode interpretar parâmetros como o efeito diretamente de cada característica sobre a probabilidade, como fazemos nos modelos lineares[12]. Assim, as probabilidades nos modelos Logit e Probit são[13]:

[11] Verbeek, **A Guide to Modern Econometrics**, 2004, p. 193.
[12] Cameron & Trivedi, **Microeconometrics**, 2005, p. 470.
[13] Greene, **Econometric Analysis**, 2018, p. 732

$$P(y_i = 1|x_i) = \int_{-\infty}^{x_i'\beta} \phi(t)dt$$

$$P(y_i = 1|x_i) = \frac{exp(x_i'\beta)}{1 - exp(x_i'\beta)}$$

Na literatura que discutimos anteriormente, há estudos que usam como características para modelar a escolha do registro o uso de computadores por cartórios, distância até o cartório, idade e educação da mãe, se a mãe coabita ou não com um cônjuge, sexo da criança, chuvas na época do nascimento, entre outras. Suponha que a variável de interesse do estudo seja distância, como a presença de unidade interligada em nosso contexto. Nesse caso o modelo Probit é:

$$P(Sem\ Registro) = \Phi\left(\beta_0 + Distancia_{ij}\beta_1 + X_{ij}\beta_2 + Y_j\right)$$

Onde:

$$\Phi(\gamma) = \int_{-\infty}^{\gamma} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right) dt$$

Outro modelo muito usado na literatura é dos dados em painel com efeitos fixos. Quando descrevemos um conjunto de dados como um painel, queremos dizer que esses dados contêm as mesmas unidades observadas ao longo do tempo. Por exemplo, uma base que contém o PIB de cada município nos últimos três anos, sendo cada linha um município em determinado ano, está em uma estrutura de painel. Essa estrutura permite resolver um dos principais desafios da inferência causal[14]. Trata-se das variáveis não observadas e relacionadas com o tratamento a variável dependente. Tais variáveis podem ser fatores fixos e intrínsecos[15] a uma unidade de observação como individuo ou município. O mais interessante é que esses efeitos podem ser constantes no tempo[16]. Suponhamos que a expressão não observável de uma variável

[14] Imbens & Wooldridge, *Recent Developments in the Econometrics of Program Evaluation* in **NBER Working Papers**, 2008.
[15] Heiji, de Boer et. al. **Econometric Methods with Applications in Business and Economics**, 2004, p. 692
[16] Por exemplo, duas cidades distintas que, apesar de mudanças econômicas, sempre terão culturas diferentes ou duas pessoas que, independentemente da educação a que tem acesso, sempre terão diferentes níveis de aptidão para uma determinada tarefa.

para uma unidade "i" no período "t" é:

$$y_{ti} = x'_{it}\beta + \varepsilon_{it} + c_i$$

Essa expressão diz que a variável dependente pode ser expressa como uma função linear de características observáveis, um erro, que descreve a variabilidade aleatória dos dados e "c_i", a caraterística intrínseca, idiossincrática e constante da unidade "i". Em geral, estamos interessados no vetor de coeficientes "β", que nos diz como cada característica, ou variável independente, afeta a variável dependente "y_{ti}". Podemos tirar a média dessa expressão ao longo do tempo:

$$\bar{y}_i = \bar{x}'_i\beta + \bar{\varepsilon}_i + c_i$$

Se subtrairmos essa média de cada observação ao longo do tempo, obtemos[17]:

$$(y_{it} - \bar{y}_i) = (x'_{it} - \bar{x}'_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i) + (c_i - c_i)$$

Assim, subtraindo a média, os efeitos fixos desapareceram e conseguimos estimar os mesmos parâmetros de interesse sem qualquer viés[18], mas desta vez modelando os desvios da média. Um resultado semelhante poderia ser obtido por meio da primeira diferença[19]. Outra forma de lidar com esse problema seria usando os chamados efeitos aleatórios. Nesse caso, assumiríamos que "c_i" é aleatório e independente "x'_{it}", de forma que seria capturado pelo termo do erro. Ocorre que esse pressuposto em geral não é verdade[20]. Suponha que fatores geográficos de um município façam com que ele consistentemente tenha mais sub-registro. Por exemplo, se a população mora às margens do rio Amazonas, em um local que o deslocamento é difícil e feito exclusivamente de barco. Nesse caso, mesmo que ofereçamos a esse município os mesmos recursos a que outros municípios têm acesso, sua performance em termos de sub-registro ainda assim será pior. Ocorre que esse fator idiossincrático do município não é independente das demais características. Provavelmente a característica geográfica influenciaria outras variáveis como indicadores de pobreza e saneamento básico. Ou seja, é difícil afirmar com segurança que predisposições locais ou individuais em relação ao sub-registro não afetam

[17] Heiji, de Boer et. al. **Econometric Methods with Applications in Business and Economics**, 2004, p. 694
[18] Angrist & Pischke. **Mostly Harmless Econometrics**, 2009, p. 223.
[19] Angrist & Pischke. **Mostly Harmless Econometrics**, 2009, p. 224.
[20] Verbeek, **A Guide to Modern Econometrics**, 2004, p. 348.

também outras variáveis de controle, como educação, renda etc. De qualquer forma, isso pode ser testado estatisticamente e, havendo evidências de que o pressuposto é plausível, podemos usar efeitos aleatórios.

Uma consequência indesejada dessa solução, i.e., trabalhar com desvios das médias para lidar com os mesmos coeficientes, mas sem efeitos fixos, é que a análise acaba por se restringir a variações intraunidades, aumentando a variância dos erros[21]. Mais importante ainda é o fato de que não conseguimos encontrar coeficientes para variáveis estáveis ao longo do tempo. Por exemplo, a localização de um município não muda ao longo do tempo. No exemplo que descrevemos acima, poderíamos querer adicionar uma variável identificando municípios próximos a rios para avaliar esse impacto. Ocorre que, assim como qualquer efeito fixo não observado, variáveis fixas observadas desaparecem e não conseguimos estimar seu impacto, apenas nos livrar dele para estimar o impacto das demais sem viés[22]. No nosso caso, poderíamos fazer uma regressão da taxa municipal de sub-registros em controles e no número de unidades interligadas em cada município. Assim, removeríamos os efeitos fixos de cada município.

Outro método que pode ser empregado é o da regressão descontinua ou RDD. Esse método não foi explorado na literatura. Isso ocorre porque esses estudos foram gerais e não se debruçaram sobre cortes ou limites regulatórios. Isto é, o RDD é aplicável quando podemos comparar observações um pouco acima ou um pouco abaixo de um corte, em termos de uma variável[23]. Esse tipo de situação é comum com limites regulatórios, como notas de corte em educação ou faixas de renda em questões tributárias. O interessante dessas abordagens é que observações perto de um corte tendem a ser muito parecidas em termos das suas características não observáveis. Por exemplo, alunos que obtiveram um ponto a mais ou um ponto a menos que a nota de corte para uma bolsa, provavelmente estudaram com a mesma intensidade, diferindo tão somente em termos de receber ou não a bolsa. Restringir a análise estatística a esses alunos ajuda a tornar o estudo dos impactos da bolsa mais precisos, uma vez que, nesse subconjunto, ter ou não a bolsa é aleatório[24].

---

[21] Allison, **Fixed Effects Regression Methods for Longitudinal Data Using SAS**, 2005, p. 4.

[22] Donald & Lang, *Inference With Difference in Differences and Other Panel Data* in **The Review of Economics and Statistics**, 2007.

[23] Cook, *Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics* in **Journal of Econometrics**, 2008.

[24] Por exemplo, alunos que fizeram um ponto a mais ou um ponto a menos de uma nota de corte, empresas na fronteira de estados com regulações diferentes, famílias com uma renda um pouco acima ou um pouco abaixo da renda máxima para participação em um programa social etc.

Ou seja, restringir a amostra dessa forma seria o mesmo que atribuir a bolsa de maneira aleatória em um experimento. Vale destacar que, como esse método para descobrir o efeito de um tratamento envolve restringir a amostra, ele não pode ser generalizado, sendo de efeito local. Note também que a escolha da banda, ou seja, quantos pontos acima e quantos pontos abaixo do corte, impõe uma escolha entre precisão e viés. Uma banda maior implica incluir na amostra observações não tão semelhantes, aumentando o viés. Contudo uma banda muito pequena restringe a amostra a um número pequeno de observações, aumentando a variância. Uma forma aproximada de verificar se há viés para determinada banda é realizar um teste de balanceamento e avaliar se as observações na banda se parecem em termos das demais características. Isso não garante, contudo, que elas serão semelhantes em termos de características não observadas. Vale destacar também que esse método não exige um corte preciso. O método pode ser flexibilizado em um modelo chamado de regressão descontínua "sem nitidez", ou FRD[25]. Nesse caso, o modelo assume que pontos acima do corte aumentam, mas não determinam, a probabilidade de receber o tratamento[26]. Além dos riscos já descritos com relação a variância e ao viés, um risco adicional que temos aqui é de se interpretar relações não lineares como descontínuas. Formalmente, podemos descrever a estratégia como[27]:

$$\mu_l(x) = \lim_{z \uparrow x} E[Y(0)|X = z] \quad \mu_r(x) = \lim_{z \downarrow x} E[Y(1)|X = z]$$

E o efeito causal do tratamento como:

$$\tau_{RDD} = \mu_r(c) - \mu_l(c)$$

Se usarmos uma regressão no lugar de esperanças condicionais, teremos[28]:

$$y_{it} = \beta_0 + \beta_1 D_{it} + f(Z_{it}) + \gamma' x_{it} + u_{it}$$

Sendo a variável "D_i" o indicador de observações acima do corte, ou seja, da descontinuidade, seu coeficiente será, então, o salto na descontinuidade, ou seja, o

[25] Pischke, *Regression Discontinuity Design Notes*, **London School of Economics,** 2018, pp. 14-15.
[26] Porter, *Estimation in the Regression Discontinuity Model* in **Harvard University Working Papers**, 2003.
[27] Imbens & Lemieux, *Regression Discontinuity Designs : A Guide to Practice* in **Journal of Econometrics**, 2007.
28 Angrist & Krueger, *Empirical Strategies in Labor Economics* in **Handbook of Labor Economics,** V. 3, 1999, p. 1307.

efeito do tratamento. O termo "$f(Z_i)$", indica uma função, em geral linear[29], capturando efeitos direto da variável de corte. Voltando ao exemplo da bolsa, a variável nota impõe um corte a partir do qual o indivíduo pode receber a bolsa, mas as notas em si também afetam a variável dependente diretamente, sem ser pela bolsa.

Como discutiremos adiante, no presente caso o corte pode ser 18 horas para o horário do nascimento ou dias antes e durante um feriado. Isso porque parte das unidades interligadas deixa de funcionar após as 18h ou durante feriados, de forma que se compararmos o sub-registro em um hospital ou município, durante e antes de um feriado, estaríamos comparando grupos de nascimentos com perfil semelhante, diferindo apenas na exposição a uma unidade interligada.

O uso de um corte não é a única forma de compararmos observações semelhantes. Outra forma de buscar essa semelhança é pelo pareamento por escore de semelhança ou propensão, e.g. *Propensity Score Matching* ou *Mahalanobis Score Matching*. Basicamente, esses métodos usam características observadas para encontrar unidades com a maior semelhança possível, diferindo apenas no tratamento. Naturalmente, o cálculo desse escores de semelhança pode ser feito de maneiras diferentes. Isso permite lidar com tratamentos que não são atribuídos de maneira aleatória[30]. Comparando semelhantes, conseguimos resolver parcialmente a questão da dependência entre tratamento e variável dependente[31]. No caso do PSM, o escore é a probabilidade condicional de tratamento, estimada com um modelo discreto, em geral logístico[32], a partir das características:

$$PS = P(T_i|x_i)$$

$$P(T_i = 1|x_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} = \Lambda(x_i'\beta)$$

29 Angrist & Pischke, **Mastering Metrics: The Path From Cause to Effect**, 2014, eq. 4.2.

30 Rosenbaum & Rubin, *The Central Role of The Propensity Score in Observational Studies For Causal Effects* in **Biometrika**, 1983.

31 Rose & Laan, **Targeted Learning: Causal Inference for Observational and Experimental Data**, 2011, p. 348.

32 Greene, **Econometric Analysis**, 2018 p. 732.

A função de verossimilhança é, então[33]:

$$L = \prod_{i=1}^{n} [\Lambda(x_i'\beta)]^{y_i} [1 - \Lambda(x_i'\beta)]^{1-y_i}$$

Quando duas unidades são igualmente propensas a receber o tratamento podemos, assim como fizemos na regressão descontinua, assumir que atribuição do tratamento entre elas equivale a uma atribuição aleatória. Logo, buscamos minimizar a diferença entre propensões[34] para encontrar um par ideal[35]:

$$C^0\big(P(T_i = 1|x_i)\big) = \{j: \big|P(T_i = 1|x_i) - P\big(T_j = 1|x_j\big)\big| = min\{\big|P(T_i = 1|x_i) - P\big(T_j = 1|x_j\big)\big|\}\}$$

Outra maneira de fazer isso é usar a distância de Mahalanobis[36], que basicamente usa uma variação da distância euclidiana entre os dois vetores de características:

$$X_{MDM} = M\left(X|\sqrt{(X_i - X_j)S^{-1}(X_i - X_j)} < \delta\right)$$

Uma vez feito o pareamento, calculamos a diferença média entre as variáveis dependentes[37]:

$$\hat{ATE} \approx \frac{1}{S}\sum_{s=1}^{S}(\bar{y}_{is} - \bar{y}_{js})$$

Sendo "S" é o número de pares e "$\hat{ATE}$" o efeito do tratamento. Ao longo da pesquisa cogitamos a possibilidade de usar o algoritmo Perceptron para esse tipo de pareamento, mas vimos que o método de Mahalanobis era a alternativa mais adequada ao escore de propensão.

[33] Rosenbaum, **Observational Studies**, 2002, pp. 78, 296.
[34] Sianesi, *Implementing Propensity Score Matching Estimates With Stata* in **UK Stata Users Group Meetings**, 2001, p. 6.
[35] Chamado de *Nearest Neighbor Matching*.
[36] King & Nielsen, Why Propensity Scores Should Not Be Used for Matching, 2018, p.
[37] Gelman & Meng, **Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspective**, 2004, pp. 28 – 29.

4. MODELOS DE PAREAMENTO DE BASES DE DADOS

Com relação ao pareamento de bases de dados, usamos a biblioteca de *Python Record Linkage Toolkit*[38]. Essa biblioteca basicamente cria candidatos a pares e um vetor de similaridades a eles associado. Um problema dessa abordagem é que, para grandes números de nascimentos não pareados, o número de pares em potencial é muito grande. Para contermos o número de combinações, criamos algumas restrições. Buscamos todas as combinações entre nascimentos não pareados que: (1) ocorreram no mesmo município, (2) no mesmo ano e (3) no mesmo mês. Na nomenclatura da biblioteca essa restrição é denominada de bloqueio. O bloqueio por ano, mês e município faz com que o número de candidatos seja computacionalmente razoável, mas ele tem um custo. O bloqueio é limitante, uma vez é possível que, por exemplo, no momento do registro, o município de nascimento seja confundido com o de residência. Em um caso desse poderíamos ter uma criança que aparece nas duas bases, mas que não é pareada porque os municípios de nascimento não são o mesmo nas duas bases. Infelizmente o custo de não usar o bloqueio em termos de memória seria proibitivo. Com todos os candidatos a par, o algoritmo calcula um vetor de características usando fórmulas de similaridade, sobretudo de similaridades entre textos. São usados três métodos, Damerau-Levenshtein, Jarowinkler e Levenshtein[39]. Esses métodos dão diferentes medidas da distância entre duas palavras ou blocos de texto. Extensões deles são comumente aplicados para corrigir erros de digitação. No caso da distância de Levenshtein, por exemplo, a distância se dá pelo número de operações necessárias para transformar um bloco de texto em outro, sendo essas operações a deleção, inserção e substituição[40]. Dessa forma, treinamos os algoritmos com pares de observações pareadas pela DNV. Assim, observamos como se comportam os índices de similaridade quando as observações em cada base descrevem a mesma pessoa. Usando esse comportamento, o algoritmo classifica observações não pareadas.

---

[38] https://recordlinkage.readthedocs.io/en/latest/about.html

[39] *Levenshtein distance*, Algorithms and Theory of Computation Handbook, in **Dictionary of Algorithms and Data Structures**, 1999.

[40] O algoritmo usa programação dinâmica para encontrar o número mínimo de passos necessários para a transformação. O algoritmo constrói uma matriz que descreve subconjuntos dos textos em questão. A partir dela,9 ele encontra o conjunto mínimo de passos. Chamemos a matriz de "M". Preenchimento do elemento "$M_{a,b}(i,j)$", que corresponde a transformação do subtexto terminado em "i" no subtexto terminado em "j" da seguinte forma: $min\left\{M_{a,b}(i-1,j)+1, M_{a,b}(i,j-1)+1, M_{a,b}(i-1,j-1)+1_{(a_i \neq a_j)}\right\}$. Basicamente, esses passos correspondem aos passos que antecedem uma deleção, uma inserção e uma substituição.

O método de aprendizagem e classificação dos pares pode ser qualquer método de aprendizagem de máquina. No plano de trabalho propusemos usar Gaussian Naive Bayes para o aprendizado do peso relativo de cada índice similaridade. Ocorre que, após teste, encontramos melhores resultados usando Perceptron. Seguimos com uma breve discussão de cada modelo.

As diversas versões do algoritmo Perceptron são as aplicações básicas de redes neurais. Trata-se de um modelo linear da seguinte forma[41]:

$$f(x) = \langle w \cdot x \rangle + b$$

Que pode ser expresso como:

$$f(x) = \theta \cdot x + \theta_0 = \sum_{i=1}^{n} \theta_i x_i + \theta_0$$

Assim como em outros modelos, definimos uma perda a ser minimizada. Definimos a classificação observada no conjunto de dados para treino de "$y^{(i)}$". A perda é definida como:

$$\begin{cases} y^{(i)} f\left(x^{(i)}\right) > 0 \rightarrow loss = 0 \\ y^{(i)} f\left(x^{(i)}\right) \leq 0 \rightarrow loss = 1 \end{cases}$$

Seguimos então buscando parâmetros "θ, θ_0" que minimizem a perda, i.e., os erros na classificação:

$$D(\theta, \theta_0) = -\sum y^{(i)}\left(x^{(i)} \cdot \theta + \theta_0\right)$$

A obtenção dos parâmetros é feita computacionalmente com o algoritmo de gradiente estocástico. O processo se resume no seguinte[42]: Inicialmente, os parâmetros recebem zero. Itera-se, então sobre o conjunto de dados de treino. O algoritmo calcula em cada etapa a perda "$loss\left(\theta, \theta_0, x^{(i)}, y^{(i)}\right)$". Havendo uma perda diferente de zero, os

[41] Cristianini & Shawe-Taylor, **An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods**, 2000, 2.1.
[42] Hastie, Tibshirani & Friedman, **The Elements of Statistical Learning**, 2009, p. 150.

parâmetros são melhorados:

$$\begin{pmatrix} \theta \\ \theta_0 \end{pmatrix} + \begin{pmatrix} y^{(i)}x^{(i)} \\ y^{(i)} \end{pmatrix} \rightarrow \begin{pmatrix} \theta \\ \theta_0 \end{pmatrix}$$

Esse processo é feito para todos os dados num ciclo. Repetimos o ciclo "T" vezes. Após esses ciclos, calcula-se a média dos valores dos parâmetros em cada ciclo:

$$\theta_{final} = \frac{1}{nT}\left(\theta^{(1)} + \theta^{(2)} + \cdots + \theta^{(nT)}\right)$$

Uma vez que os valores definitivos dos parâmetros são calculados, podemos classificar os dados de teste e, havendo uma performance adequada, os dados não pareados:

$$\hat{y}^{(i)} = sign\left(f\left(x^{(i)}\right)\right)$$

O algoritmo Gaussian Naive Bayes[43] foi utilizado em testes, mas foi substituído pelo Percpetron por ter resultados menos satisfatórios nesse contexto. De fato, ele tem pressupostos muito restritivos, i.e., independência entre características, mas em geral produz bons resultados[44]. O algoritmo funciona, basicamente, por meio de probabilidades condicionais e o teorema de Bayes. Com o pressuposto de independência e o cálculo das probabilidades de cada característica condicional a variável dependente. Assim, conseguimos, por meio do teorema de Bayes, encontrar a probabilidade de um resultado, por exemplo de um par ser correto ou não, dado um vetor de características, usando as probabilidades condicionais:

$$p(y|x_1, \ldots, x_n) = \frac{p(y)\prod_{i=1}^{n} P(x_i|y)}{p(x_1, \ldots, x_n)}$$

Assim, o algoritmo busca a classificação mais provável usando o vetor de similaridades, calculando essa probabilidade usando as probabilidades condicionais

[43] https://recordlinkage.readthedocs.io/en/latest/ref-classifiers.html
[44] Hastie, Tibshirani & Friedman, **The Elements of Statistical Learning**, 2017, p. 211.

calculadas com o conjunto de treino. Assim, tomamos a classificação para a qual:

$$\hat{y} = arg\,max\, p(y) \prod_{i=1}^{n} P(x_i|y)$$

Há várias formas de calcular a probabilidade condicional. Para variáveis discretas podemos simplesmente usar contagem. Contudo, como temos variáveis contínuas, podemos usar a distribuição normal, que é a versão Gaussiana do classificador Bayesiano:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

O Perceptron não foi usado isoladamente. Também foram construídos pesos baseados na experiência prática de pareamento de outras bases pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Essas experiencias foram discutidas em reunião com a equipe responsável pelo pareamento de certidões de óbito. Esses pesos foram combinados com os resultados do aprendizado de máquina da seguinte forma: só aceitamos pares reconhecidos como corretos tanto pelos critérios do aprendizado de máquina quanto pelos critérios baseados no IBGE. Assim, se um par de duas observações, uma oriunda de cada base, não for reconhecido como correto, ou seja, como se tratando do mesmo nascimento, de acordo com os dois cálculos, assumimos que o pareamento foi equivocado. Essa opção foi conservadora uma vez que nos foram disponibilizadas poucas informações comuns entre as duas bases para realizar o pareamento. Na base que acessamos, o nome da criança e o horário de nascimento, por exemplo, só estavam disponíveis no SIRC. Se tais variáveis existissem tanto no SIRC quanto no SINASC, o pareamento seria feito de maneira mais segura. Sem um conjunto diversificado de informações comuns, o risco de ocorrer pareamentos errados é grande sem um método conservador. Obviamente, e escolha por critérios conservadores e o uso de vários índices de similaridade certamente impõe um custo. Provavelmente pares na realidade corretos foram considerados incorretos pelo algoritmo. Um exemplo disso é que o SIRC oferece tanto o nome do pai quanto o nome das mães em uma tabela separada para filiações. Assim, unificamos a tabela de certidões de nascimento com a de filiações, filtrando somente o nome da mãe, tendo em vista que o SINASC não dispõe do nome do pai. Contudo, infelizmente observamos muitos erros na entrada de gênero da tabela de filiações. Assim, mesmo com o filtro, para muitos nascimentos recebemos o nome do pai no lugar do nome da mãe. Para situações como

essa um método menos conservador poderia reconhecer os pares apesar do erro, já que para a maioria dos pais casados o sobrenome seria similar. Ocorre que ter sobrenomes em comum também pode ocorrer entre pessoas sem qualquer relação de parentesco. Dessa forma, o custo de um método menos conservador em termos de pareamentos falsos não seria justificável. Uma vez aplicados tais critérios e encontrados os pares, completamos os números das Declarações de Nascidos Vivos (DNVs) faltantes. Havendo o número da DNV em uma das bases para o par, aplicamos o mesmo número para as duas. Se ninguém no par tem o número, geramos aleatoriamente um número e preenchemos a DNV nas duas bases. Com as DNVs geradas pelo pareamento probabilístico, aplicamos a união determinística usando a DNV como chave.

Esse processo, bem como vários outros, foram implementados por meio de funções em Python. Essas funções podem ser aplicadas a diferentes bases de dados. A primeira no código enviado é *calcular_painel_cartorios(Municipios,Cartorios,Diretorio)* . A função usa a lista de todos os cartórios do país para elaborar outra lista, esta com o número de cartórios em cada município, por ano. Também fizemos uma função semelhante para elaborar a mesma lista para unidades interligadas, *calcular_painel_unidades_interligadas(Municipios,Unidades,Diretorio)*. A função que segue é a *unir_bases_de_dados(Base_Referencia,Bases,Diretorio)*, que implementa o método *merge* da biblioteca Pandas, o método determinístico para unir duas bases, já mencionado acima. Em seguida, há uma função para carregar as bases de dados que unimos, *carregar_dados(Diretorio)*, e as transforma num vetor de bases a ser pareado[45]. Essa função é usada na anterior para facilitar o pareamento das variáveis de controle obtidas de diferentes fontes. Em seguida temos duas funções que calculam agregados municipais com base em uma lista de nascimentos, *resumir_sinasc (Ano,Municipios,SINASC,Diretorio)* e *calcular_sub_registro(Diretorio),* sendo que a primeira calcula o nascimento por características como educação ou idade da mãe, e o segundo o sub-registro. Por fim, para o painel de controles alguns trabalhos manuais foram necessários em conjunto com a aplicação das funções, pois nem todas as bases tinham o código do município e as grafias dos nomes variaram. A função seguinte é *limpar_bases(Diretorio)* que faz a limpeza das bases do SIRC e do SINASC, ajustando nomes, tipos e formatos de variáveis. Essa função foi usada na função seguinte, *parear_registros_det(Diretorio)*, que aplica o pareamento

[45] O nome dos arquivos tem que estar correto para que sejam carregados.

determinístico nas bases limpas. Obviamente, como muitas observações não tem os números da DNV, o pareamento determinístico é insuficiente. A função *separar_dados_pareados(Diretorio)* separa, então, as observações pareadas pela DNV das não pareadas. Seguimos então com uma função chamada *parear_registros_prob(ano,mes,Diretorio)* que, como descrevemos no produto dois, é a mais relevante delas. Essa é a função que faz o pareamento probabilístico dos dados não pareados. Nela foram implementadas as bibliotecas Record Linkage Toolkit e Scikit-learn.

Dada sua importância na pesquisa, vale destacar alguns pontos de seu funcionamento. A função abre a base de pareada deterministicamente pela DNV que mencionamos acima. Ela toma as observações sem par e uma amostra de 10000 observações das pareadas. A amostra utilizada é passada pela função do Record Linkage que constrói pares e vetores de características e usada como conjunto de treino no algoritmo de Machine Learning. Esse procedimento de carregar apenas uma amostra inicialmente é utilizado para economizar memória. Carregar e, pior ainda, treinar com a base de todos os nascimentos do país seria inviável em termos de memória. O número de combinações seria muito grande. Como descrevemos acima, tanto o treino quanto a classificação exigem que o Record Linkage Toolkit gere os candidatos a par. Implementamos isso na função *gerar_pares(dn,cn).* Mais uma vez, a biblioteca gera pares em potencial, e.g. a primeira observação do SIRC e a quarta observação do SINASC. Associado a cada um dos pares, testa um vetor de similaridades, e.g. similaridade do nome da mãe segundo Levenshtein, diferença entre os códigos do município de residência etc. Com os pares gerados para a amostra, podemos treinar o algoritmo de 'aprendizado máquina' e, em seguida, usar o aprendizado para classificar pares gerados para os dados não pareados. Vale ressaltar novamente, que essa não seria a única forma de proceder. Suponhamos que a observação "x_i" de uma base e a "y_j" da outra têm "$k \in K$" campos em comum, gerando um vetor com "k" índices de similaridade entre as duas. Poderíamos simplesmente atribuir pesos para cada medida calcular a norma euclidiana e estabelecer um corte, uma espécie de média mínima das similaridades a partir da qual consideramos um par como coreto. É o que fizemos com a abordagem inspirada na metodologia do IBGE, descrita acima. Usar exclusivamente essa abordagem é, contudo, uma decisão problemática, pois os pesos níveis de corte seriam exclusivamente *ad hoc*, e não inferidos a partir dos dados. Para evitar esse viés, combinamos as duas abordagens. Na parte de aprendizado de máquina usamos a aprendizagem supervisionada por meio do Perceptron, como descrito.

5. DADOS OBTIDOS

Aqui descrevemos brevemente os dados utilizados. Foram utilizadas diversas variáveis de controle vindas de bases de dados oficiais. Elaboramos um painel com variáveis municipais e estaduais, obtidas de diferentes fontes oficiais. Controles relativos à economia e finanças públicas foram obtidos no IPEADATA[46] e do IBGE[47]. Entre esses controles estão PIB, PIB per capita, importações e exportações, deflator do PIB[48], gastos municipais em saúde, educação, transporte e justiça. O IPEADATA também oferece o total de homicídios por município em um ano. Esse controle busca refletir a qualidade das instituições na região. O DATASUS[49] por sua vez oferece, com algumas omissões, dados de saúde, de infraestrutura hospitalar e de saneamento básico. Além disso, combinando informações do DATASUS e do IBGE[50,] obtivemos o índice de Gini[51] e população desocupada nos estados. Por fim, usamos os dados abertos do SINASC para o número total de nascimentos por raça, escolaridade, idade, educação e estado civil da mãe. Esta esse não é a mesma base do SINASC fornecida pelo MDHC, mas os organizados nas páginas do SUS[52] e da Secretaria de Vigilância em Saúde[53]. Isso ocorreu porque inicialmente os dados disponibilizados pelo MDHC tinham inconsistências, como descrevermos adiante. Os controles vindos no SINASC foram obtidos antes da base do MDHC ser corrigida, mas refletem os mesmos números.

Outra fonte de informações foi a Pesquisa de Informações Básicas Municipais (MUNIC)[54]. Organizada pelo IBGE, ela reúne informações administrativas sobre os governos municipais. Nosso uso da base, foi, contudo, restrito. Isto porque grande parte das variáveis disponíveis na base não se aplicavam ao estudo, como informações sobre políticas de meio ambiente. Mais importante, contudo, é o fato de

[46] http://ipeadata.gov.br/Default.aspx
[47] https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?=&t=downloads
[48] Além do deflator implícito usamos também o IPCA.
[49] http://tabnet.datasus.gov.br/
[50] https://sidra.ibge.gov.br/
[51] Algumas omissões do índice de Gini foram completadas usando os dados disponíveis em http://infodf.codeplan.df.gov.br/?page_id=23
[52] http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinasc/cnv/nvbr.def
[53] http://svs.aids.gov.br/dantps/centrais-de-conteudos/dados-abertos/sinasc/
[54] https://www.ibge.gov.br/estatisticas/sociais/saude/10586-pesquisa-de-informacoes-basicas-municipais.html?=&t=downloads

que para as variáveis relevantes há muitas omissões e mudanças de metodologia entre as coletas. As variáveis que coletamos da MUNIC foram o (1) total de funcionários da administração direta municipal, (2) a existência de um ou mais postos de cartório ou unidade interligada em estabelecimento de saúde que realiza partos e (3) o número de tais postos ou unidades.

Assim, a versão final das variáveis que indicam o número e localização de unidades interligadas em municípios foi construída a partir de diversas bases. Isso porque o número de unidades interligadas é coletado separadamente pelo Conselho Nacional de Justiça (CNJ) e pela Associação Nacional dos Registradores de Pessoas Naturais (ARPEN). Como dissemos, a existência de um ou mais cartórios ou unidades interligadas em estabelecimentos de saúde que realizam partos foi coletada pelo IBGE em 2009 e 2011. A partir dessas coletas, o IBGE parou de tratar a questão como binaria e registrou o número de unidades e cartórios em 2014. O CNJ também dispõe da informação, mas na forma de uma lista de postos de cartório e unidades interligadas em estabelecimentos de saúde. Ocorre que a lista não informa a data de instalação e sua última atualização foi em 2019. Dessa forma, o máximo que podemos extrair dessa lista é o total de cartórios e unidades em 2019. Finalmente, a ARPEN, por meio do MDHC, enviou sua própria lista com unidades interligadas, ano de instalação, localização e número de serventia. O ano de instalação, na verdade, é inferido a partir da data do primeiro registro. Combinando essas listas, podemos calcular quantas unidades interligadas cada município tinha em cada ano. Ocorre que o número de unidades segundo a ARPEN é menor do que o do CNJ e muito menor do que o do que a coleta do IBGE em 2014. Nossa hipótese de explicação da disparidade é que nem todos os cartórios enviam dados a ARPEN. Discutiremos as dificuldades oferecidas por essas incongruências, bem como os métodos de solução do problema, mais adiante.

Por fim, as medidas de sub-registro têm duas origens. Uma é o pareamento de 2015 a 2017 já realizado pelo IBGE[55]. A outra, naturalmente, é o pareamento realizado no

[55] Costa, Trindade & Oliveira, *Pareamento de Dados das Estatísticas do Registro Civil e das Estatísticas Vitais in* **Sistemas de Estatísticas Vitais: Avanços, Perspectivas e Desafios**, IBGE, 2015, p. 26.

bojo desse estudo. Já descrevemos amplamente o processo de pareamento. Trata-se, em suma, de (1) uniformizar códigos, formatos e nomes de variáveis nas duas bases, (2) parear pela DNV onde possível, (3) onde não for possível[56], criar candidatos a pares no município, mês e ano e calcular, para aquele par, índices de similaridade, (4) usar aprendizado de máquina sobre uma amostra de dados pareados pela DNV para entender como se comportam os índices de similaridade, (5) usando o algoritmo de aprendizado de máquina treinado e pesos inspirados em estudos anteriores, classificar os pares como corretos ou incorretos, (6) para os pares corretos, atribuir a mesma DNV nas duas bases, indicando que tratam do mesmo nascimento, (7) parear deterministicamente usando as DNVs atribuídas, (8) usando a lista de nascimentos computar, para cada município e em cada ano, a taxa de sub registro.

[56] https://biblioteca.ibge.gov.br/visualizacao/livros/liv101575.pdf

6. ETAPAS DO DESENVOLVIMENTO

Aqui faremos uma breve discussão das etapas do desenvolvimento do estudo. Aprofundaremos as discussões sobre dificuldades e limitações no tópico seguinte. A primeira etapa foi a elaboração do plano de trabalho. Nessa etapa analisamos em linhas gerais as estruturas do SIRC e do SINASC e escolhemos os modelos de inferência causal aplicáveis. Em seguida, entramos em contato com a equipe de Tecnologia da Informação (TI) do MDHC para acessar as bases. Inicialmente, achamos que as bases poderiam ser acessadas de qualquer máquina usando DBeaver, mas, por questões de segurança e proteção dos dados, fomos orientados a acessar remotamente uma máquina física, dentro do MDHC.

Com essa máquina, pudemos acessar as bases. Nesse momento as bases do SINASC não tinham qualquer tabela. Entrando em contato com a TI novamente, conseguimos trazer os dados do SINASC. Infelizmente, esses dados vieram incompletos. Provavelmente em algum momento da extração pela TI o Excel foi usado, pois todos os anos no SINASC tinham o mesmo número de linhas, que era o número máximo de linhas aceito pelo Excel. Além disso, algumas variáveis que estavam presentes nas documentações não vieram para a base.

Feitas as correções possíveis, seguimos na elaboração, em paralelo, do pareamento e da construção do painel de controles. A construção do painel envolveu, como descrevemos, baixar dados de sites oficiais e usar, na medida do possível, anos e códigos de municípios como chaves para uni-los numa única base. Algumas incongruências entre bases foram encontradas, de forma que nos limitamos às que cobriam um período maior, com menos omissões. A MUNIC em particular tem uma lista grande de variáveis, com diversos códigos e mudanças de metodologia ao longo do tempo. Assim, foi necessário algum trabalho manual para adequá-la.

No que tange ao código, o primeiro passo foi extrair os dados usando SQL, com os filtros apropriados. Uma vez obtidos os dados, seguimos para planejar a estrutura de cada função. Vimos quais tabelas precisavam ser unidas, quais variáveis precisavam ser ajustadas etc. Uma vez preparados os dados de controle e as tabelas brutas do SIRC e do SINASC, aplicando as funções de adequação das variáveis, procedemos a

implementar o pareamento probabilístico.

Infelizmente, a máquina que usávamos naquele momento não conseguia lidar com o volume de dados. Assim, a TI do MDHC mais uma vez, gentilmente nos auxiliou, criando uma máquina virtual com 32GB de memória. Ainda assim, o código de pareamento não era executado pois, na sua estrutura original, exigiria mais de 10TB de memória.

De volta ao código, resolvemos restringir quais trechos das bases usar. Escolhemos usar integralmente apenas os dados que tinham que ser pareados, de forma que os dados já pareados não ocupassem memoria em vão. Ainda assim, precisaríamos dos dados pareados para o treino. Decidimos então por treinar com uma amostra de 10.000 indivíduos. A escolha desse número foi, em certa medida *ad hoc*. O resultado não mudaria muito com uma amostra de 9.000 ou 11.000 indivíduos. A escolha desse patamar tem, contudo, uma razão matemática. Pelo princípio da casa dos pombos[57], se no Brasil há 5.570 municípios, numa amostra de 10.000 nascimentos certamente teremos municípios com mais de um nascimento. Isso é importante porque, como afirmamos acima, calculamos as combinações possíveis de nascimentos dentro de um município. Como para o aprendizado de máquina precisamos de exemplos de pares corretos e de pares incorretos, precisamos, pelo menos, de mais de um indivíduo nascendo em cada município. Por exemplo, imagine que no município "a" nasceu apenas o indivíduo "i". Seja "$cn_i^{(a)}$" seu registro na base do SIRC e "$dn_i^{(a)}$" seu registro no SINASC. Como não há outros nascimentos nesse município, o único par possível é "$\left(cn_i^{(a)}, dn_i^{(a)}\right)$" que é o par correto. Como o algoritmo, nesse caso, vê apenas um exemplo de pareamento correto, ele não consegue aprender. Imagine, contudo, que há dois indivíduos no município, individuo "i" e individuo "j". Nesse caso, as combinações possíveis são:

$$\begin{bmatrix} \left(cn_i^{(a)}, dn_i^{(a)}\right) & \left(cn_i^{(a)}, dn_j^{(a)}\right) \\ \left(cn_j^{(a)}, dn_i^{(a)}\right) & \left(cn_j^{(a)}, dn_j^{(a)}\right) \end{bmatrix}$$

Agora, o algoritmo tem dois exemplos de pareamento correto e dois de pareamento

[57] Aigner & Ziegler, **Proofs From The Book**, 2003, p. 130.

incorreto. A representação matricial é conveniente, pois os pares na diagonal principal são os verdadeiros. Dessa forma, escolhemos um número que é, de certa forma, arbitrário, mas cuja escolha foi informada pela teoria. Esse número permite que existam exemplos dos dois tipos de pares. Traçada essa estratégia de otimização, vimos que era possível executar o código. Para cada mês de cada ano, a execução dura de 6 a 20 horas na máquina virtual do MDHC, de forma que o pareamento completo leva alguns dias. Feito o código e iniciado o pareamento, verificamos em amostras seu sucesso. Inspecionamos alguns dos pares e todos estavam corretos. Seguimos então com o pareamento e com análise estatística.

Diversas variações dos modelos descritos acima foram executadas. Os resultados serão relatados adiante. Como discutimos aqui e no plano de trabalho, uma das formas de lidar com variáveis omitidas seria nos restringirmos a nascimentos ou altas realizadas próximas as 18:00 ou próximos a feriados, com um desenho inspirado em regressão descontínua. Nesse caso a variável dependente seria o sub-registro em cada hospital. Em tese, isso seria possível pois após o corte muitas unidades interligadas fecham. A descontinuidade média de cada hospital indicaria o efeito da presença de uma unidade interligada. A ideia subjacente a essa estratégia é que, num mesmo hospital, a única diferença entre nascimentos antes e depois do corte é a probabilidade de acesso a uma unidade interligada, de forma que o tratamento, i.e., exposição à unidade interligada, seria ditado por um fator exógeno, i.e., data e horário do nascimento. Infelizmente, não pudemos implementar essa abordagem.

A primeira limitação foi a restrição de tempo para a realização da pesquisa, de modo que tivemos que escolher qual análise fazer. A implementação do estudo com dados em painel e com pareamento por escore de propensão são mais rápidas de se implementar e mais usadas na literatura. O uso aqui proposto do modelo de descontinuidade na regressão desviaria um pouco da estrutura da literatura, nossa proposta era aplicar um modelo de probabilidade linear para cada hospital e município, com um indicador do corte temporal, e depois calcular o efeito agregado. Essa estrutura é nova e exigiria mais análises para verificar sua adequação.

Além disso, uma limitação dessa abordagem é a variação no horário de funcionamento das unidades interligadas. Idealmente, essa estratégia teria que ser

implementada com horários, pois só os feriados imporiam uma amostra muito restrita. Contudo, segundo estudo realizado pelo Ministério Público do Estado do Rio de Janeiro[58], das unidades interligadas no estado, aproximadamente 18% funcionam além do horário comercial, mais de 40 horas semanais, sendo que 64% funcionam por 35 ou horas ou menos durante a semana[59]. Assim, há o risco de que a escolha do horário de funcionamento pelo poder público esteja correlacionada com outras variáveis. Se, por exemplo, nas unidades interligadas mais eficazes, ou seja, com maior impacto sobre o sub registro, o horário de funcionamento vai além das 18:00, nenhuma descontinuidade será observada naquele ponto, mesmo que o efeito exista.

Outra dificuldade seria o uso do tipo do parto, já que médicos e pacientes podem influenciar datas e horários de cesárea, sendo escolha correlacionada com outras variáveis. Assim, a descontinuidade na regressão seria mais adequada à amostra de partos normais. Claro, poderíamos testar se a quantidade e perfil dos nascimentos próximo ao corte temporal é a mesma.

Por essas dificuldades em termos de tempo e implementação decidimos abandonar, pelo menos para efeitos desse projeto, o uso da descontinuidade. Ele se aplica aos modelos de escolha discreta. Como toda a construção dos dados foi orientada ao painel municipal, para fazermos modelos individuais, sejam eles de escolha discreta ou de pareamento por escore de propensão, teríamos que trabalhar os dados e dessa forma adicionar alguns controles municipais na base de nascimentos individuais. Essas alterações na estrutura da base e a execução dos modelos leva tempo, de forma que nós restringimos às aplicações mais clássicas da inferência causal.

Seguimos, portanto, com as análises que usam o painel municipal. Trata-se da regressão com efeitos fixos e pareamento com escore de propensão. Ambas as abordagens buscam controlar para efeitos de outras variáveis e comparar municípios semelhantes, com e sem unidades interligadas[60]. Isso porque, mais uma vez, se compararmos municípios diferentes em outras variáveis além das unidades

[58] CENPE & MPRJ, Diagnóstico das Unidades Interligadas do Estado do Rio de Janeiro, 2019, p. 13
[59] O mesmo estudo aponta como fator importante na postergação do registro a falta da presença do pai para o reconhecimento de paternidade.
[60] Imbens & Wooldridge, *Recent Developments in the Econometrics of Program Evaluation* in **NBER Working Papers**, 2008.

interligadas, poderemos observar níveis de sub-registro diferentes, mas que refletem, na verdade, outras diferenças subjacentes. Essencial, nos dois casos, é adicionar também o efeito do tempo, já que a implementação de uma unidade interligada pode ocorrer em um momento cuja tendencia de redução do sub-registro já vinha acontecendo. Por fim, buscamos enfrentar o viés de auto seleção, veja que a implementação e aplicação do Provimento CNJ nº 13 de 03/09/2010 é dos Governos locais. Assim, variações no sub-registro no grupo de municípios e estados que escolheram essa política pode se dar em virtude de tais entes federados, por possuir mais recursos já terem promovido outras políticas de registro civil.

7. LIMITAÇÕES DO ESTUDO

Há diversas limitações e desafios nesse estudo. Algumas dessas questões são parcialmente resolvidas pelos métodos estatísticos, outras não. Estimativas anteriores do sub-registro no Brasil sugerem uma queda sistemática do sub-registro civil de nascimento nos últimos anos, passando de 17,4% em 2004[61] para 6,7% em 2012[62] e para apenas 2.6% em 2017[63]. Dessa forma, a identificação de efeitos causais da política de unidades interligadas requer a separação do efeito do tempo. Ocorre que essa separação é dificultada pelo ruído, ou seja, a grande variabilidade dos dados, especialmente em virtude de imprecisões nas medidas tanto do número de unidades quanto do sub registro.

Além disso, há o viés de auto seleção, já descrito. Podemos observar efeitos, na verdade, oriundos das características que levam um governo local a demandar esse tipo de política, nos termos do Provimento CNJ nº 13 de 03/09/2010[64].

Mais importante, contudo, é a validade das medidas de sub-registro civil de nascimento. O pareamento das duas bases, em tese, nos dá o sub-registro, pois assumimos que uma criança cujo nascimento foi observado pelo sistema de saúde e registrado no SINASC, mas não observado pelos cartórios no SIRC, tem de ser um sub-registro. Ocorre que nem todos os cartórios digitalizaram todo seu acervo e nem todos os cartórios estão no SIRC. Isso fica claro quando comparamos em números absolutos o número de entradas nas duas bases. Pelo menos nas bases a que tivemos acesso, a disparidade é muito grande para que seja uma questão apenas de sub-registro. Assim, a presença de um nascimento em apenas uma das bases não garante que houve sub-registro. O nascimento pode ter sido registrado, mas o dado simplesmente não fazer parte da base do SIRC ainda. Assim, em vez de explicarmos o sub-registro, é possível que estejamos explicando a falta de informações no SIRC. Isso prejudicaria conclusões sobre o número absoluto de sub-registro, mas, a priori,

[61] https://www.ibge.gov.br/estatisticas/sociais/populacao/26176-estimativa-do-sub-registro.html?edicao=26182&t=resultados
[62] https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/14542-asi-entre-2002-e-2012-sub-registro-de-nascimentos-caiu-de-203-para-67
[63] https://www.gov.br/mdh/pt-br/navegue-por-temas/registro-civil-de-nascimento/registro-civil-do-nascimento
[64] https://atos.cnj.jus.br/atos/detalhar/1298

não inviabilizaria o estudo do impacto causal das unidades sobre o sub-registro. Isso porque a variação na medida, digamos, poluída do sub-registro ainda assim seria influenciada pelas unidades se a falta de envio de informações do SIRC não for correlacionada como sub-registro. Esse é o maior problema. É muito difícil assumir que o sub-registro e a falta de informações do SIRC não sejam variáveis correlacionadas. Isso porque a carência de recursos e estrutura em um município provavelmente impactam tanto os cartórios do local quanto a população. Dito de outra forma, se a falta de estrutura e recursos impede que o cartório envie informações, provavelmente essa mesma carência afeta a capacidade de mães e pais registrarem. Ademais, é possível que nas regiões onde o benefício das unidades é maior, ele não seja observado pela falta de envio de informações do cartório.

O uso que fizemos dos métodos de pareamento por escore de propensão não foi rigoroso. Isso porque esses métodos em geral são aplicáveis a dados em painel. Outras especificações teriam que ser rodadas. Por fim, cabe destacar que há uma limitação de tempo. Dadas as restrições de tempo da consultoria, não pudemos explorar todas os usos possíveis dos dados. Há diversas especificações dos modelos que não foram testadas. Discutiremos algumas delas no tópico de possíveis extensões.

8. ANÁLISE DE RESULTADOS

Seguimos, então, com a discussão dos resultados, iniciando pelas regressões com efeitos fixos e depois com os métodos de pareamento[65].

Como descrevemos acima, buscamos lidar com as variáveis omitidas para fazer a inferência causal[66] já que os fatores intrínsecos[67] a uma unidade desaparecem com a diferença da média[68]. Logo, nossas especificações são variações da seguinte:

$$SUB_{ti} = \gamma_0 + \gamma_1 UNIDADES_{ti} + \beta_0' ECONOMIA_{ti} + \beta_1' GASTOS_{ti} + \beta_2' INDICES_{ti} + \beta_3' SAUDE_{it} + NASC_{it} + c_i + \alpha_i t + \varepsilon_{it}$$

Para um município "i" no período "t". Destacamos, mais uma vez, que assumimos que o processo gerador das omissões do SIRC é ortogonal aos demais dados, validando a inferência causal sobre os impactos das unidades, ainda que não a previsão de determinado nível de sub-registro.

Como já destacamos, a medida de unidades interligadas apresenta alguns desafios. Como já indicamos, construímos a variável usando as bases do IBGE, CNJ e ARPEN em conjunto. Como o IBGE mudou a metodologia, temos uma versão binária e uma versão inteira da variável. A versão binária pode ser gerada a partir do número de unidades, i.e. 1 (um) se há uma ou mais unidades e zero se não há nenhuma. Outra opção é ignorar 2009 e usarmos a variável inteira. Um problema adicional é que os números reportados pelas três instituições são contraditórios. Assim, aceitamos como número de unidades num município num ano o maior dos disponíveis nas três bases.

No que tange o vetor $ECONOMIA_{ti}$ podemos incluir variáveis como o índice de Gini, taxa de desocupação estadual, PIB e PIB per capita. No caso de $GASTOS_{ti}$ há as variáveis de

[65] Pareamento aqui se refere ao pareamento de grupos de tratamento e controle e não ao pareamento do SIRC e do SINASC. Infelizmente, em português os dois processos usam o mesmo termo técnico, o que gera confusão. Em inglês, os métodos de pareamento para inferência causal são chamados de "matching", ao passo que o pareamento de bases de dados diferentes se chama "linkage".

[66] Imbens & Wooldridge, *Recent Developments in the Econometrics of Program Evaluation* in **NBER Working Papers**, 2008.

[67] Heiji, de Boer et. al. **Econometric Methods with Applications in Business and Economics**, 2004, p. 692

[68] Por exemplo, duas cidades distintas que, apesar de mudanças econômicas, sempre terão culturas diferentes ou duas pessoas que, independentemente da educação a que tem acesso, sempre terão diferentes níveis de aptidão para uma determinada tarefa.

gastos municipais em educação, saúde, transporte e justiça. O vetor $INDICES_{ti}$ pode incluir índices de criminalidade e de mortalidade infantil. O vetor $SAUDE_{it}$ incorpora variáveis da infraestrutura de saúde, como leitos de internação. Por fim, temos o vetor $NASC_{it}$ que contém porcentagens que indicam o perfil dos nascimentos em termos de escolaridade da mãe, idade da mãe e raça.

Há variáveis que usamos, mas que por serem constantes ao longo do tempo, não são incorporadas à regressão. São elas: a localização da cidade no domínio amazônico, aglomeração urbana, o total de cartórios e o número de unidades interligadas segundo a coleta do IBGE em 2014.

As estatísticas descritivas abaixo cobrem todo o painel de controles. Ele cobre um período maior do que as medidas de sub-registro, 2005 a 2019. As medidas de sub-registro cobrem o período de 2015 a 2019. Por essa razão, o número de observações é maior para a maior parte dos controles.

Variável	Observações	Média	Desvio-Padrão	Mínimo	Máximo
Sub registro	16,710	3.24%	5.55%	0.00%	95.62%
Número de unidades interligadas	83,546	0.05	0.73	0.00	63.00
Número de cartórios	79,470	2.06	3.05	0.00	87.00
Taxa estadual de desocupação	44,560	0.09	0.03	0.03	0.18
Gini estadual	77,977	79.18%	4.77%	55.53%	90.65%
PIB real per capita, ano-base 1993 (RS$)	66,807	425.13	466.66	8.30	21262.23
Homicídios	72,366	9.73	60.45	0.00	2925.00
Homicídios per capita	61,237	0.0002	0.0002	0.00	0.0023
Funcionários da administração pública direta municipal per capita	61,029	0.05	0.03	0.00	3.57
Gastos municipais com per capita educação	58,223	17.86	11.11	0.00	1927.94
Gastos municipais per capita com saúde	58,223	14.40	9.06	0.00	1195.71
Gastos municipais per capita com transporte	58,223	2.60	4.46	0.00	146.26
Gastos com municipais per capita com justiça e segurança	58,223	0.07	0.77	-4.00	100.99
Leitos hospitalares para alta complexidade	6,761	9.38	33.22	1.00	516.00

Leitos de alojamento conjunto para mãe e recém-nascido	8,393	32.21	82.96	1.00	1629.00
Óbitos infantis	9,286	35.00	109.29	1.00	2312.00
Taxa de domicílios com água encanada no estado	55,700	92.02%	9.44%	53.29%	99.70%
Total de nascimentos	79,682	548.46	3139.83	1.00	179025.00
Porcentagem de nascimentos de negros e pardos	79,682	51.5%	32.9%	0.0%	100.0%
Porcentagem de Nascimento de indígenas	79,682	1.0%	5.8%	0.0%	97.9%
Porcentagem de nascimentos em que a mãe tinha menos de 19 anos	79,682	21.1%	7.2%	0.0%	66.7%
Porcentagem de nascimentos em que a mãe tinha 11 ou mais anos de estudo	79,682	61.2%	19.0%	0.0%	100.0%
Porcentagem de nascimentos em que a mãe é solteira	79,682	45.2%	19.7%	0.0%	100.0%

A tabela a seguir nos permite visualizar as incongruências entre os diferentes registros do número de unidades interligadas ou postos de cartório em estabelecimentos de saúde que realizam partos. Os dados abaixo mostram o número de municípios para o número de unidades interligadas em um ano:

	Número de unidades interligadas								
Ano	0	1	2	3	4	5	6	7+	Total
2005	5,570	0	0	0	0	0	0	0	5,570
2006	5,570	0	0	0	0	0	0	0	5,570
2007	5,570	0	0	0	0	0	0	0	5,570
2008	5,570	0	0	0	0	0	0	0	5,570
2009	5,570	0	0	0	0	0	0	0	5,570
2010	5,570	0	0	0	0	0	0	0	5,570
2011	5,503	38	11	9	5	3	0	1	5,570
2012	5,491	46	12	9	6	3	1	2	5,570
2013	5,484	51	14	8	7	3	1	2	5,570
2014	4,228	1,045	209	26	34	7	7	14	5,570
2015	5,460	71	17	9	6	4	1	2	5,570
2016	5,452	77	18	9	6	4	1	2	5,569
2017	5,447	82	17	10	6	4	1	2	5,569
2018	5,441	86	19	10	6	4	1	2	5,569
2019	5,182	280	53	19	18	3	3	11	5,569
Total	81,108	1,776	370	109	94	35	16	13	83,546

Como se vê na tabela, o número de municípios em cada nível de unidades interligadas em 2014 diverge consideravelmente.

Com relação às variáveis fixas ao longo do tempo, quisemos restringir a amostra para selecionar locais em que as unidades provavelmente têm um maior impacto. Nossa hipótese, nesse sentido, é que em zonas rurais ou de difícil aceso, em que as condições de transporte e o acesso a quaisquer bens públicos é restrito, a utilidade do registro é menor. Isso seria sobretudo válido para regiões com menos acesso à educação e onde o trabalho infantil no campo é mais frequente. Dito de outra forma, em centros urbanos as chances de uma criança acessar bens públicos como creches, escolas públicas ou programas de transferência de renda é maior, o que incentiva os pais a registrarem. Podemos ver a distribuição dos municípios abaixo.

Tipo de Concentração Urbana	Freq.	Porcent.
Grande	4,155	4.97%
Média	5,745	6.88%
Rural	73,650	88.15%

Localizado na Amazônia Legal	Freq.	Porcent.
Não	71,970	86.14%
Sim	11,580	13.86 %

Cartórios no Municípios	Freq.	Porcent	Cum.
0	20,246	25.5%	25.5%
1	26,123	32.9%	58.4%
2	11,887	15.0%	73.3%
3	7,369	9.3%	82.6%
4	3,981	5.0%	87.6%
5	3,074	3.9%	91.5%
6	2,277	2.9%	94.3%
7	1,485	1.9%	96.2%
8	983	1.2%	97.4%
9	725	0.9%	98.3%
10	285	0.4%	98.7%
11	180	0.2%	98.9%
12	135	0.2%	99.1%
13	141	0.2%	99.3%
14+	579	0.7%	100.0%

Municípios selecionados para receber uma unidade interligada em qualquer ano da base, segundo o IBGE	Freq.	Porcent.
Sim	58,181	69.64%
Não	25,365	30.36%

Aqui cabe destacar que desviamos um pouco de algumas definições do IBGE. Os municípios classificados grandes ou médios seguem a definição do IBGE, mas classificamos qualquer outro município fora de aglomerações urbanas simplesmente como “Rural”, sem outras distinções.

Var. Dependente: Sub-registro	(1)	(2)	(3)	(4)	(5)
Tem Unidade					-0.4486617
					(.5009727)
Número de Unidades Interligadas	-1.050625**	-0.0584005	-0.4526405	-0.1521281	
	(.4325428)	(.4099949)	(.5273378)	(.4178166)	
2017		-1.370873*	-1.379949*	-1.242489*	-1.240654*
		(.0614099)	(.0639044)	(.115739)	(.1156333)
2016		-0.9876144*	-0.9975574*	-0.8865469*	-0.8854993*
		(.0552098)	(.0574886)	(.0996801)	(.0996195)
Constante	3.292849*	4.033*	4.10324*	4.163301*	4.169996*
	(.0200376)	(.0386179)	(.03781)	(.0828735)	(.0692922)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.					

Os resultados iniciais das regressões com efeitos fixos não são conclusivos. Na especificação (1) a única variável independente é o número de unidades interligadas. Nesse caso, cada unidade interligada reduziria o sub-registro em um ponto percentual. Tal efeito seria estatisticamente significante. Ocorre que, ao adicionarmos efeitos fixos dos anos, especificação (2), o efeito das unidades interligadas se torna estatisticamente insignificante. Na especificação (3) nós restringimos a municípios que não estão em áreas de grande concentração urbana e na especificação (4) a municípios que, segundo o IBGE, receberam algum tipo de política de promoção do registro civil de nascimento. Nessas duas especificações, o efeito das unidades ainda não é significativo. Por fim, em (5) usamos uma variável binária indicando se o município tem ou não unidade interligada no lugar do número de unidades interligadas. O efeito da presença de uma ou mais unidades também não é significativo. Dados tais resultados, seguimos para a análise de especificações com mais controles.

Na especificação (A) usamos como variável de interesse a presença ou não de uma ou mais unidades interligadas, variável binária, controlando o efeito do tempo, PIB per capita, porcentagem de recém-nascidos negros ou pardos, porcentagem de recém-nascidos cuja mãe tinha menos de 19 anos, Gini estadual, desocupação estadual, crime per capita e mortalidade infantil. A especificação (B) é praticamente idêntica, mas sem mortalidade infantil, uma vez que essa variável tem muitas omissões e reduz muito as observações usadas. A especificação (C) é a mesma a especificação (A), mas usando o número de unidades interligadas no lugar de uma variável binaria. Por fim, a especificação (D) é a mesma especificação de (C), mas sem mortalidade infantil e restringindo a amostra apenas para municípios fora de grandes centros urbanos.

Var. Dependente: Sub-registro	(A)	(B)	(C)	(D)
Tem Unidade	-0.7055242	-0.2387827		
	(.679425)	(.4790525)		
Número de Unidades Interligadas			-0.0419826	-0.3334808
			(.5286478)	(5114385)
2016	0.5426293	-0.2349737	0.5439951	-0.2723749
	(.4089066)	(.1333258)	(.4088609)	(.138919)
2017	0.215116	-0.4648539*	0.2129326	-0.50902
	(.3965853)	(.1428897)	(.3966981)	(.1486818)
PIB per capita	0.0001601	-0.0001291	0.0001562	-0.0000998
	(.0004456)	(.0002285)	(.0004476)	(.000238)
% Nascimentos de negros e pardos	-0.0514251	-0.2543462	-0.0670641	-0.2227684
	(.9645291)	(.3878155)	(.9644181)	(.3924255)
% Nascimentos cuja mae tem menos de 19 anos	-7.267284	-0.3345418	-7.226938	-0.3420044
	(5.718755)	(.8204311)	(5.719906)	(.8286403)
Gini Estadual	8.729671	-7.084983	8.452096	-7.065605
	(17.96963)	(6.224647)	(17.98919)	(6.530775)
Desocupação Estadual	-0.5070933*	-0.2914504*	-0.5083983*	-0.2824352*
	(.1326068)	(.0469611)	(.1325659)	(.0492943)
Crime per capita	207.2788	145.8946	197.6841	117.5909
	(580.3997)	(177.9067)	(580.5171)	(181.3452)
Mortalidade Infantil	-17.30535		-17.05423	
	(13.60586)		(13.59926)	
Observações	1,774	15,889	1,774	15,114

*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.

Como vemos, em todas as especificações o coeficiente das unidades é negativo, ou seja, as unidades diminuiriam o sub-registro. Ocorre que, com a variância dos dados, esse efeito não é significativo, mesmo usando os municípios como clusters para o cálculo dos desvios. Mesmo restringindo a amostra a municípios fora de grandes centros urbanos e removendo mortalidade infantil, o efeito ainda não pode ser capturado com precisão. A única variável significativa é a desocupação estadual. Uma hipótese para explicar isso é que, com o desemprego, mais famílias recorrem a programas de transferência de renda, tal como o Bolsa Família, o que exige que registrem filhos e os matriculem na escola.

Com relação aos métodos de pareamento, os resultados foram mais robustos. Como já descrevemos, o pareamento por escore de propensão usa de probabilidade do indivíduo, hospital ou município serem expostos a uma unidade interligada. Como na análise preliminar nos restringimos aos municípios, selecionamos municípios para o grupo de controle usando dois métodos. O primeiro é o escore de propensão[69], ou seja, calculando a probabilidade de que tal município receba uma unidade interligada:

$$\boldsymbol{P(Unidades\ Interligadas_{it} = 1|x_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}}$$

Onde:

$$\boldsymbol{x_i'\beta = \beta_1' ECONOMIA_{ti} + \beta_2' INDICES_{ti} + \beta_3' SAUDE_{it} + NASC_{it}}$$

$$\hat{ATE} \approx \frac{1}{S}\sum_{s=1}^{S}\left(\overline{SUB}_{is} - \overline{SUB}_{js}\right)$$

Onde "S" é o número de pares e "$\hat{ATE}$" o efeito estimado da introdução de uma ou mais unidades interligadas. Uma alternativa é o método de Mahalanobis[70], que fundamentalmente busca como controle de uma observação que recebeu o tratamento, i.e., a unidade interligada, o mais semelhante em termos de suas características, ou seja, o com a menor distancia euclidiana entre as variáveis normalizadas. Uma vez pareadas as observações, comparamos o sub-registro médio

[69] Rubin & Rosenbaum, *The Central Role of the Propensity Score in Observation Studies for Causal Effects*, **Biometrika**, 1983, p. 42.
[70] King et. Al., *Comparative Effectiveness of Matching Methods For Causal Inference*, **The Institute for Quantitative Social Science, Harvard,** 2011, p. 4

entre os dois grupos. Mesmo variando o grupo de variáveis e o método que usamos para fazer o pareamento, o efeito do tratamento e consistentemente negativo.

Na primeira especificação usamos o método PSM e como variáveis PIB per capita, porcentagem de nascimentos de negros e pardos, porcentagem de nascimentos em que a mãe tinha menos de 19 anos, Gini estadual, desocupação, crime per capita e mortalidade infantil. Nesse caso, os resultados são:

(PSM) Observações: 588	2015	2016	2017
Efeito da presença de unidades	-1.03**	-1.25*	-.165
	(.65)	(.40)	(.32)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.			

Mais uma vez, o efeito das unidades interligadas é negativo, ou seja, a presença de uma ou mais unidades diminui o sub-registro. Aqui encontramos resultados mais robustos, uma vez que o efeito para 2015 e 2016 são estatisticamente significantes. Mais uma vez, a inclusão de mortalidade infantil reduz muito a amostra em virtude das omissões. Retirando mortalidade infantil obtemos resultados melhores. Os efeitos continuam negativos e, desta vez, estatisticamente significantes para todos os anos.

(PSM) Observações: 5296	2015	2016	2017
Efeito da presença de unidades	-1.90***	-1.03*	-.474***
	(1.16)	(.37)	(.27)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.			

Em seguida, separamos a amostra entre municípios em áreas urbanas e em áreas rurais. Os resultados são:

(PSM) Observações: 4671	2015	2016	2017
Efeito da presença de unidades	-1.12**	-1.18*	-.58***
	(.58)	(.28)	(.45)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.			

Restringindo a amostra para municípios em áreas urbanas não obtemos nenhum resultado significativo. Isso oferece evidências de que a nossa hipótese pode estar correta. Seguimos usando o método de Mahalanobis. Os resultados são semelhantes:

(MNN) Observações: 4671	2015	2016	2017
Efeito da presença de unidades	-1.24*	-1.15*	-.67***
	(.40)	(.28)	(.42)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.			

Adicionalmente, especificamos um modelo usando a amostra de municípios fora de regiões urbanas, mas de estados cuja adesão dos cartórios ao SIRC é de 100%. Esse nível de adesão só ocorreu em 2017. Nesse caso o coeficiente é de "-.62" com significância de 1%. Resultados melhores são obtidos se restringimos a amostra aos municípios que receberam alguma política de promoção de sub-registro em 2014. O efeito continua negativo, mas dessa vez todos são significantes:

(MNN) Observações: 1471	2015	2016	2017
Efeito da presença de unidades	-1.35*	-1.25*	-1.05*
	(.41)	(.25)	(.37)
*, ** e *** indicam, respectivamente, significância estatística de 1%, 5% e 10%.			

Os resultados dos métodos de pareamento são interessantes. Eles sugerem que, ao introduzirmos uma ou mais unidades interligadas em um município, o sub-registro pode cair aproximadamente de 2 a 0.5 pontos percentuais. Como já dissemos, o uso do valor absoluto do sub-registro é discutível, mas o efeito benéfico das unidades interligadas em áreas rurais se mantém.

Resta, contudo, explicarmos a incongruência dos resultados da análise de dados em painel. Seria necessária uma pesquisa mais profundada para compreender o resultado. De qualquer forma, os resultados sugerem um efeito modesto das unidades, tendo em vista que são muito sensíveis a especificações e modelos. Ou seja, são resultados interessantes, mas não são robustos o suficiente. Para aferir com mais segurança o impacto das unidades, seria necessário testarmos mais modelos e buscarmos outras referências teóricas. Discutiremos isso nas extensões.

9. POSSÍVEIS EXTENSÕES E MELHORIAS

Há diversas extensões e melhorias possíveis. Esse estudo deve ser visto como introdutório, visto que ele organizou os dados e testou os modelos mais fundamentais. Podemos aprofundar com outros modelos e o uso de outros dados.

A primeira extensão que precisa ser destacada é o uso de modelos de pareamento para dados longitudinais. Como os métodos de pareamento, tanto PSM quando Mahalanobis, não podem ser aplicados diretamente ao conjunto total de dados longitudinais, estimamos o efeito para cada ano. Há, contudo, métodos de pareamento específicos para dados longitudinais. Seu uso seria um teste interessante das conclusões obtidas.

Em discussões com professores de economia e econometria da Universidade de Delaware e da Universidade do Tennessee sobre o caso, foi levantada a possibilidade de que o contraste radical entre os resultados das regressões e dos métodos do pareamento se deve ao nível das variáveis. Foi sugerido que, talvez fosse mais apropriado trabalhar com variáveis em nível, ao invés de usarmos variáveis per capita, ou até mesmo usar as variáveis em log. Essa escolha, contudo, não pode ser arbitrária, com vistas apenas a melhorar os resultados. Um caminho seria buscar referências na literatura que justifiquem essa mudança e então testar seus impactos no resultado[71].

Outra extensão necessária é implementar os métodos sugeridos no plano de trabalho que foram descartados pela questão do tempo, como os métodos de escolha discreta e de regressão descontinua. Como os primeiros usariam nascimentos como nível de observação e os segundos hospitais como nível de observação, evitaríamos imprecisões resultantes do processo de agregar os dados para construir o painel. Eles poderiam validar o que foi obtido aqui pelos métodos de pareamento.

Outra melhoria importante, embora computacionalmente mais complexa, seria incorporar distâncias na análise. No caso dos modelos de escolha discreta essa necessidade fica clara. Ocorre que mesmo para o painel municipal isso deveria ser

[71] Uma variável interessante que poderia ser adicionada e que foi discutida na literatura é o gênero. A literatura sugere que meninas são menos registradas que meninos.

levado em consideração. Isso porque os métodos de pareamento buscam na base municípios semelhantes como grupo de controle. Ocorre que, se esses municípios forem municípios vizinhos dos que receberam a unidade interligada, na prática, eles também se beneficiam de seus efeitos. Assim, ao comparar os dois, estaríamos subestimando o efeito das unidades. Usando distâncias poderíamos impor restrições nesse sentido.

Outra melhoria que deve ser investigada é o uso do número da serventia para identificar nascimentos. A lista de unidades interligadas da ARPEN e do CNJ tem o código dos cartórios. O SIRC também dispõe de códigos para os cartórios, e.g. número da serventia, embora não seja o mesmo. Se conseguirmos identificar os nascimentos que podem ter ocorrido em um hospital ou estabelecimento que tem unidade interligada e quais os cartórios, a análise pode ficar mais precisa. Isso deixaria a análise da regressão descontínua mais precisa, uma vez que poderíamos estimar o modelo para estabelecimentos de saúde com e sem unidades interligadas, embora os detalhes da implementação disso ainda tenham que ser estudados.

Há também melhorias que podem ser aplicadas ao processo de pareamento das bases. Podemos testar outros algoritmos de aprendizado de máquina, bem como uma regressão logística, para tentar melhorar o pareamento.

Por fim, há algumas melhorias na apresentação e na operacionalização do estudo. Em primeiro lugar, os modelos foram estimados usando STATA, por conveniência. Para esses modelos a sintaxe do STATA é simples, o retorno dos resultados é claro e a implementação dos métodos é consolidada. Essa parte do estudo poderia, contundo, ser feita em Python, assim como o pareamento. Assim, com mais tempo, seria possível e desejável realizar tudo usando Python apenas, por ser uma linguagem aberta.

Por fim, tanto esse relatório quanto as tabelas de resultados poderiam ser automatizados. O código, tanto em Python quanto em STATA poderia exportar automaticamente tabelas em LaTeX, o que facilitaria atualizações e correções do estudo, bem como a criação de outros relatórios ou apresentações derivadas.

Apoio:

Realização:

MINISTÉRIO DOS
DIREITOS HUMANOS
E DA CIDADANIA

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO