Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Avaliação de impacto do Programa Bolsa Família na inserção de jovens no mercado formal de trabalho por meio do método de Regressão com Descontinuidade (RDD)

Aloísio Dourado Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Rommel N. Carvalho

Coorientador
Prof. Dr. Donald Matthew Pianto

Brasília
2017

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

DAL453a

Dourado Neto, Aloísio
Avaliação de impacto do Programa Bolsa Família na inserção de jovens no mercado formal de trabalho por meio do método de Regressão com Descontinuidade (RDD) / Aloísio Dourado Neto; orientador Rommel N. Carvalho; co-orientador Donald Pianto. -- Brasília, 2017.
83 p.

Disserta??o (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2017.

1. Avaliação de Impacto de programas Sociais. 2. Bolsa Familia. 3. Regressão com Descontinuidade. I. N. Carvalho, Rommel, orient. II. Pianto, Donald, co orient. III. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Avaliação de impacto do Programa Bolsa Família na inserção de jovens no mercado formal de trabalho por meio do método de Regressão com Descontinuidade (RDD)

Aloísio Dourado Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Rommel N. Carvalho (Orientador)
CIC/UnB

Prof. Dr. Pedro Henrique Melo Albuquerque
ADM/UnB

Prof.a Dr.a Maristela Terto de Holanda
CIC/UnB

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 28 de Julho de 2017

Dedicatória

Dedico este trabalho a Mila, Luzinho, Ju e Lucas que suportaram com compreensão longos períodos de ausência durante as aulas, estudos, pesquisas, preparações de artigos, trabalhos em laboratório e elaboração dessa dissertação.

Agradecimentos

Agradeço aos orientadores, professores, colegas, familiares e amigos que de alguma forma contribuíram para que eu chegasse a esse ponto. Agradeço especialmente a meu Pai e a minha Mãe pelo incentivo que sempre deram para que eu nunca parasse de estudar. Agradeço também aos colegas da Setic/TCU, Geraldo e José Renato, pela compreensão e apoio no início dessa jornada e aos colegas da SecexPrevi/TCU Fábio, Melchior, Jorge e Ângelo pelo incentivo e importantes contribuições.

Resumo

Com a massificação do uso da tecnologia de informação e o recente aumento no compartilhamento de bases de dados entre as organizações, o TCU, órgão responsável pelo controle externo da União, passou a realizar auditorias com forte apoio de análises baseadas em dados. Entretanto, o volume de auditorias operacionais baseadas em dados governamentais para verificar a efetividade de políticas públicas ainda é baixo quando comparado às auditorias de conformidade de mesma natureza. Os principais dificultadores para a realização de análises de impacto a partir de bases governamentais são a complexidade inerente à manipulação de grandes bases de dados e a carência de softwares econométricos em código aberto capazes de realizar estimações com amostras da ordem de dezenas de milhões de observações. Visando contribuir para a mudança deste cenário, o presente trabalho disponibiliza um processo de integração de dados flexível e um *software* de estimação RDD capaz de realizar análises em bases de dados com dezenas de milhões de registros e alta variança, com boa performance, aproveitando os recursos de processamento paralelo dos computadores atuais. O presente trabalho demonstra que é possível avaliar a efetividade do Programa Bolsa Familia, uma importante política pública brasileira, utilizando bases de dados disponíveis no TCU, ao realizar uma análise do impacto do programa na inserção de jovens no mercado formal de trabalho, por meio da abordagem RDD, utilizando 13 milhões de registros extraídos de algumas dessas bases. Os resultados obtidos indicam que os jovens que permanecem no PBF por mais tempo apresentam menor nível de acesso ao mercado formal que aqueles que saem mais cedo do programa. Na avaliação realizada, foi obtida uma redução no tempo de processamento de 88% em relação ao software original. A abordagem de avaliação aqui apresentada não é restrita ao TCU e pode contribuir para a melhoria do processo de avaliação e fiscalização de programas sociais no Brasil se utilizada pelo órgão de controle interno da União ou pelos gestores da política pública.

Palavras-chave: Bolsa Família, Transferência Condicionada de Renda, RDD, Avaliação de Programas Sociais

Abstract

The *Big Data Era* has brought a huge amount of data to The Brazilian Federal Court of Accounts (TCU), which has used this data with great success in a variety of compliance audits. In contrast, the number of operational audits conducted with this new paradigm is still small. The main obstacles to carrying out impact analyzes from government bases are the complexity inherent in the manipulation of large databases and the lack of open source econometric software capable of making estimates with samples with tens of millions of observations. In order to help change this scenario, the present work evaluates the effectiveness of the *Bolsa Familia Program*, the main conditional cash transfer program in Brazil, by performing an impact analysis of the program's contribution to young Brazilian workers' access to the formal labor market, using governamental databases available at the TCU, using 13 million records extracted from some of these databases. This work provides a data integration workflow for governamental databases and a RDD estimation tool, capable of dealing with datasets with ten million observations or more and high variance, which takes advantage of modern computers' multi thread capabilities. The results indicate that young people who remain in the PBF for a longer time have a lower level of formal market access than those who leave the program earlier. The developed RDD tool presented a 88% better execution time than the original software. The expectation is that this work will contribute to the improvement of the process of evaluating and auditing social welfare programs in Brazil.

Keywords: Brazil's Bolsa Familia, Conditional Cash Transfer, Public Policies Evaluation, RDD

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CAGED Cadastro Geral de Empregados e Desempregados.

CGU Ministério da Transparência, Fiscalização e Controladoria-Geral da União.

CPF Código de Pessoa Física.

CPUs *Central Processing Units*.

DDD *Differences in Differences in Differences*.

DER Diagrama Entidade-Relacionamento.

DID *Differences in Differences*.

EAI *Enteprise Application Integration System*.

ENEM Exame Nacional do Ensino Médio.

ETL *Extract-Transform-Load*.

FIES Fundo de Financiamento Estudantil.

FRDD *Fuzzy Regression Discontinuity Design*.

GPU *Graphic Processing Unit*.

IBGE Instituto Brasileiro de Geografia e Estatística.

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

MDSA Ministério do Desenvolvimento Social e Agrário.

MR Modelo Relacional.

NIS Número de Identificação Social.

NIT Número de Identificação do Trabalhador.

PBF Programa Bolsa Família.

PIS Programa de Integração Social.

PLANFOR Plano Nacional de Qualificação do Trabalhador.

PNAD Pesquisa Nacional por Amostra de Domicílios.

PROGRESSA *Programa de Educación Salud y Alimentación.*

PROSPERA *Programa de Inclusión Social.*

PSM *Propensity Score Matching.*

RAIS Relação Anual de Informações do Ministério do Trabalho e Previdência Social.

RCM *Rubin Causal Model.*

RDD *Regression Discontinuity Design.*

SAGI Secretaria de Avaliação e Gestão da Informação.

SecexPrevi Secretaria de Controle Externo de Previdência e Assistência Social.

SRDD *Sharp Regression Discontinuity Design.*

TCU Tribunal de Contas da União.

Capítulo 1

Introdução

O presente capítulo tem por objetivo principal apresentar a motivação desse projeto de pesquisa. Inicialmente, é apresentado o contexto no qual o projeto se insere, dos pontos de vista de sua aplicabilidade no âmbito do Tribunal de Contas da União (TCU), de sua relação com o tema de avaliação de políticas públicas no Brasil e a forma como procura aproximar as áreas de econometria e ciência da computação, ao utilizar métodos econométricos de avaliação de programas, com apoio de técnicas de processamento paralelo e integração de bases de dados, comuns entre cientistas de dados, para resolver limitações de desempenho e dificuldades de acesso a dados dispersos. Na sequência, são apresentados os objetivos, a justificativa para o projeto, sua contribuição esperada e a estrutura do documento.

1.1 Contextualização

Nesta seção é apresentado, de forma resumida, o contexto no qual este trabalho se insere, tanto do ponto de vista da missão institucional do TCU quanto em relação à avaliação de políticas públicas e de programas sociais no Brasil. É feita também uma breve apresentação sobre a necessidade de aproximação entre as áreas de econometria, ciência de dados e tecnologia da informação.

1.1.1 Auditoria Operacional e Análise de Dados no TCU

O TCU é o órgão de controle externo brasileiro, responsável por julgar as contas dos administradores públicos e proceder à fiscalização contábil, financeira, orçamentária, operacional e patrimonial das unidades dos poderes e das demais entidades da União [1].

Desde a década de 90, de forma alinhada à reforma do estado e a gestão por resultados, o TCU vem desenvolvendo trabalhos que tratam da questão da eficiência, economicidade

e eficácia da gestão pública sob a modalidade de auditorias operacionais [2]. A modalidade de auditoria instituída no TCU em 1990, veio a complementar as auditorias de conformidade, focadas em questões legais, financeiras e contábeis, e, em maio de 1996, o TCU desenvolveu, em colaboração com a Fundação Getúlio Vargas do Rio de Janeiro, um programa de capacitação de seus técnicos com o objetivo de prover ao TCU os meios necessários à prática de avaliação de programas públicos orientada a resultados [2].

Mais recentemente, como forma de otimização do seu trabalho, o TCU passou a realizar auditorias de conformidade e operacionais com forte apoio de análises baseadas em dados conforme está formalizado em seu planejamento estratégico para o período de 2015 a 2021 [3]:

> "A atividade de controle externo tem como insumo e produto informação e conhecimento, elementos altamente dependentes de tecnologia da informação. O tratamento de dados e informações é condição indispensável para alavancar as atividades de controle.
>
> Dessa forma, desenvolver capacidade organizacional ampla para trabalhar com recursos tecnológicos emergentes e analisar grandes bases de dados é condição imprescindível para ampliar a capacidade de fiscalização e, consequentemente, dar resposta às demandas apresentadas ao TCU".

A Secretaria de Controle Externo de Previdência e Assistência Social (SecexPrevi) é uma das unidades do TCU que estão trabalhando intensivamente com auditoria baseada em dados. A secretaria dispõe de ampla massa de dados e, mesmo com equipe reduzida, realiza análises de conformidade sobre número significativo de benefícios previdenciários, trabalhistas e assistenciais, conforme está evidenciado nos acórdãos 718/2016 - Plenário [4], 1009/2016 - Plenário [5], 1181/2016 - Plenário [6], expedidos pelo TCU no ano de 2016. Esses acórdãos representam o resultado de trabalhos de verificação de conformidade, que analisaram aproximadamente R$424 bilhões em benefícios previdenciários [4], R$75 bilhões em benefícios assistenciais [5] e R$33 bilhões em benefícios trabalhistas [6] em 2015.

Apesar do ganho de produtividade que a auditoria por meio da análise de dados tem proporcionado, até o presente momento, o número de auditorias operacionais baseadas em dados para verificação da efetividade, eficiência ou eficácia de políticas públicas ainda é relativamente pequeno nessa unidade do Tribunal, quando comparado ao volume de auditorias de conformidade realizadas.

O presente projeto de pesquisa de mestrado tem por objetivo realizar uma análise de impacto do Programa Bolsa Família (PBF) na inserção de jovens no mercado formal de trabalho, utilizando bases de dados disponíveis no TCU, contribuindo, pelo exemplo, para a mudança daquele cenário de baixo volume de auditorias operacionais. Esse programa foi escolhido em razão do volume de recursos envolvidos, sua importância social e por

não possuir uma sistemática consistente de avaliação, conforme é detalhado a seguir. Tal proposta também está alinhada ao direcionamento estratégico do TCU, por tratar da aferição de eficácia de política pública (auditoria operacional) por meio de análise de bases de dados com utilização de ferramental tecnológico emergente, de forma inovadora, com ampla possibilidade de aplicação no âmbito da SecexPrevi.

1.1.2 Bolsa Família e Avaliação de Políticas Públicas no Brasil

A partir da Constituição Federal de 1988 [7], observou-se uma ampliação da cobertura da Política Social do Brasil [8]. O Programa Bolsa Família (PBF) é um dos principais exemplos do esforço governamental nesse sentido. O programa foi instituído pela Lei 10.836, de 9 de janeiro de 2004 [9], com objetivo de unificar as ações de transferência de renda do Governo Federal, especialmente as do Programa Nacional de Renda Mínima vinculado à Educação - Bolsa Escola[1], do Programa Nacional de Acesso à Alimentação - PNAA [2], do Programa Nacional de Renda Mínima vinculada à Saúde - Bolsa Alimentação[3] e do Programa Auxílio-Gás[4].

Segundo o Ministério do Desenvolvimento Social e Agrário (MDSA)[14], o PBF é o programa social que contribui para o combate à pobreza e à desigualdade no Brasil. Desde sua criação, o número de famílias atendidas pelo programa vem crescendo, refletindo no aumento do volume de recursos aplicados ao programa. Em 2015, conforme dados extraídos do Portal da Transparência[5] , foram distribuídos cerca de 75 bilhões de reais às mais de 14 milhões de famílias participantes.

Em um trabalho de 2001 [16], os Marinho e Façanha destacaram a importância da observação da efetividade, eficiência e eficácia das políticas públicas na avaliação da ação governamental. Apesar dessa importância e do grande volume de recursos destinados aos programas sociais, não se observa no Brasil uma sistemática de avaliação compatível com o volume de recursos aplicados conforme foi observado em artigo publicado em 2011 [8]. No trabalho, Jannuzzi afirmou que a sistemática de avaliação da ação governamental na área social vinha amadurecendo em ritmo insuficiente para acompanhar o crescimento dos recursos aplicados. O autor ainda destacou que grande parte do esforço empreendido para avaliar programas sociais no Brasil é baseado na realização de pesquisas de avaliação que frequentemente conduzem a constatações empíricas ambíguas ou pouco consistentes sobre a efetividade dos programas, gerando grande insatisfação com os estudos avaliativos.

[1]Lei nº 10.219, de 11 de abril de 2001[10]

[2]Lei nº 10.689, de 13 de junho de 2003[11].

[3]Medida Provisória nº 2.206-1, de 6 de setembro de 2001[12].

[4]O programa Auxílio-Gás foi extinto em 31 de dezembro de 2008 e absorvido pelo PBF[13]

[5]O portal da transparência [15] é um sítio mantido pelo Ministério da Transparência, Fiscalização e Controladoria-Geral da União (CGU), que disponibiliza dados sobre a execução financeira e orçamentária da União

Em artigo datado de 2005 [17], Jannuzzi já havia chamado a atenção para a imensa quantidade de dados disponíveis e como tais dados poderiam ser utilizados na avaliação das políticas públicas para o direcionamento da ação governamental. Mourão e Laros [18] também levantaram questionamentos sobre a utilização de pesquisas como única fonte de dados para avaliação de programas sociais e propuseram um método de avaliação baseado em análises estatísticas de indicadores sociais para o Plano Nacional de Qualificação do Trabalhador (PLANFOR).

Se, por um lado, a avaliação por meio de pesquisas pode não conduzir a resultados satisfatórios, por outro lado, a realização de experimentos estatísticos de avaliação de programas sociais, que normalmente produzem os melhores resultados, não é eticamente possível, pelo menos na maioria dos casos. A avaliação experimental de programas sociais, assim como as avaliações de medicamentos na indústria farmacêutica, requer a seleção aleatória de grupos para receberem e não receberem o tratamento (grupo de controle). A avaliação é feita por meio da comparação estatística dos resultados obtidos pelos dois grupos. No caso de programas sociais de transferência de renda como o PBF, a aplicação de experimentos exigiria a seleção aleatória de indivíduos ou famílias para receberem e não receberem o benefício, o que certamente geraria grande insatisfação entre os não beneficiados.

Essa questão é discutida por Rubin em seu trabalho de 1974 [19]. O autor defende que os experimentos aleatórios devem ser preferidos, sempre que possível. Entretanto, o autor também defende que técnicas quasi-experimentais [20], quando bem aplicadas, geralmente conduzem a resultados satisfatórios sem suscitar questões éticas como os métodos experimentais. Essa questão também é abordada no trabalho de avaliação do *Programa de Educación Salud y Alimentación (PROGRESSA)* [21] do México[6], tendo obtido resultados comparáveis aos do método experimental com a técnica *Regression Discontinuity Design (RDD)* [22].

O Decreto 5.209 [23] que regulamenta o PBF estabelece que dentre os objetivos principais do programa está estimular a emancipação sustentada das famílias que vivem em situação de pobreza e extrema pobreza. A condição de auto-suficiência é alcançada por meio das ditas portas de saída[7], dentre as quais destaca-se o acesso ao mercado de trabalho [24]. Além disso, um dos focos do programa são jovens entre 16 e 17 anos, que possuem, inclusive, um benefício específico, o benefício variável vinculado ao adolescente [9]. Assim sendo, uma avaliação de impacto que considere o ingresso no mercado formal por

[6]Em setembro de 2014, o PROGRESSA mexicano foi reformulado e passou a se chamar *Programa de Inclusión Social (PROSPERA)* (`https://www.prospera.gob.mx/swb/es/PROSPERA2015/Quees_PROSPERA`).

[7]Porta de Saída é uma expressão utilizada para representar os mecanismos de superação da condição de pobreza ou extrema pobreza e emancipação dos beneficiários em relação ao programa.

parte de jovens que tenham recebido o benefício tende a proporcionar uma boa medida de efetividade do programa.

O impacto do PBF sobre os jovens tem despertado interesse de alguns pesquisadores, mas tais pesquisas se concentram na questão da escolaridade. Por exemplo, de Brauw et al. [25] usaram dados de pesquisas relizadas em 2005 e 2009 pelo Centro de Desenvolvimento e Planejamento Regional para avaliar o impacto do programa na permanência dos jovens na escola, utilizando uma amostra de 15.426 famílias. Reynolds [26] também avaliou o impacto do PBF sobre a escolaridade, lançando mão de dados obtidos pela Pesquisa Nacional por Amostra de Domicílios (PNAD), envolvendo aproximadamente 150.000 famílias. O impacto do PBF sobre o trabalho formal também já foi estudado. Barbosa e Courseil [27] usaram dados da PNAD para avaliar o impacto do programa sobre a opção pelo trabalho formal ou informal, mas o foco do estudo foram os adultos, e não os jovens.

Outro aspecto comum observado nos trabalhos relacionados foi a utilização de dados provenientes de pesquisas com número limitado de amostras quando comparado ao universo de famílias beneficiadas pelo programa, como em [25] (15.426 amostras), [26] (150.000 amostras) e [27](145.547 amostras).

Além de apresentar um tema pouco explorado, que é o impacto de médio prazo do programa sobre o acesso dos jovens ao mercado formal, o presente projeto de pesquisa introduz uma inovação: a utilização de ampla massa de dados governamentais sobre o Bolsa Família, incluindo o Cadastro Único e a Folha de Pagamentos além de dados sobre o mercado formal da Relação Anual de Informações do Ministério do Trabalho e Previdência Social (RAIS), todos disponíveis no TCU, associada à aplicação da técnica quasi-experimental RDD de avaliação de impacto.

1.1.3 A necessidade de aproximação entre a econometria e a tecnologia de informação

Como já mencionado, métodos quasi-experimentais, a exemplo do RDD exercem um importante papel na avaliação de programas sociais [19]. Apesar da importância e do sucesso da abordagem [21] [28] [27] [29] [30] [31] [32], seu uso, evolução e provimento de ferramental tem estado restrito a economistas e estatísticos envolvidos com econometria, como pode ser verificado pela formação e atuação dos autores dos trabalhos aqui citados. De acordo com H. R. Varian [33], Economista Chefe da Google, historicamente, os economistas foram acostumados a lidar com massas de dados que cabem em uma planilha. Em um ensaio mais detalhado sobre este tema, Einav [34] afirma que aprender a lidar com grandes bases de dados é um dos grandes desafios para os economistas modernos. Tomando por exemplo o RDD, embora vivamos em uma era de grande produção de dados, algumas

das ferramentas disponíveis para realizar as análises não estão preparadas para trabalhar com grandes massas de dados. No mesmo trabalho citado anteriormente, ao discorrer sobre a necessidade de aproximação entre os economistas e os cientistas de dados, Varian [33] afirma que ferramentas e técnicas de manipulação de dados desenvolvidas para pequenas amostras por economistas se tornarão cada vez mais inadequadas para lidar com problemas que os pesquisadores de aprendizagem de máquina já estão acostumados a enfrentar. Como exemplo do distanciamento que existe entre as duas áreas, podemos citar os pacotes de estimação RDD de código aberto disponíveis em linguagem de programação R [8], conforme levantamento e análise comparativa realizada em 2016 [36]: os pacotes *rdd* [37], *rdrobust* [38] e *rddtools* [39]. Apesar de estarem atualizados com as últimas técnicas de estimação e análise RDD, nenhum dos três pacotes possui capacidade de processamento paralelo, sendo, portanto, incapazes de aproveitar os recursos de processamento *multithread* dos computadores atuais. A capacidade de processamento em paralelo torna-se especialmente mais importante, à medida em que o tamanho da base aumenta, em função da possibilidade de redução dos tempos de processamento. Cumpre destacar que, conforme relatório técnico publicado em 2014 pelo *National Bureau of Economic Research* [40], as linguagem de programação R, $MATLAB$ [9] e $Mathematica$ [10] são as linguagens de *script* mais conhecidas por economistas, sendo que o R é a única *opensource*. Assim sendo, os três pacotes R aqui citados seriam as principais opções de estimação RDD para os economistas que preferem trabalhar com ferramentas livres.

Justin Grimmer [43], em artigo de 2014 também afirma que Cientistas Sociais e Cientistas de Dados podem se beneficiar de trabalhar juntos quando estiverem lidando com inferência causal em um cenário de Big Data. Entretanto, o autor acrescenta uma análise do ponto de vista contrário, ao afirmar que assim como os economistas não têm o hábito de lidar com grandes bases de dados, os cientistas de dados, via de regra, não estão habituados com métodos de inferência causal. O autor ainda destaca as diferenças que existem entre os métodos usuais de aprendizagem de máquina e os experimentos de inferência causal, deixa claro que uma abordagem não substitui a outra, mas apresenta os benefícios que trabalhos combinados podem apresentar que vão além da incorporação da habilidade de lidar com grandes massas de dados pelos métodos de inferência causal.

O presente trabalho procura, então, aproximar estes dois universos, lançando mão de ferramentas de integração de dados para lidar com a heterogeneidade dos ambientes e

[8] R [35] é uma linguagem e ambiente para computação estatística e desenvolvimento de gráficos de alta qualidade para publicação.

[9] $MATLAB$ [41] é uma plataforma otimizada para resolver problemas científicos e da área de engenharia que possui uma linguagem de programação baseada em matrizes adequada para representar expressões matemáticas em ambiente computacional.

[10] $Mathematica$ [42] é um ambiente integrado para computação técnica que prove uma interface que possibilita organizar texto, código e gráficos em um único documento.

técnicas de processamento paralelo adequadas a altos volumes de dados em um algoritmo de análise de impacto por RDD.

1.2 Objetivos

O objetivo geral do projeto de pesquisa é realizar uma avaliação de impacto do PBF no que tange à sua contribuição para o acesso de jovens beneficiários ao mercado formal de trabalho por meio da abordagem quasi-experimental RDD, utilizando a totalidade dos dados aplicáveis das bases governamentais sobre trabalho e bolsa família, disponíveis no TCU de forma dispersa e heterogênea. Os objetivos específicos desse trabalho são os seguintes:

- integrar em um único ambiente as bases de dados governamentais Cadastro Único, Folha de Pagamentos do PBF e RAIS, disponíveis no TCU de forma dispersa em ambientes heterogêneos e preparar os dados para estimação RDD;
- desenvolver e disponibilizar para a comunidade *opensource* uma versão de *software* de estimação RDD capaz de manipular grandes massas de dados com bom desempenho, lançando mão de recursos de paralelismo em *multi-thread* disponíveis nos computadores atuais;
- quantificar o impacto do PBF no acesso de jovens ao mercado formal de trabalho por meio da abordagem RDD, usando como fontes de informação grandes bases de dados disponíveis no TCU como o Cadastro Único, a Folha de Pagamentos do PBF, e a RAIS;
- verificar se existe variação dos impactos entre subpopulações considerando critérios de distribuição regional, local de residência (urbana ou rural) e gênero.

1.3 Justificativa do Tema

Conforme apresentado na Seção 1.1.2, a avaliação de programas sociais, embora seja um tema de extrema importância para direcionar corretamente a aplicação de recursos públicos, ainda é pouco realizada no Brasil. Visando suprir esta carência, o presente trabalho apresenta uma proposta de avaliação de impacto do PBF, um dos mais importantes programas de transferência condicionada de renda do mundo, com foco nos jovens, um dos principais alvos do programa.

Adicionalmente, o tema proposto vem suprir uma carência que hoje existe na SecexPrevi que é a realização de auditorias operacionais baseadas em dados. A auditoria

baseada em dados já se mostrou um importante instrumento de aumento de produtividade no controle externo, entretanto, o volume de auditorias operacionais baseadas em dados ainda é pequeno quando comparado às auditorias de conformidade da mesma natureza.

Subsidiariamente, este trabalho ainda aborda a importância da aproximação entre as áreas economia e ciência de dados, como indicado em [33] e [43].

Por fim, o tema do presente projeto de pesquisa também está alinhado ao planejamento institucional do TCU e tem possibilidade de aplicação prática na SecexPrevi sem requerer investimentos em pesquisas de avaliação.

1.4 Contribuição Esperada

Com o presente trabalho, pretende-se gerar contribuições em duas áreas: melhoria do processo de avaliação e fiscalização de programas sociais no Brasil, e aproximação entre as áreas de economia e tecnologia da informação por meio da disponibilização de um processo de avaliação de impacto que envolve utilização de ferramental de integração de dados e utilização de processamento paralelo para estimação RDD em grandes volumes de dados.

No que se refere ao processo de avaliação e fiscalização de programas sociais, com o presente trabalho, pretende-se disponibilizar um modelo de aplicação do método RDD totalmente baseado em dados preexistentes na Administração Pública, que não requer investimentos em pesquisas de avaliação. Apesar de ter sido originalmente pensado para atender uma necessidade do TCU, o modelo proposto também pode ser aplicado por outros órgãos. No TCU, o modelo poderá ser incorporado à sistemática de fiscalização contínua [44] em implementação na SecexPrevi[11], contribuindo para a identificação de oportunidades de melhoria na gestão do programa, em sintonia com a missão institucional do TCU. Tal uso do modelo pode também ser feito igualmente pelo órgão de controle interno da União, no exercício de sua atribuição. Adicionalmente, o modelo proposto pode ser utilizado como ferramenta de avaliação sistemática da ação governamental pelo órgão gestor da política pública como instrumento de melhoria contínua e otimização da aplicação de recursos. Além disso, o modelo pode ser facilmente adaptado para outros programas.

Acerca da aproximação entre a Economia e a Tecnologia da Informação, apresentaremos como o ferramental de integração de dados, muito comum entre os profissionais da computação, pode ser usado em um processo de avaliação de impacto a partir de bases heterogêneas. Adicionalmente, será apresentado como a adoção de técnicas de proces-

[11] Para mais detalhes sobre o projeto de fiscalização contínua da SecexPrevi consulte `http://portal.tcu.gov.br/inovatcu/projetos/fiscalizacao-continua-na-secexprevi.shtml`

samento paralelo pode contribuir para a redução de tempos de execução dos pacotes de estimação de RDD disponíveis atualmente em linguagem *R* [36], que, conforme resultados de pesquisa recente [40], é a linguagem de *script* mais conhecida por economistas na modalidade *open source*.

Como as ferramentas *open source* atualmente existentes não suportam processamento paralelo, pretende-se também contribuir com a disponibilização de um pacote *R* de análise RDD com suporte a *multi-thread* adaptado para grandes bases de dados.

1.5 Estrutura do Documento

O presente trabalho está estruturado em cinco capítulos:

- o Capítulo 1 corresponde a essa introdução;
- no Capítulo 2 é apresentada a fundamentação teórica do presente projeto, bem como uma relação de trabalhos correlatos;
- no Capítulo 3 é apresentada a metodologia utilizada e a solução proposta;
- no Capítulo 4 são apresentados os resultados obtidos;
- no Capítulo 5 são apresentados as conclusões e trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Nesse capítulo será apresentada a fundamentação teórica do presente projeto de pesquisa. Na Seção 2.1 será apresentada a teoria relativa à avaliação de efetividade de programas sociais, com ênfase na avaliação de impacto por meio do método quasi-experimental *Regression Discontinuity Design (RDD)*. Na sequência, será apresentada a fundamentação teórica sobre as técnicas de integração de dados (Seção 2.2) e paralelismo (Seção 2.3) utilizadas na solução desenvolvida. O capítulo é encerrado na Seção 2.4 com um levantamento de trabalhos correlatos e sua relação com o presente trabalho.

2.1 Avaliação de Impacto de Políticas Públicas

A análise de impacto tem por objetivo mensurar os efeitos do programa sobre a população-alvo além de estabelecer uma relação de causalidade entre a política e os efeitos sobre a sociedade [45]. Arretche [46] usa a definição anterior como sinônimo de avaliação de efetividade de programas sociais e diferencia os conceitos de efetividade e eficácia usando como exemplo um programa de vacinação: um determinado programa de vacinação pública pode ter sido eficaz ao atingir as metas de quantidade de pessoas vacinadas, entretanto, o programa pode não ter sido efetivo ao não conseguir reduzir o nível de incidência da doença que se pretendia erradicar sobre a população. Avaliação de efetividade ou impacto é, portanto, a verificação do alcance dos objetivos de mais alto nível do programa. Ainda segundo o autor, a principal dificuldade metodológica de uma avaliação de impacto reside na necessidade de demonstrar que os resultados encontrados possuem relação de causalidade com a política pública [46].

Ao buscar avaliar o impacto de um programa ou de um tratamento qualquer, o que se espera é medir a diferença entre os resultados que seriam obtidos por uma determinada unidade caso ela fosse submetida ao tratamento e caso ela não fosse tratada [47], sendo que uma unidade pode ser qualquer entidade sobre a qual deseja-se aferir o resultado

do tratamento, por exemplo, um indivíduo, um grupo de pessoas ou uma organização. Buscando estabelecer um método formal para avaliação de impacto, em seu trabalho seminal de 1974, Rubin [19] formalizou o conceito de resultados potenciais, conforme será descrito nessa subseção. Seja $Y_i(0)$ o resultado potencial da unidade i na ausência do tratamento e $Y_i(1)$ o resultado potencial da unidade i na presença do tratamento. Para a unidade i, o impacto ou efeito do tratamento seria $Y_i(1) - Y_i(0)$. O problema é que não é possível observar simultaneamente $Y_i(1)$ e $Y_i(0)$. A alternativa mais óbvia para se enfrentar esta questão é a realização de experimentos aleatórios, nos quais unidades são aleatoriamente selecionadas para receberem ou não o tratamento. Supondo que ambos os grupos tenham M elementos, o efeito médio do tratamento corresponde então à média da diferença dos resultados entre tratados e não tratados conforme a equação 2.1.

$$\frac{1}{M}\sum_{i=1}^{M}[Y_i(1) - Y_i(0)] \tag{2.1}$$

O framework proposto por Rubin em 1974 [19] ficou posteriormente conhecido na literatura como *Rubin Causal Model (RCM)* [48]. No mesmo trabalho de 1974, o autor também defendeu que, embora a seleção aleatória seja ideal para eliminar o efeito de viés de seleção, nem sempre é possível a realização de experimentos aleatórios para a determinação do efeito médio do tratamento, especialmente na área das ciências sociais [19]. Seja, por exemplo, o caso da implantação do Programa Bolsa Família (PBF). Caso a escolha das famílias para participarem do programa de transferência de renda fosse aleatória, certamente inúmeras críticas seriam levantadas. Haveria grande insatisfação entre os não selecionados e questões éticas seriam suscitadas. Outro aspecto que dificulta a realização de experimentos aleatórios para avaliação de programas é a impossibilidade de aplicação do método em programas que se encontram em fase avançada de implantação. Mais uma vez, no caso do PBF, dado seu atual estágio de universalização, é impossível a seleção aleatória de grupos de controle e tratamento para a realização de um experimento. Quando tal situação ocorre, muitos pesquisadores lançam mão do que se costuma chamar na literatura de métodos quasi-experimentais. Campbell [20] define os quasi-experimentos como sendo os desenhos de pesquisas sociais nos quais os pesquisadores utilizam algum aspecto dos verdadeiros experimentos nos seus procedimentos de coleta de dados, mesmo não tendo total controle sobre os estímulos experimentais que se obtém com a seleção aleatória. O autor também destaca que como o completo controle experimental é ausente, é imperativo que o pesquisador tenha total consciência sobre as variáveis que ele está falhando em controlar, de modo a assegurar a validade das conclusões [20]. Mesmo considerando as preocupações levantadas por Campbell [20], Rubin [19], apesar de reconhecer que os reais experimentos são preferíveis sempre que aplicáveis, defende que experimen-

tos não aleatórios (quasi-experimentos), quando aplicados corretamente, podem levar a resultados equivalentes ou muito próximos aos do método experimental.

Corroborando com o pensamento de Rubin acerca dos quasi-experimentos [19], um fato cientificamente interessante ocorreu na implantação do programa de transferência de renda Progressa do México[1]. Por decisão dos gestores, a seleção dos municípios para participarem inicialmente do programa se deu de forma aleatória [21] e isso possibilitou a realização de experimentos para avaliar o impacto do programa. Alguns pesquisadores aproveitaram a oportunidade para comparar os resultados obtidos pelos métodos experimentais com alguns métodos não experimentais. Por exemplo, Buddelmeyer e Skoufias [21] compararam o método não experimental *Regression Discontinuity Design (RDD)* com o método experimental e concluíram que os resultados foram muito próximos. Em contrapartida, Diaz e Handa [49] aproveitaram a oportunidade para avaliar o método *Propensity Score Matching (PSM)*, mas diferente de Buddelmeyer e Skoufias, não chegaram a resultados satisfatórios.

Alguns dos métodos quasi-experimentais [20] mais utilizados em avaliações de programas sociais serão apresentados com mais detalhes nas próximas subseções.

2.1.1 Diferenças em Diferenças

O estimador *Differences in Differences (DID)* é uma das mais populares ferramentas quasi-experimentais de pesquisa aplicada em economia para avaliar o efeito de políticas públicas [50]. O método pode ser aplicado quando existem dados antes e após o tratamento e pode-se observar as diferenças pré-existentes entre os grupos de tratamento e controle. O método é relativamente recente, tem por base uma ideia simples e foi inicialmente introduzido em 1994 por Card e Kruegger [51]. Os autores analisaram os impactos do aumento do salário mínimo no estado americano de Nova Jersey na oferta de empregos na indústria de *fast-food*. Como grupo de controle, foi utilizado o estado vizinho da Pensilvânia que não teve aumento no salário mínimo. O trabalho consistiu da análise das diferenças nas variações dos indicadores nos dois estados. Com os resultados obtidos, os autores concluíram pela não confirmação da hipótese original de que o aumento do salário minimo implicaria em redução na oferta de empregos. De acordo com Abadie [50] o DID usa o conceito de resultados potenciais do modelo RCM [19] já mencionado anteriormente, da forma descrita a seguir.

Seja $Y_i(t)$ o resultado potencial da unidade i no tempo t. Considere também que $t = 0$ indica o período anterior ao início do tratamento, ao passo que $t = 1$ representa

[1]Em setembro de 2014, o PROGRESSA mexicano foi reformulado e passou a se chamar *Programa de Inclusión Social (PROSPERA)* (`https://www.prospera.gob.mx/swb/es/PROSPERA2015/Quees_PROSPERA`).

o período posterior ao tratamento. Seja também $D_i(t)$ o indicador de tratamento da unidade i de forma que $D_i(t) = 1$ representa uma unidade tratada e $D_i(t) = 0$ uma unidade não tratada. Desta forma, supondo que as unidades tratadas e não tratadas seguiriam tendências paralelas na ausência do tratamento, a esperança da diferença dos resultados nos momentos final e inicial condicionada à presença do tratamento é dada pela equação 2.2, e, de modo análogo, a esperança da diferença dos resultados nos momentos final e inicial condicionada à ausência do tratamento é dada pela equação 2.3.

$$\mathbb{E}\left[Y(1) - Y(0)|D(1) = 1\right] \tag{2.2}$$

$$\mathbb{E}\left[Y(1) - Y(0)|D(1) = 0\right] \tag{2.3}$$

Assim sendo, a estimativa DID do efeito médio do tratamento é dada pela equação 2.4, que corresponde à diferença entre as esperanças 2.2 e 2.3. A equação 2.5 corresponde a um simples reagrupamento dos termos da equação 2.4, de modo a evidenciar que o viés de seleção $\mathbb{E}\left[Y(0)|D(1) = 1\right] - \mathbb{E}\left[Y(0)|D(1) = 0\right]$ está sendo subtraído da diferença entre os grupos de tratados e não tratados.

$$\mathbb{E}\left[Y(1) - Y(0)|D(1) = 1\right] - \mathbb{E}\left[Y(1) - Y(0)|D(1) = 0\right] \tag{2.4}$$

$$\{\mathbb{E}\left[Y(1)|D(1) = 1\right] - \mathbb{E}\left[Y(1)|D(1) = 0\right]\} - \{\mathbb{E}\left[Y(0)|D(1) = 1\right] - \mathbb{E}\left[Y(0)|D(1) = 0\right]\} \tag{2.5}$$

Em seu trabalho, Abadie [50] ainda apresenta os principais conceitos sobre DID, formaliza o método, fornece uma boa quantidade de referências e apresenta uma família de estimadores não paramétricos.

Uma extensão do conceito DID é o método das triplas diferenças, em inglês *Differences in Differences in Differences (DDD)*, apresentado por Gruber em 1994 [52]. Bastante similar ao DID, o DDD permite a utilização de dois grupos de controle para aumentar a robustez do método. Suponha que se deseja avaliar uma política de saúde estadual destinada a idosos acima de 65 anos, em um determinado estado da federação, e que se disponha de dados antes do início do programa. Um possível grupo de controle seriam os indivíduos do mesmo estado com menos de 65 anos, e, portanto, não afetados pela nova política. Entretanto, políticas federais também podem afetar a saúde dos idosos e comprometer o resultado da análise DID. Uma alternativa seria usar como grupo de controle, os idosos de um outro estado. Mais uma vez, mudanças na saúde dos idosos podem ser sistematicamente diferentes entre os estados, por exemplo em razão de diferenças nos níveis de renda. O método DDD permite a obtenção de estimadores mais robustos ao possibilitar a utilização dos dois grupos de controle simultaneamente.

Como exemplo de aplicação recente do método DDD na avaliação de programas sociais, é possível citar o trabalho de Reynolds [26], em 2015. A autora utiliza dados do Instituto Brasileiro de Geografia e Estatística (IBGE) e da Pesquisa Nacional por Amostra de Domicílios (PNAD) para avaliar a ampliação da cobertura dos benefícios do PBF para adolescentes, que antes era limitada à idade de 15 anos e que passou para 17 anos em 2008. A autora conclui que a ampliação do programa contribuiu para a manutenção na escola dos adolescentes que já estavam estudando aos 15 anos, entretanto não foi capaz de reverter a situação daqueles que abandonaram a escola aos 16 anos antes de 2008.

2.1.2 *Propensity Score Matching*

O *Propensity Score Matching (PSM)* é uma abordagem quasi-experimental que foi inicialmente proposta por Rosenbaun e Rubin em 2 trabalhos subsequentes, [53] e [54]. A ideia fundamental por trás do método é, dado um conjunto de características pré-tratamento, encontrar entre os não tratados, indivíduos que sejam similares aos tratados. Uma vez encontrado tal grupo (controle), diferenças nos resultados obtidos pelos grupos de controle e tratamento podem ser atribuídas ao tratamento.

Após sua introdução, a literatura acerca do método continuou a evoluir. Hirano et al. [55] apresentaram um estimador Horvitz-Thompson baseado em ponderação pelo inverso da probabilidade de atribuição ao tratamento. Abadie e Imbens [56] propuseram um método para correção assintótica de viés nos estimadores e um novo estimador para a variância condicional. Hirano e Imbens [57] introduziram um estimador flexível para construir pesos e usaram estes pesos em uma regressão ponderada do resultado.

Mesmo com sua evolução, o método PSM mantém a forte pressuposição de que as variáveis que influenciam simultaneamente a designação do tratamento e os resultados potenciais são observadas pelo pesquisador. A afirmação anterior pode ser confirmada no trabalho de Heckman et al. [58]. Os autores demonstraram que o estimador é muito sensível à qualidade do modelo de *matching* utilizado pelo pesquisar e demonstraram que a retirada de variáveis importantes do modelo aumenta significativamente o viés das estimativas.

Apesar das fortes pressuposições envolvidas, o *Propensity Score Matching* (PSM) é bastante usado na literatura. Caliendo e Kopeinig [59] oferecem uma boa revisão da literatura acerca do método, além de apresentarem um guia prático para sua aplicação por pesquisadores. Conforme já citado na abertura da Seção 2.1, Diaz e Handa [49] aproveitaram-se das características da implantação do PROGRESSA no México para fazer uma avaliação do método PSM. Os autores concluíram que o método não se mostrou satisfatório no caso do programa de transferência de renda do México, apresentando resultados enviesados em relação ao método experimental. Por outro lado, Brawn et al. [25]

aplicaram a abordagem de Hirano et al. [55] em uma avaliação de impacto do Programa Bolsa Família (PBF) nos resultados escolares de crianças entre 6 e 17 anos e foram capazes de confirmar suas estimativas por meio de análises empíricas.

2.1.3 RDD

O *Regression Discontinuity Design (RDD)* é uma abordagem quasi-experimental de avaliação do efeito de intervenções proposta inicialmente por Thistlethwaite e Campbell em 1960 [22] que busca evitar o efeito do viés de seleção explorando alguma descontinuidade que exista no critério de elegibilidade do programa. Naquele trabalho seminal, Thistlethwaite e Campbell avaliaram o efeito de um programa de concessão de bolsas de estudo, aproveitando o fato de que os prêmios eram concedidos apenas aos alunos que obtinham nota acima de um determinado valor em um teste de seleção. O princípio básico por trás do método baseia-se no fato de que alunos com notas próximas do ponto de corte seriam similares entre si e que a diferença da média dos resultados futuros entre os que que ficaram logo acima do ponto de corte e os que ficaram logo abaixo do ponto de corte poderia ser atribuída à concessão da bolsa. O método não chamou muita atenção até os anos 90 [47] quando então começaram a surgir na literatura alguns trabalhos de aplicação como Angrist e Lavy [60] que estimaram o efeito do tamanho das turmas no resultado acadêmico dos alunos e Black [61] que avaliou a disposição dos pais de pagar mais por melhores escolas ao comparar preços das casas próximas a diferentes escolas. Segundo Hanh, Todd e Van der Klaauwn [31], a grande vantagem do RDD frente a outras abordagens quasi-experimentais é que ela torna irrelevantes algumas questões acerca da especificação do modelo, como a escolha de variáveis para incluir no modelo e a identificação de sua forma funcional, como detalharemos a seguir.

A Figura 2.1 ilustra o princípio básico do método. Na ausência do tratamento, a distribuição do resultado em função de X é suave. Na presença do tratamento, a existência de uma descontinuidade no ponto de corte indica o efeito causal do tratamento. O efeito médio do tratamento no ponto de corte é dado pela amplitude dessa descontinuidade. Por estar diretamente relacionada ao presente projeto de pesquisa, a abordagem RDD será apresentada com um pouco mais de detalhes nas próximas subsessões, incluindo a apresentação de alguns avanços recentes acerca do método, presentes na literatura. Para detalhamento do método, utilizaremos a notação matemática utilizada por Imbens e Lemieux em [47]. Além disso, o trabalho de Imbens e Lemieux [47] também apresenta um guia para aplicação prática do método e uma boa relação de referências.

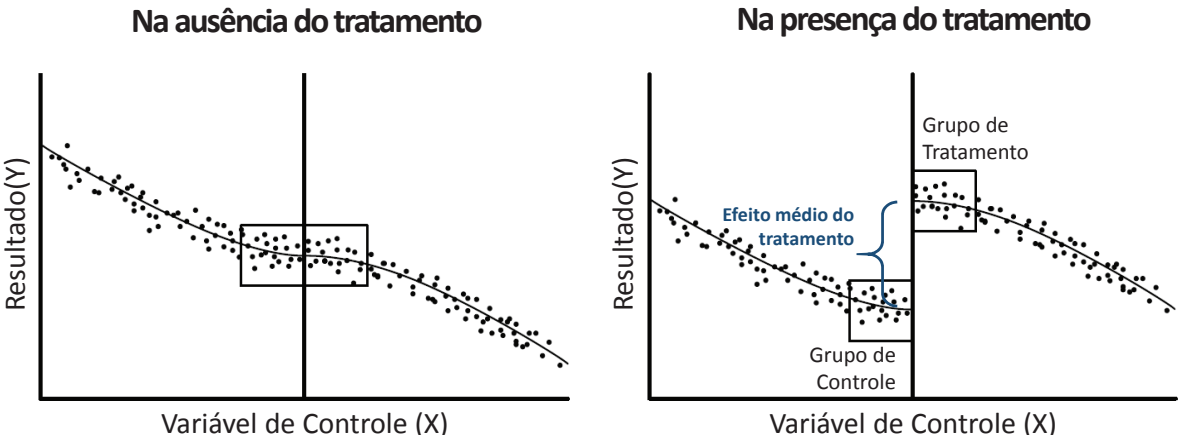

Figura 2.1: Ilustração do princípio do método RDD (adaptado) (Fonte: [62]).

Sharp RDD versus Fuzzy RDD

No RDD, a elegibilidade para participação no programa é determinada pelo valor da covariável X_i (conhecida na literatura como *running variable* e traduzida para variável de controle nesse trabalho). O tipo de relação entre a variável de controle e a probabilidade das unidades receberem o tratamento determina o tipo de abordagem a ser seguida. No *Sharp Regression Discontinuity Design (SRDD)* a probabilidade salta de 0 para 1 no ponto de corte. No *Fuzzy Regression Discontinuity Design (FRDD)* a probabilidade pode saltar de um valor maior que zero para um valor menor que 1. Conforme a notação utilizada em [47], seja $W_i \in \{0, 1\}$, a variável que representa a presença do tratamento na unidade i, onde $W_i = 1$ quando a unidade foi tratada e $W_i = 0$ em caso contrário. O resultado obtido pela unidade i pode ser expresso pela equação 2.6.

$$Y_i = (1 - W_i) \cdot Y_i(0) + (W_i) \cdot Y_i(1) = \begin{cases} Y_i(0) \mid W_i = 0 \\ Y_i(1) \mid W_i = 1 \end{cases} \tag{2.6}$$

Ou seja, se X_i está acima de um determinado ponto de corte, a unidade i participa do programa ($W_i = 1$), caso contrário, não ($W_i = 0$). O exemplo anterior corresponde ao modelo *Sharp* RDD, onde o tratamento é totalmente determinado pela covariável (X_i). Entretanto, a participação no tratamento ou programa pode ser apenas parcialmente determinada pela variável de controle. Nesse caso, tem-se o que é chamado na literatura de *Fuzzy* RDD. Na Figura 2.2, a diferença entre as variações *Sharp* e *Fuzzy* é ilustrada. Em ambos os casos, se a distribuição de X_i é suave e contínua, qualquer descontinuidade na distribuição condicional do resultado como uma função de X_i pode ser interpretada como uma evidência de um efeito causal do tratamento.

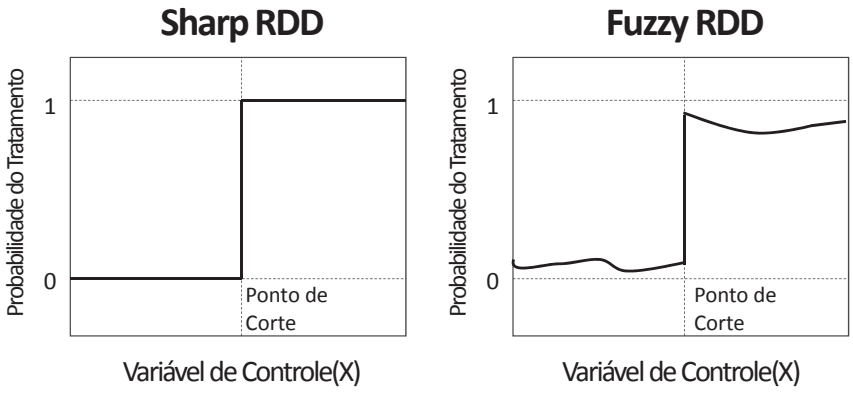

Figura 2.2: Sharp e Fuzzy RDD.

No SRDD, a atribuição do tratamento W_i é uma função determinística da variável de controle X_i conforme a expressão 2.7.

$$\begin{cases} W_i = 1 \mid X_i \geq c \\ W_i = 0 \mid X_i < c \end{cases} \tag{2.7}$$

Assim sendo, a esperança condicional do resultado que representa o efeito médio do tratamento no ponto de corte é dado pelo estimador apresentado na equação 2.8 [47], sendo que Y representa o resultado de interesse, X representa a variável de controle e c é o ponto de corte.

$$\tau_{SRD} = \lim_{x \to c^+} \mathbb{E}[Y \mid X = x] - \lim_{x \to c^-} \mathbb{E}[Y \mid X = x] \tag{2.8}$$

Na equação 2.8, $\lim_{x \to c^+} \mathbb{E}[Y \mid X = x]$ representa o limite da esperança condicional do resultado do grupo de tratamento quanto a variável de controle está logo acima de c. De modo análogo, $\lim_{x \to c^-} \mathbb{E}[Y \mid X = x]$, é o limite da esperança condicional do resultado do grupo de controle, quando X está logo abaixo de c. Em outras palavras, este estimador corresponde à diferença da esperança condicional do resultado no ponto de corte, à direita e à esquerda.

No FRDD, é possível haver uma mudança menor na probabilidade do tratamento no ponto de corte. Por exemplo, seja um determinado programa social no qual pessoas acima de determinada idade (ponto de corte) são elegíveis, mas a efetiva participação no programa é opcional. Alguns indivíduos elegíveis podem optar por não participar do programa. neste caso, a probabilidade saltaria de zero para um número menor que um. Considerando esta probabilidade, obtêm-se o seguinte estimador representado ela equação 2.9 [47].

$$\tau_{FRD} = \frac{\lim_{x \to c^+} \mathbb{E}[Y \mid X = x] - \lim_{x \to c^-} \mathbb{E}[Y \mid X = x]}{\lim_{x \to c^+} \mathbb{E}[W \mid X = x] - \lim_{x \to c^-} \mathbb{E}[W \mid X = x]} \tag{2.9}$$

Este estimador (τ_{FRD}) é bastante similar ao τ_{SRD}, mas, no denominador, ele considera a descontinuidade na probabilidade da atribuição ao tratamento (W).

Abordagens paramétricas *versus* não paramétricas

No método RDD, a avaliação do impacto no ponto de corte passa pela determinação dos limites presentes nas expressões de τ_{SRD} ou τ_{FRD} vistas anteriormente. Segundo Jacob et al. [62], os métodos disponíveis para calcular estes limites incluem abordagens paramétricas e não paramétricas, entretanto, as abordagens paramétricas trazem consigo a necessidade do pesquisador estipular a forma funcional da relação entre o resultado que se pretende avaliar e a variável de controle. No caso das abordagens não paramétricas, essa necessidade não existe.

Em 2001, Hanh, Todd e Van der Klaauwn [31] propõem uma abordagem para determinação do efeito médio do tratamento no ponto de corte baseada na então recente técnica de regressão linear local, que evita a baixa capacidade de estimativa na borda da técnica de regressão por *kernel* que vinha sendo usada até então. Essa abordagem ficou conhecida como HTV (referência ao nome dos autores) e, segundo Imbens e Lemieux [47], tornou-se a abordagem recomendada de estimativa RDD.

Sobre a ordem da polinomial a ser utilizada na regressão linear local, estudos recentes [63] advogam que polinomiais de baixa ordem são preferíveis frente às polinomiais de alta ordem, em razão da sensibilidade em relação à ordem da aproximação polinomial.

Determinação do *bandwidth*

Para aplicar a regressão linear local é necessário escolher uma faixa (conhecida na literatura como *bandwidth*) para restringir as observações a serem utilizadas na regressão [47, 62]. Por um lado, uma faixa muito larga, que inclua observações muito distantes do ponto de corte, pode conduzir a estimativas enviesadas. Por outro lado, faixas muito estreitas tendem a reduzir drasticamente a precisão das estimativas [62].

Até 2010, a maioria dos estudos com RDD usavam um critério baseado em validação cruzada (na literatura conhecido como *cross-validadion*) para selecionar a largura de faixa que mais ajusta os dados, entretanto, em 2011, Guido Imbens e Karthik Kalyanaraman apresentaram um procedimento para escolha do *bandwidth* para o estimador RDD (conhecido na literatura como procedimento IK) [64] que foi largamente aceito pelos pesquisadores do tema. A abordagem anterior baseada em *cross-validation* tentava escolher o *bandwidth* que minimizava uma aproximação do erro médio integrado (MISE) que é ótimo para ajustar uma curva sobre todo o suporte dos dados. Ao invés disso, Imbens e Kalyanaraman propuseram um procedimento que minimiza o erro quadrático médio

(MSE) do estimador na fronteira do ponto de corte. Segundo os autores, a nova abordagem se mostra mais adequada para o RDD porque o estimador τ_{RDD} consiste justamente na diferença do resultado de duas regressões em lados opostos de um determinado ponto no limite desse ponto e não em todo o suporte dos dados. Em comparações com diferentes versões do método anterior realizadas pelos autores por meio de simulações e análises com dados realistas [64], o novo procedimento se mostrou adequado. O novo procedimento apresentou erro padrão e viés equivalentes ou melhores que os métodos vigentes à época.

Sharp e *Fuzzy Kink Regression Designs*

Uma característica comum de muitas políticas públicas é a presença de uma dobra (*kink*) ou múltiplas dobras na relação entre a variável de controle e o resultado de interesse [65]. O valor dos benefício por desemprego, por exemplo, é tipicamente definido por uma fórmula que depende dos ganhos prévios. Card et al. [65] discutem esta questão e caracterizam uma ampla classe de modelos nos quais o "*Regression Kink Design*" (RKD, or RK Design) provê inferências válidas sobre o efeito médio do tratamento. As figuras Figura 2.3 e Figura 2.4 extraídas do trabalho de Card et al. [65] ilustram a ocorrência de dobras na relação empírica entre o valor diário médio em euros dos benefícios por desemprego na Áustria e os ganhos anuais. Nos dois gráficos a escala horizontal foi ajustada para que a dobra ocorra no ponto zero. É possível observar um aumento da inclinação quando o rendimento ultrapassa o ponto de corte inferior ($Tmin$) e uma redução quando ultrapassa o limite superior ($Tmax$).

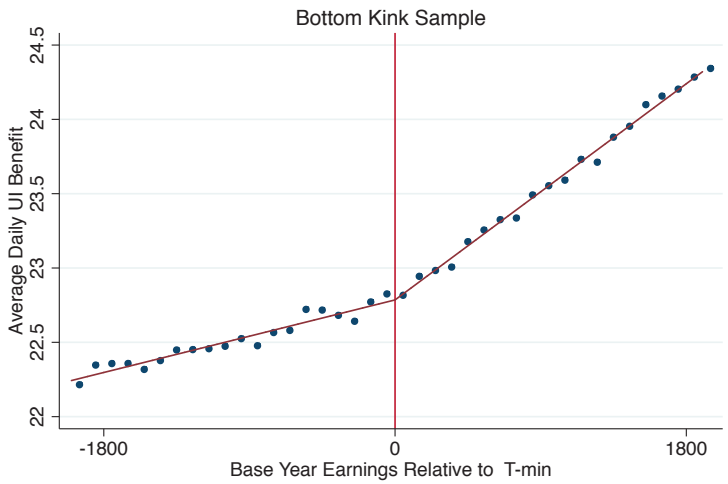

Figura 2.3: Kink RDD - Exemplo de dobra inferior (Fonte: [65]).

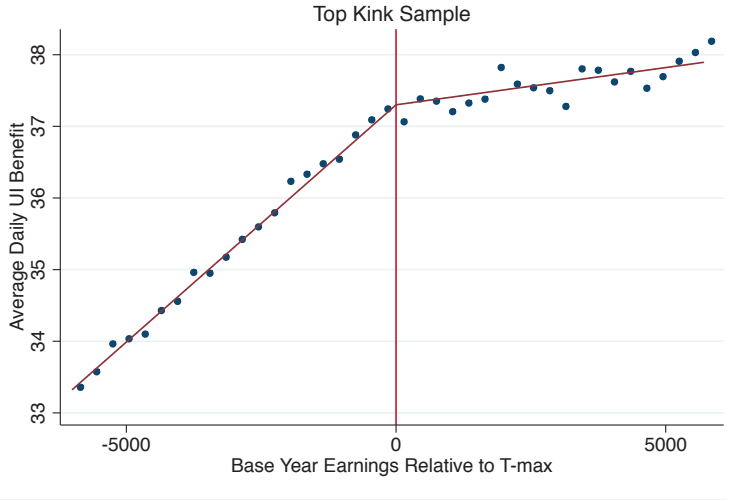

Figura 2.4: Kink RDD - Exemplo de dobra superior (Fonte: [65]).

Intervalos de confiança robustos

O método para seleção ótima do *bandwidth* é baseado na minimização do erro quadrático médio no ponto de corte [64], como visto anteriormente. Calonico et al. [66] alegam que este seletor produz intervalos de confiança que podem ser enviesados, com cobertura empírica abaixo do seu valor nominal. Eles propuseram um intervalo de confiança mais robusto, baseado em argumentos teóricos que se aplica a SRDD, FRDD e suas versões *Kink*, usando um estimador com correção de viés em conjunto com um novo estimador de erro padrão. Estes procedimentos estão disponíveis em um pacote[2] *R* descrito em [38].

Covariáveis

Segundo Imbens e Lemieux [47], é possível que, em determinados estudos, existam covariáveis adicionais (além da variável de controle) que possam ser usadas para reduzir o viés e aumentar a precisão das estimativas RDD. Os autores demonstram que apesar da redução no viés não ser significativa quando o *bandwidth* é pequeno, ele pode se tornar importante à medida em que o *bandwidth* aumenta. Adicionalmente, os autores afirmam que a inclusão de covaráveis pode aumentar a precisão, se elas forem correlacionadas com o resultado potencial.

Em 2016 Calonico et al. [67] examinaram as propriedades de um estimador polinomial local que incorpora covariáveis discretas e contínuas e mostraram que esse estimador ajustado por covariáveis se mantém consistente com o estimador padrão para o efeito do

[2]O pacote *rdrobust: Robust Data-Driven Statistical Inference in Regression-Discontinuity Designs* está disponível no repositório oficial do *R* em `https://cran.r-project.org/web/packages/rdrobust/index.html`

tratamento. Adicionalmente, eles conseguiram uma redução de 5% a 10% no tamanho dos intervalos de confiança. Esses procedimentos também se aplicam a SRDD, FRDD e suas versões *Kink*.

Validade Externa e Interna

Segundo Imbens e Lemieux [47], o grande atrativo do RDD é o alto grau de validade interna que pode ser obtido com um relativamente baixo número de pressupostos, quando comparado a outras abordagens quasi-experimentais. Em contrapartida, os autores também destacam que a validade externa do método é inferior a outras abordagens e está limitada à região próxima ao ponto de corte. Tais características tornam o método especialmente útil para realizar avaliações de impacto em determinadas subpopulações ou para avaliar a pertinência de pequenos ajustes no ponto de corte de políticas (alteração da idade limite para recebimento de um determinado benefício, por exemplo).

O alto grau de validade interna do RDD também é mencionado por Jacob et al. [62]. Entretanto, os autores destacam que, para que tal validade interna seja obtida, com resultados próximos ao dos métodos experimentais, alguns cuidados precisam ser tomados pelo pesquisador:

- a variável de controle não pode ser manipulada, causada, nem influenciada pelo tratamento;
- o ponto de corte é exógeno, ou seja, é determinado independentemente da variável de controle;
- nada além do tratamento e do resultado pode ser descontínuo no intervalo de análise, caso contrário, o resultado poderia ser atribuído a outro fator que não o tratamento em estudo;
- na ausência do tratamento, a relação entre o resultado e a variável de controle deve ser contínua.

Segundo McCrary [68] a questão da não manipulação da variável de controle é um importante requisto. O valor do ponto de corte não deve ser definido em um determinado ponto de modo a incluir ou excluir indivíduos específicos. Da mesma forma, os interessados não devem adulterar o seu valor para ser incluído ou excluído do grupo de tratamento. Em seu trabalho, o autor demonstra que o uso de variável de controle manipulada pode conduzir a resultados fortemente enviesados e propõe um teste[3] para verificar

[3] O teste de McCrary é implementado no pacote *rdd: Regression Discontinuity Estimation* está disponível no repositório oficial do *R* em `https://cran.r-project.org/web/packages/rdd/index.html`

a ocorrência de manipulação na variável de controle, baseado em um estimador para a descontinuidade da função de densidade da variável de controle no ponto de corte. Esse teste é implementado por meio da estatística do teste de Wald com a hipótese nula de que não há descontinuidade. A negação da hipótese nula neste caso indica que houve manipulação da variável de controle tendo em vista alterar artificialmente a probabilidade de participação no tratamento.

2.2 Integração e Pareamento das Informações

Em um trabalho de avaliação de impacto a partir de bases de dados governamentais, um dos primeiros desafios que precisará ser enfrentando é a integração dos dados disponíveis e a sua disponibilização em um ambiente computacional capaz de realizar as estimativas de acordo com a abordagem escolhida. Outra dificuldade importante para o pesquisador que pretende trabalhar com bases governamentais será o pareamento dos registros (em inglês, *record linkage*) das diversas fontes, antes da aplicação do método, uma vez que sistemas distintos muitas vezes não compartilham as mesmas chaves de identificação. Para exemplificar estes problemas, pode-se citar o trabalho de 2016 de Paiva et al. [69] no qual os autores abordam as dificuldades inerentes à análise de dados governamentais, principalmente no que se refere à diversidade das origens e propõem uma arquitetura capaz de controlar todas as fases do processo de integração dos dados de um portal corporativo. Nessa Seção será apresentada a fundamentação teórica acerca de integração e pareamento de informações, dois temas bastante relacionados entre si.

2.2.1 Integração das Informações

Em diversas situações nas quais o pesquisador deseje realizar avaliações a partir de dados governamentais, as bases de dados a serem utilizadas poderão estar em ambientes de bancos de dados distintos e heterogêneos. No caso do presente trabalho, por exemplo, as bases de dados do Cadastro Único e a folha de pagamento do PBF encontram-se disponíveis no TCU em banco de dados Oracle, enquanto a base RAIS encontra-se em banco de dados SQL Server. O tamanho das bases envolvidas também é um aspecto que requer atenção nesse e muito possivelmente em outros trabalhos baseados em dados governamentais. Por exemplo, a maior tabela do Cadastro Único possui 300 milhões de registros, a folha de pagamentos do PBF gera 40 milhões de registros por mês e a RAIS possui 400 milhões de registros por ano, conforme consultas realizadas nas próprias bases.

Trata-se, portanto, de um problema complexo de integração de dados. Segundo Bernstein e Haas [70], a integração de informações é uma das atividades que mais consomem recursos das empresas que lidam com tecnologia da informação, correspondendo até a

40% de seus orçamentos. Os autores destacam também que o gasto com aquisição de *software* representa apenas uma parte dos custos, uma vez que muito esforço é dispendido em atividades de movimentação, e padronização de dados. Ainda segundo Bernstein e Haas [70], existem diversas abordagens para lidar com este tipo de problema. A seguir são relacionadas, dentre as abordagens identificadas pelo autores, as mais relevantes no contexto do presente trabalho.

- **Carga de *Data Warehouse***. Um *Data Warehouse* é um banco de dados que consolida informações de diversas fontes [71]. Normalmente, ferramentas de *Extract-Transform-Load (ETL)* são utilizadas para resolver este problema, simplificando a tarefa de escrever *scripts*. Segundo Kabiri e Ciadmi [72], o ETL desempenha um papel crítico em um ambiente de *Data Warehouse*. Ainda segundo os autores, o ETL é responsável por coletar dados de fontes diversas e realizar a uniformização de formato quando os dados provem de fontes heterogêneas, realizar as transformações lógicas necessárias conforme as especificações do negócio e carregar os dados no banco de dados de destino.
- ***Virtual Data Integration***. Enquanto as ferramentas de ETL transferem fisicamente os dados de um ou mais ambientes de banco para outro, na abordagem *Virtual Data Integration* é fornecida uma visão integrada dos dados, sem necessidade de realizar movimentações físicas. Tais soluções de *Virtual Data Integration* fornecem um mediador [73] que recebe as consultas dos usuários, repassa as consultas para as diversas fontes de dados, consolida as respostas e devolve o resultado consolidado.
- ***Message Mapping***. Segundo Bernstein e Haas [70] *middlewares* orientados a mensagem ajudam a integrar aplicações desenvolvidas independentemente, movendo mensagens entre elas. Duas linhas de produtos com essa abordagem bastante conhecidas no mercado são o *Enteprise Application Integration System (EAI)* e o *Enterprise Service Bus* [74].

Bernstein e Haas [70] ainda citam as seguintes abordagens: Mapeamento Objeto-Relacional e Gerenciamento de Portais. Dada a necessidade de disponibilização dos dados integrados em um ambiente propício à execução dos algoritmos de estimação de impacto, a abordagem baseada em ETL se mostra a mais adequada para o presente projeto de pesquisa.

2.2.2 Pareamento das Informações

No caso do projeto em questão, as bases de dados do Cadastro Único e a Folha de Pagamentos compartilham uma mesma chave comum, entretanto, não existe uma chave

confiável comum entre a RAIS e as outras duas bases e, para a realização da análise de impacto, é importante determinar com segurança se beneficiários do PBF conseguiram alcançar o trabalho formal, ou seja, se possuem declaração da RAIS no ano sob análise ou não. Trata-se de um problema típico de pareamento de registros, ou *Record Linkage*, em inglês.

Winkler [75] define *Record linkage* ou *matching* computadorizado como sendo a ciência de relacionar mesmas entidades usando identificadores imperfeitos como nome, endereço e data de nascimento. No Brasil, esse problema de pareamento de dados entre bases governamentais é, infelizmente, muito comum, pela falta de um identificador único de pessoas que seja compartilhado entre todas as bases de dados de políticas públicas. Por exemplo, nas bases da Receita Federal a chave de identificação de pessoas físicas é o Código de Pessoa Física (CPF). Já nas bases da Previdência Social, a chave de identificação é o Número de Identificação do Trabalhador (NIT). Nas bases referentes às políticas de trabalho, a chave de identificação é o Programa de Integração Social (PIS). Algumas políticas públicas ainda possuem identificação própria como é o caso do PBF. No caso do PBF, apesar de a base de dados possuir os atributos CPF e NIT, a informação ali presente não é confiável, com ocorrência de valores nulos e até mesmo valores repetidos entre pessoas diferentes. O mesmo ocorre com a RAIS, que também traz o atributo CPF, além do PIS.

Segundo Winkler [75], a preparação prévia dos dados possibilita estruturar uniformemente nomes, endereços e outros campos em seus componentes de modo possibilitar aplicar os métodos teóricos de pareamento. Ainda segundo Winkler [75], a padronização pode trazer mais resultados que o uso de algoritmos sofisticados de pareamento. Um exemplo de padronização aplicado ao problema do presente projeto de pesquisa é a uniformização do campo CPF presente tanto no Cadastro Único quanto na RAIS em um campo de 11 posições com zeros a esquerda.

Em relação a campos textuais, conforme [76] dentre as abordagens para comparação existentes destacam-se os tratamentos fonéticos como o Soundex e o NYSIIS.

Em seu trabalho, Winkler [75] também provê uma visão geral sobre os métodos estatísticos que se mostraram efetivos no pareamento de informações. Segundo o autor, as ideias básicas por trás dos métodos de *matching* foram formalizadas por Fellengi e Suntera [77] e passam pela determinação de uma nota (*score*) e 2 pontos de corte (*threshold*). Para pares de registros cujo *score* é igual ou superior que ao maior ponte de corte, tem-se um *match*. Para pares cujo *score* fica entre os dois pontos de corte, tem-se uma possibilidade. Para os demais casos tem-se o não *match*.

Os principais métodos de pareamento utilizam abordagens não-supervisionadas, uma vez que dados para treinamento normalmente não estão disponíveis em projetos de *record*

linkage [75]. Para estimação da taxa de falsos *matchs* são utilizados métodos estendidos com abordagens não-supervisionadas e semi-supervisionadas[4], tendo as últimas maior acurácia.

2.3 Processamento Paralelo e RDD

Dado o volume de dados envolvido no presente trabalho e a complexidade dos algoritmos de estimação de impacto de programas, é importante que os softwares utilizados nas estimativas aproveitem eficientemente os recursos de hardware disponíveis, para otimizar o tempo de processamento. De acordo com levantamento e análise comparativa realizados em 2016 [36], os *softwares* de estimação RDD de código aberto disponíveis em linguagem de programação R[5] são os pacotes *rdd* [37], *rdrobust* [38] e *rddtools* [39]. Conforme pode ser facilmente verificado pela análise dos softwares, nenhum deles suporta nativamente processamento *multithread* e, portanto, não aproveitam de forma eficiente os recursos dos processadores atuais.

Segundo Cerroti [78], a evolução dos processadores levou ao desenvolvimento das atuais *Central Processing Units (CPUs) multicore* (em português, Unidades Centrais de Processamento multi-núcleo), que são confiáveis e estão disponíveis a baixo custo. Cerrotti acrescenta que para que possam aproveitar ao máximo tamanho poder de processamento, os desenvolvedores de sistemas precisam conhecer a fundo o funcionamento da arquitetura *multicore*. A plataforma R [35], por exemplo, possui o pacote *doMC* [79] que possibilita a realização de *loops* em paralelo em máquinas *multicore*, entretanto, os algoritmos de estimação RDD não utilizam este recurso.

Segundo Hasanov [80], a multiplicação de matrizes é muito frequente em algoritmos numéricos de álgebra linear e é um dos mais estudados problemas na computação de alta performance. De fato, por utilizar regressão linear local [31] os algoritmos de estimação RDD lançam mão de rotinas internas de multiplicação de matrizes, especialmente para o cálculo dos intervalos de confiança das estimativas. Quanto maior for a quantidade de dados utilizados na regressão linear local, maior será a ordem das matrizes multiplicadas, resultando em significativo consumo de processamento. Conforme Hasanov [80], a

[4]Os termos supervisionado e não-supervisionado são amplamente utilizados em mineração de dados para descrever a forma com o algoritmo aprende com base nos dados [75]. Quando o aprendizado ocorre a partir de um conjunto dados pré-classificados tomados como exemplo (os chamados dados de treinamento), têm-se o aprendizado supervisionado. Quando o aprendizado ocorre sem que sejam fornecidos dados de treinamento, têm-se o aprendizado não supervisionado. O termo semi-supervisionado é utilizado em [75] para descrever situações nas quais subconjuntos de pares são revisados por por especialistas e as probabilidades de *match* são reestimadas em um processo iterativo

[5]R [35] é a linguagem e ambiente para computação estatística *open-source* preferida por economistas [40]. Ver discussão na subseção 1.1.3.

otimização de tais procedimentos é fundamental para redução do tempo total de processamento.

Seja a multiplicação de matrizes representada na expressão 2.10. Um método simples de multiplicação de matrizes em paralelo consiste em dividir a primeira matriz por linhas e depois multiplicar cada bloco pela segunda matriz (equações 2.11 e 2.12). O resultado final é obtido combinando as matrizes resultantes em 2.11 e 2.12 por linha, na mesma ordem em que os blocos foram divididos

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \tag{2.10}$$

Bloco 1:

$$\begin{bmatrix} a_{11} & a_{12} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \end{bmatrix} \tag{2.11}$$

Bloco 2:

$$\begin{bmatrix} a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \tag{2.12}$$

No exemplo anterior, os blocos têm o mesmo número de linhas, mas isto não é um requisito. Para melhores resultados, a primeira matriz deve ser dividida em tantos blocos quanto forem os processadores disponíveis.

No presente trabalho, foi utilizado o algoritmo descrito acima em conjunto com a infraestrutura de processamento paralelo provida pelo pacote *doMC* para dotar um algoritmo de estimação de recursos de processamento paralelo. O algoritmo e os resultados obtidos em estimações RDD sobre dados simulados e do Bolsa Família foram descritos em um artigo [81] que foi publicado nos anais da *International Joint Conference on Neural Networks* de 2017 (IJCNN 2017), na sessão especial *Large Datasets and Big Data Analytics: Theory, Methods, and Applications*. Além de detalhar o algoritmo utilizado, o artigo demonstra, por meio de dados simulados e reais, que em um cenário de alta variância, a utilização de grandes bases de dados é muito importante para melhorar a qualidade das estimatitivas. Os resultados indicam que, com um *dataset* de mais de 13 milhões de registros com informações sobre trabalhos de jovens do PBF e com um servidor com 10 *cores*, o algoritmo paralelo apresentou um tempo de execução 6 vezes melhor que o algoritmo sequencial, sem alterar as estimativas.

2.4 Trabalhos Relacionados

O RDD é um método que, quando corretamente aplicado, pode gerar resultados bastante próximos aos da abordagem experimental [62]. Especialmente na avaliação de programas

de transferência condicionada de renda como é o caso do PBF, o método é muito utilizado.

No trabalho *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESSA* [21], já mencionado anteriormente, os autores aproveitaram a forma utilizada para implantação do programa de transferência condicionada de renda PROGRESSA no México, para fazer uma avaliação da performance do RDD. O PROGRESSA foi implantado de forma gradativa, e a escolha das cidades nas quais o programa seria inicialmente implantado foi feita de forma aleatória. Essa característica da implantação do programa mexicano, não muito comum em outros países, favoreceu sobremaneira a elaboração de estudos experimentais. Assim sendo, os autores aplicaram o método *Sharp Regression Discontinuity Design (SRDD)*, aproveitando uma descontinuidade no critério de elegibilidade do programa e compararam o resultado obtido com o método experimental. O trabalho concluiu que os resultados dos dois métodos foram equivalentes, sendo o RDD um excelente estimador do resultado experimental.

Existem muitos trabalhos recentes de avaliação de programas de transferência condicionada de renda utilizando RDD. Filmer e Schady [29] analisaram o impacto do programa de transferência de renda do Camboja na frequência escolar, utilizando *Sharp Regression Discontinuity Design* (SRDD). Barrientos e Villa [28] examinaram o impacto do *Familias en Accion* da Colômbia no mercado de trabalho. Scarlato et al. [82] investigaram os efeitos do programa *Chile Solidario* no mercado de trabalho, considerando uma perspectiva por gênero.

De todos os trabalhos localizados durante a pesquisa bibliográfica que subsidiou a elaboração desse projeto, o trabalho "*Conditional Cash Transfer and informality in Brazil*" [27] foi o mais correlato. Nele, os autores realizam uma avaliação de impacto do PBF por meio do método *Fuzzy Regression Discontinuity Design* (FRDD) para avaliar em que medida o programa afeta a escolha ocupacional dos beneficiários no sentido de torná-los mais propensos a optar por uma ocupação formal ou informal. Para aplicação do FRDD, foi explorada a descontinuidade presente na regra de elegibilidade do programa relacionada com o limite de idade das crianças. A análise foi feita com base nos microdados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2006, do Instituto Brasileiro de Geografia e Estatística (IBGE). Os resultados encontrados sugerem que o programa não tem impactos sobre a escolha ocupacional dos beneficiários entre os postos formais e informais. Este trabalho é particularmente interessante, pois está estritamente relacionado com o tema do presente projeto de pesquisa. Entretanto, existem algumas diferenças entre as abordagens que merecem destaque:

1. os autores utilizaram dados de pesquisa (PNAD) como fonte de informações, enquanto que na presente proposta de trabalho, pretende-se lançar mão de bases de dados governamentais disponíveis no ambiente do TCU, como o próprio Cadastro

Único, a Folha de Pagamentos do PBF e a RAIS para verificar a o ingresso no trabalho formal;

2. os autores avaliaram o impacto do programa na opção de trabalho do responsável familiar, enquanto que aqui pretende-se avaliar os efeitos do programa na inserção no mercado de trabalho dos jovens beneficiários ao alcançarem a idade adulta, configurando uma avaliação de impactos no médio prazo;

3. no trabalho de Barbosa e Corseuil [27], foi utilizada a idade do filho mais novo do grupo familiar em bimestres como variável de controle, enquanto que no presente trabalho pretende-se utilizar a idade dos próprios indivíduos. Adicionalmente, pretende-se utilizar a idade em dias, de modo a suavizar a distribuição do resultado em função da variável de controle.

Dentre as diferenças relacionadas, a utilização de bases governamentais como fonte de informações representa um diferencial significativo do presente trabalho. O Cadastro Único contempla informações sobre beneficiários de programas sociais implementados pelo governo brasileiro. Acerca do bolsa família, ele contempla a totalidade das famílias atendidas pelo programa, com informações pormenorizadas sobre os grupos familiares e indivíduos, incluindo dados sobre moradia e renda declarada. A Folha de Pagamentos registra os pagamentos realizados a todos os beneficiários com detalhamento dos dependentes familiares. Já a RAIS apresenta as informações declaradas pelos empregadores sobre seus empregados em um período de um ano. A declaração anual da RAIS é uma obrigação de todos os empregadores brasileiros [83], entretanto, analisando a base de dados disponível no TCU é possível verificar que existem alguns erros cadastrais e omissões. A forma como tais problemas foram enfrentados será apresentada na seção 3.1. Mesmo com esses problemas, a utilização de tais bases trouxe vantagens para a abordagem de estimação RDD como será apresentado no capítulo 4.

Capítulo 3

Metodologia

A metodologia utilizada no presente projeto de pesquisa corresponde a uma adaptação do processo recomendando por Imbens e Lemieux [47] com a incorporação de um detalhamento dos aspectos relativos à fase de obtenção e preparação dos dados, em razão da complexidade relativa à utilização de grandes bases governamentais. Foi incluída também uma fase de avaliação e adaptação de ferramentas, em razão do grande volume de dados envolvido.

As seções desse capítulo apresentam as etapas da metodologia utilizada no presente projeto de pesquisa, contemplando todas as atividades do projeto, incluindo a obtenção dos dados (Seção 3.1), avaliação e *refactoring* de ferramentas (Seção 3.2) e a realização das análises RDD (Seção 3.3).

3.1 Obtenção e Preparação dos Dados

Nesta seção é descrito todo o processo de obtenção e preparação de dados até a montagem dos *datasets* adequados à realização da análise RDD sobre o PBF. O processo aqui descrito foi totalmente automatizado utilizando o ferramental de integração de dados disponível no TCU, especificamente o *Informatica PowerCenter*[1]. Optou-se por desenvolver uma sistemática parametrizável, de modo a possibilitar a montagem de *datasets* com diferentes combinações de períodos, viabilizando a realização de uma série de análises RDD diferentes. Por exemplo, é possível parametrizar o processo para combinar dados do Cadastro Único de 2013 com dados da RAIS de 2014 e assim verificar o impacto do programa sobre os jovens que completam 18 anos na transição entre 2013 e 2014. Do mesmo modo, pode-se facilmente parametrizar o fluxo para verificar a situação dos jovens

[1] O *PowerCenter* é uma plataforma de integração de dados corporativos, dimensionável e de alto desempenho fornecida pela *Informatica*. Detalhes em `https://www.informatica.com/br/products/data-integration/powercenter.html`.

que completam 18 anos entre 2014 e 2015. A abordagem utilizada possibilita ainda a realização de análises de mais longo prazo, por exemplo, combinando dados do Cadastro Único de 2013 com os da RAIS de 2015, para verificar o impacto do programa sobre os jovens que completam 18 anos entre 2013 e 2014 em relação ao mercado formal de 2015. Essa característica do processo é especialmente útil para sua incorporação à sistemática de fiscalização contínua do TCU, uma vez que pode ser reaproveitado para novas análises ao longo do tempo, à medida em que novas massas de dados refrentes a novos períodos vão chegando ao Tribunal.

Inicialmente, no tópico 3.1.1, será apresentado o modelo de dados das bases envolvidas. Na sequência (tópico 3.1.2), será apresentado o processo desenvolvido em *PowerCenter* que contempla a integração e o pareamento de informações.

3.1.1 Modelo de Dados

A Figura 3.1 apresenta um Diagrama Entidade-Relacionamento (DER) com as entidades necessárias para a obtenção dos dados. No DER são apresentados os volumes das principais entidades. A Figura 3.2 apresenta um Modelo Relacional (MR) simplificado, no qual apenas os atributos utilizados de cada tabela são apresentados, para facilitar a visualização. É importante destacar que as tabelas são oriundas de sistemas diferentes, e foram internalizadas no ambiente do TCU mediante importação de arquivos texto, disponibilizados pelos órgãos responsáveis pela gestão das respectivas políticas públicas. Por esse motivo, a integridade referencial entre as tabelas não é assegurada. As tabelas com prefixo TB são oriundas do sistema Cadastro Único. A tabela FOLHA_BF corresponde à folha de pagamentos do PBF, que é gerada por um sistema específico. A tabela RAIS_EMPREGADOS contém a declaração anual da RAIS.

O Cadastro Único e a folha de pagamentos são mantidos pela Caixa Econômica Federal, assim sendo, as relações entre estas entidades possuem grau de confiabilidade elevado. Já a RAIS é mantida pelo Ministério do Trabalho, de forma que o relacionamento dessa com as demais não é confiável e, portanto, não foi representado no MR. Para vinculação dos dados da RAIS com os dados da Caixa Econômica, foram utilizadas as estratégias de *record linkage* apresentadas na Seção 3.1.2. Os atributos ANO_MES_CARGA (Cad. Único e folha) e ANO_RAIS foram adicionados pelo TCU para separar os dados conforme a periodicidade de envio de cada arquivo. Esses dois atributos não serão repetidos nos detalhamentos a seguir, mas fazem parte da chave lógica das entidades, que contém dados de mais de um período.

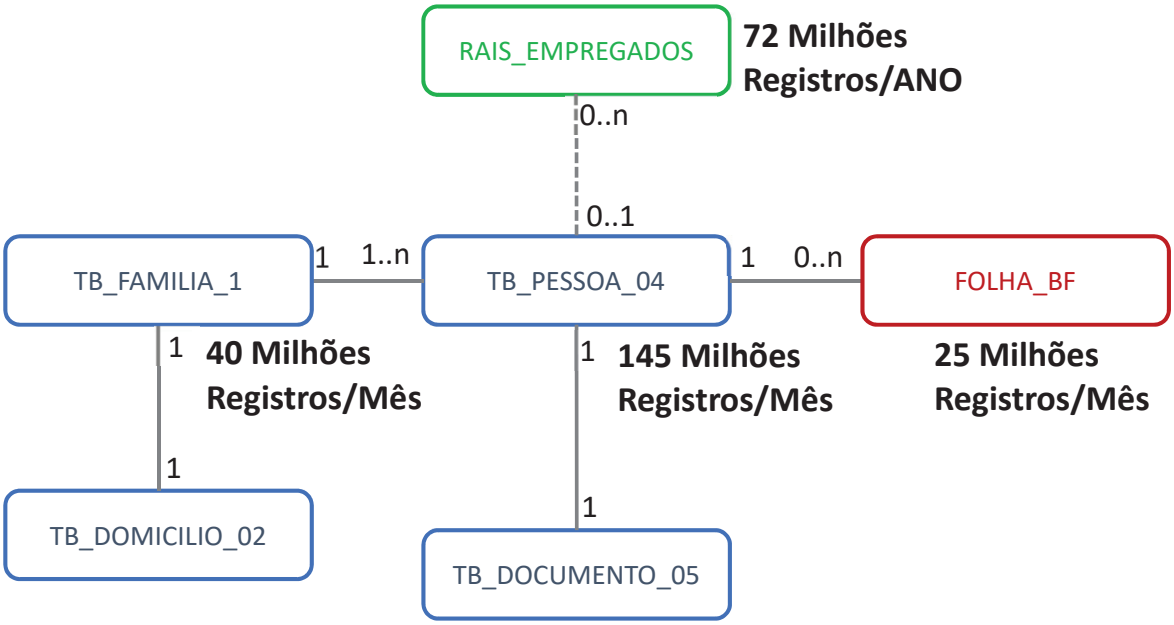

Figura 3.1: Diagrama Entidade-Relacionamento (DER).

TB_FAMILIA_1

Contém os dados principais das famílias cadastradas no Cadastro Único, incluindo o valor da renda média declarada. É importante destacar que nem toda família do Cadastro Único é beneficiária do PBF, a participação no programa deve ser verificada na folha de pagamento.

O atributo COD_EST_CADASTRAL_FAM=3 indica que os dados da família estão devidamente atualizados conforme as regras do cadastro único. A chave de identificação da família é o atributo COD_FAMILIA.

TB_DOMICILIO_02

Contém informações sobre o domicílio da família, que serão utilizadas para análises RDD por segmentos (urbano ou rural, por exemplo). A chave de vinculação com a entidade TB_FAMILIA_1 é o atributo COD_FAMILIA.

TB_PESSOA_04

Entidade principal dos membros da família. Além do nome e data de nascimento, contém o campo NUM_NIS_PESSOA_ATUAL, que representa o PIS/PASEP/NIT da pes-

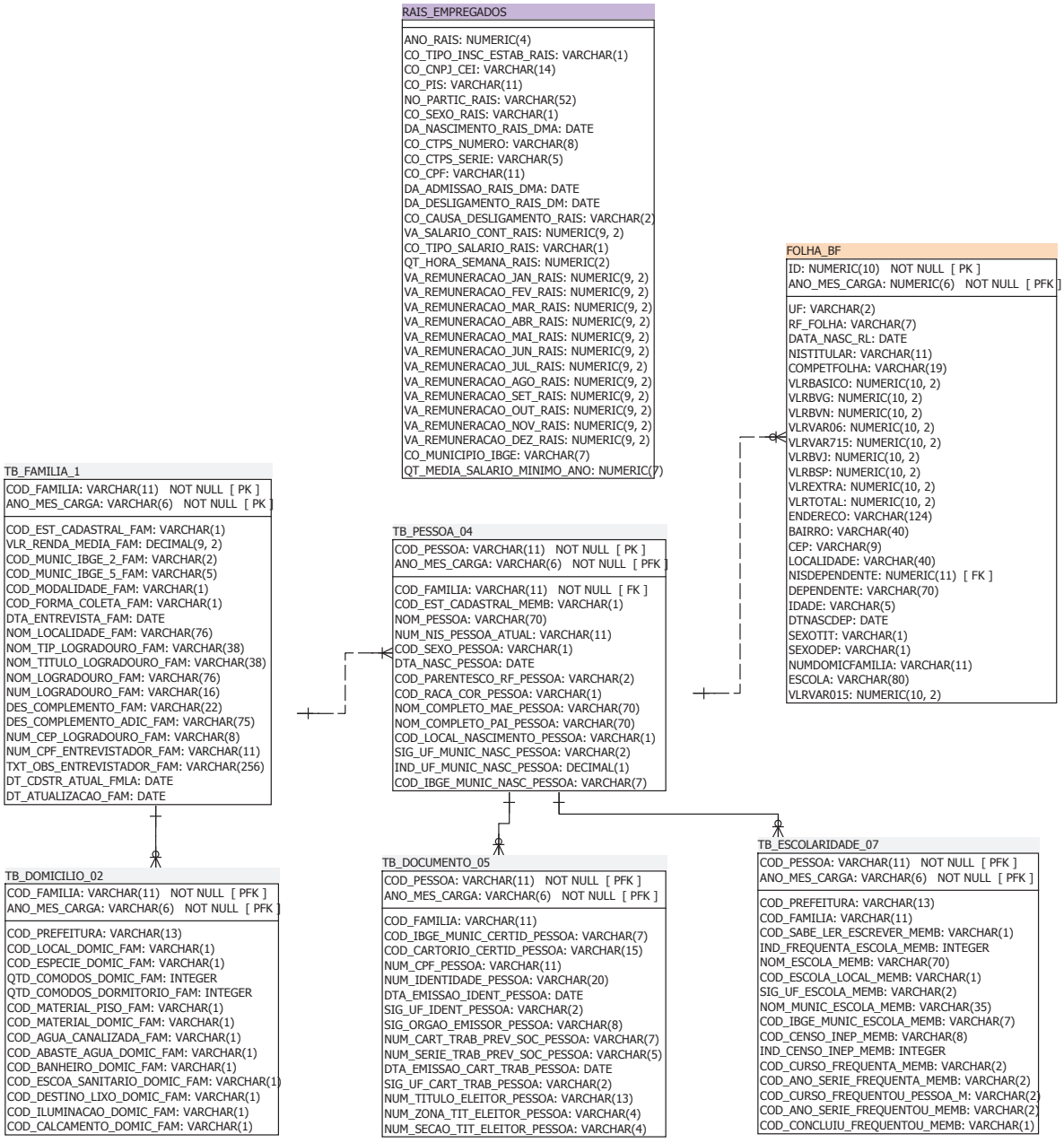

Figura 3.2: Modelo Relacional (MR).

soa, principal campo de identificação pessoal utilizado pela Caixa Econômica, e, consequentemente, utilizado para fazer a vinculação com a folha de pagamentos. O atributo COD_PARENTESCO_RF_PESSOA determina o grau de parentesco do membro com o responsável familiar (1-responsável, 2-Conjuge, 3-Filho(a), 99-Outros). O atributo COD_EST_CADASTRAL_MEMB=3 indica que os dados da pessoa estão devidamente atualizados conforme as regras do cadastro único. A vinculação com a entidade

TB_FAMILIA_1 é feita por meio do atributo COD_FAMILIA. A chave de identificação da pessoa é o atributo COD_PESSOA.

TB_DOCUMENTO_05

Dados dos documentos das pessoas. Contém o CPF, que juntamente com NIS, nome, data de nascimento, sexo e município, será usado para realizar o pareamento com a RAIS. A chave de vinculação com a entidade TB_PESSOA_04 é o atributo COD_PESSOA.

TB_ESCOLARIDADE_07

Informações sobre escolaridade das pessoas. Será usada em análises RDD segmentadas por grau de escolaridade e nível de defasagem escolar. A vinculação com a entidade TB_PESSOA_04 é feita pelo atributo COD_PESSOA.

FOLHA_BF

Contém os valores de benefício devidos ao responsável e a cada dependente das famílias beneficiárias do PBF. É por meio desta entidade que é possível verificar se um dependente recebeu ou não sua cota do benefício naquele período. A vinculação com a entidade TB_PESSOA_04 é feita com o atributo NISDEPENDENTE. Em um mesmo ANO_MES_CARGA, pode haver mais de um conjunto de lançamentos por família devido a lançamentos retroativos, o processo de integração deve considerar isso.

RAIS_EMPREGADOS

Contém informações sobre a declaração anual da RAIS acerca do trabalho formal no Brasil. Além de dados de identificação, contém os valores recebidos mês a mês no ano declarado. Uma mesma pessoa pode ter mais de um registro na RAIS no mesmo ano, caso mude de emprego, por exemplo. Para verificar o valor formal total recebido mês a mês pela pessoa, o processo de integração deve tratar as eventuais duplicidades. A ligação com o PBF foi feita por meio de técnicas de *record linkage*.

3.1.2 Integração e Pareamento de Dados

Conforme já mencionado, a integração de informações é uma atividade que consome muitos recursos em projetos de tecnologia da informação [70]. No caso do presente trabalho, que envolve bases de dados heterogêneas e de grande volume, a integração de dados requereu especial atenção. Nesse tópico, será descrito, passo a passo, o processo de integração desenvolvido em *PowerCenter*, que tratou tanto da integração lógica das entidades quanto

da questão da heterogeneidade dos ambientes, uma vez que a RAIS encontra-se em banco de dados *SQL Server* e as demais entidades estão em banco de dados *Oracle*.

Dado o tamanho das bases, o processo foi segmentado em estágios, com armazenamento de dados intermediários, de forma a possibilitar o acompanhamento e reduzir o consumo de recursos de memória nas operações de junção (*joins*). A Figura 3.3 apresenta a visão do *workflow* de integração e pareamento que contempla estágios de parametrização, seleção, integração, padronização, pareamento e agregação. Na figura, os estágios foram numerados e uma descrição de cada um deles é fornecida a seguir.

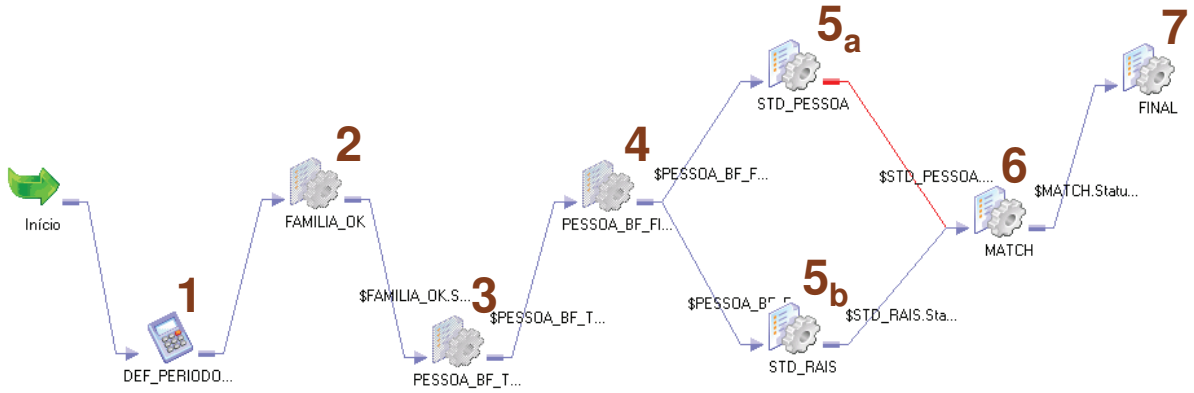

Figura 3.3: *Workflow* de integração e pareamento.

Estágio 1 - Parametrização

O primeiro estágio, representado na Figura 3.3 como uma calculadora (notação do *PowerCenter*) possibilita selecionar os períodos que serão utilizados para montagem do *dataset* final. Possibilita indicar o ano de referência que será utilizado para verificar o acesso ao mercado formal na RAIS, além do mês e ano de referência que será utilizado para obtenção dos dados cadastrais das famílias e dependentes no Cadastro Único. Adicionalmente, possibilita indicar os meses de referência que serão utilizados para verificação do recebimento do benefício pelos dependentes no ano base do Cadastro Único e no ano subsequente.

Esse estágio não realiza acesso a dados, mas define variáveis que serão usadas em todos os estágios subsequentes.

Estágio 2 - Seleção Inicial de Famílias

O primeiro passo do processo que envolve acesso a dados é a seleção das famílias segundo os critérios definidos para melhor aplicação da abordagem RDD. No caso, serão consideradas apenas as famílias com dados cadastrais atualizados e nas quais todos os membros

também estejam com dados em dia. Adicionalmente, optou-se por incluir apenas famílias compostas exclusivamente pelo responsável familiar, filhos e, opcionalmente, um cônjuge, para evitar a inclusão de famílias com outros tipos de dependentes, nas quais o efeito do programa pode ser dissipado por pessoas sem direito a parcelas específicas ou não cobertas pelas condicionalidades do programa (avós, agregados, etc). Apenas dados correspondentes ao mês de referência do Cadastro Único escolhido no passo de parametrização são considerados.

A Figura 3.4 representa este estágio. Pela imagem, pode-se perceber que é feita a junção das tabelas TB_FAMILIA_1 e TB_PESSOA_04 e na sequência é feita uma agregação que realiza a contagem de cada tipo de membro e da quantidade de membros desatualizados. A exclusão das famílias com dados cadastrais desatualizados (COD_EST_CADASTRAL_FAM <> 3) é feita na junção inicial, enquanto que a exclusão de famílias com membros desatualizados ou com membros que não atendem aos critérios é feita pelo filtro colocado após a agregação. Esse filtro exclui as famílias que possuem algum membro com dados desatualizados (COD_EST_CADASTRAL_MEMB <> 3) ou com algum membro que não seja responsável, filho ou cônjuge (COD_PARENTESCO_RF_MEMB = 99). O estágio é concluído com a gravação das famílias que atendem aos critérios na tabela temporária FAMILIA_OK.

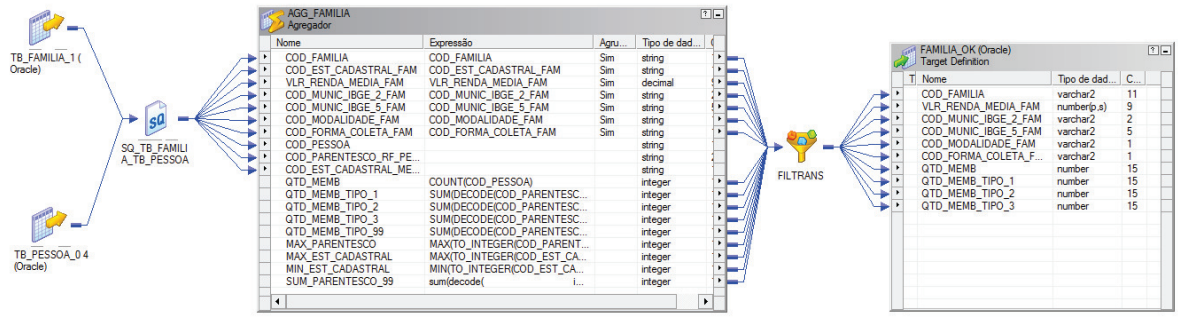

Figura 3.4: Seleção de famílias.

Estágio 3 - Obtenção de Dados dos Membros e Informações sobre Benefícios

Esse estágio é responsável por coletar os dados cadastrais de todos os membros das famílias selecionadas, além de informações sobre o recebimento do benefício no ano base utilizado para o Cadastro Único. O estágio é representado pela Figura 3.5, na qual se verifica que inicialmente é feita uma junção envolvendo a tabela temporária das famílias selecionadas, a tabela com informações sobre o domicílio da família, as tabelas de dependentes e a tabela com a folha de pagamentos do programa. A junção realizada é uma junção externa

à esquerda que assegura que todas as famílias selecionadas anteriormente serão incluídas na seleção, mesmo que algum dado complementar não esteja disponível ou que não haja pagamentos de benefício para o dependente. Além disso, a seleção de todas as tabelas respeita a informação dos períodos indicada no estágio de parametrização.

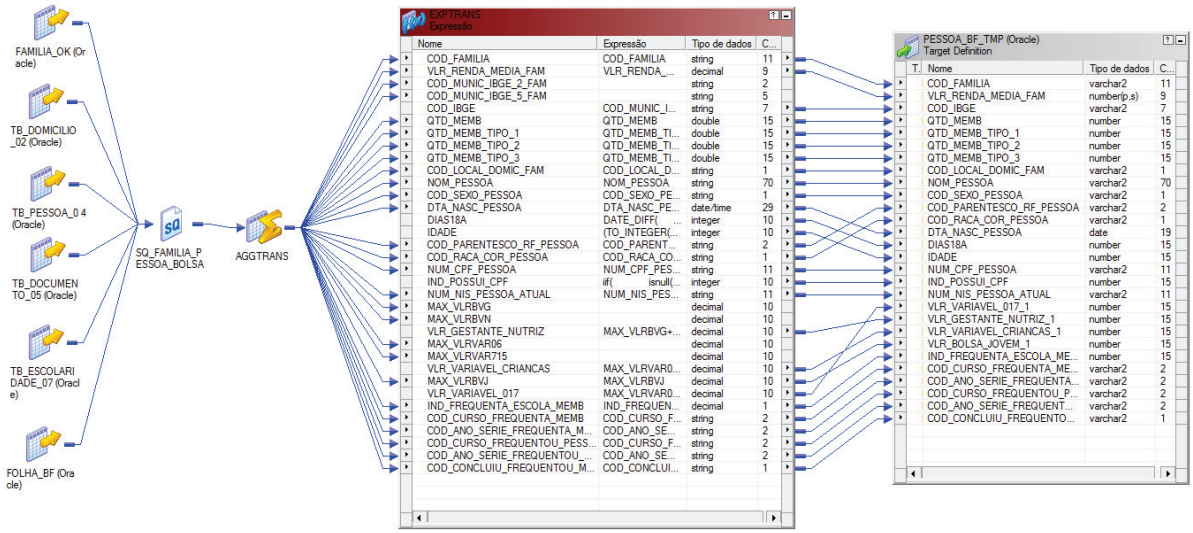

Figura 3.5: Obtenção dos dados dos dependentes das famílias selecionadas.

Após a junção, é realizada uma agregação para tratar a ocorrência de duplicidades decorrentes de pagamentos retroativos presentes na folha de pagamentos do programa. O valor final considerado corresponde ao maior valor da parcela destinada ao dependente no mês de referência considerado. Esse valor é posteriormente utilizado para determinar se o dependente recebeu ou não o benefício no período.

Após a agregação, é calculada a idade dos dependentes em anos e em dias. Além disso, os valores são consolidados em categorias, a saber: valor destinado a crianças de zero a 15 anos, a adolescentes de 16 a 17 anos, aos filhos de maneira geral (de zero a 17) e à gestante ou nutriz. Por fim, os dados são armazenados na tabela temporária PESSOA_BF_TMP.

Estágio 4 - Informação sobre Benefícios no ano Subsequente

Nesse estágio (Figura 3.6), os dados resultantes do estágio anterior são complementados com informações sobre os benefícios no ano seguinte. É feita uma junção da tabela temporária com a folha, mas desta vez, são usados os dados do mês parametrizado com a referência da folha para o ano subsequente.

Aqui também é necessária uma agregação para tratar pagamentos retroativos e, assim como no passo anterior, os valores são consolidados em categorias.

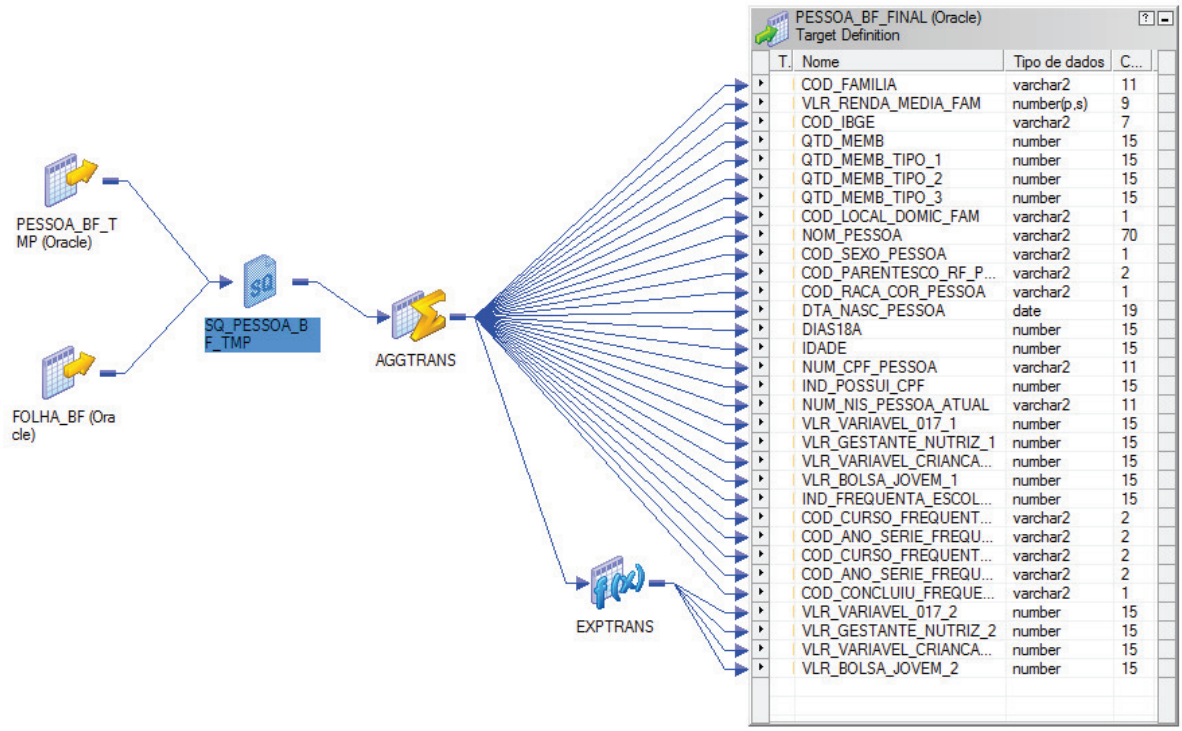

Figura 3.6: Obtenção de informação acerca do recebimento no ano subsequente.

Estágio 5 - Padronização de Nomes

Nesse estágio é realizada a padronização recomendada por Winkler [75] para maximizar os resultados do *matching* a ser realizado no próximo estágio. Os atributos CPF e PIS/-PASEP/NIT presentes tanto no Cadastro Único (incluindo folha de pagamentos do PBF) quanto na RAIS já foram padronizados para 11 caracteres quando da importação para o ambiente do TCU.

Entretanto, estes dois atributos não são suficientes para a realização do pareamento (*record linkage*) entre as duas bases. Apesar de possuir boa confiabilidade na RAIS, o CPF possui baixo nível de preenchimento entre jovens no Cadastro Único. Já o PIS/PA-SEP/NIT, apesar de possuir bom nível de preenchimento nas duas bases, não garante boa qualidade de cruzamento, pois uma mesma pessoa pode possuir dois números diferentes. Muitos dos códigos NIS existentes no Cadastro Único são gerados pela Caixa Econômica Federal especificamente para cadastro das pessoas no Bolsa Família. Apesar do NIS poder ser reaproveitado com identificador do trabalhador, para muitas destas pessoas, quando da ocorrência do primeiro emprego, é gerado um número NIT.

Dessa forma, faz-se necessária a utilização de outras informações para criar o critério pareamento. No caso especifico, utilizou-se também o nome da pessoa, a data de nascimento, o município, o estado da federação e o sexo para definir a nota do pareamento ou

match score [75].

A data de nascimento já se encontra armazenada no formato nativo do banco de dados, não sendo necessária padronização complementar. Município, estado da federação e sexo também já estão padronizados. Resta, então, uniformizar o nome. Como será mostrado nos resultados do pareamento, existem muitos erros de grafia nos nomes em ambas as bases, dificultando sobremaneira o pareamento. A estratégia de uniformização adotada visa a enfrentar a questão dos erros de grafia, adotando a conversão fonética conforme citado em [76], mais especificamente o algoritmo Soundex disponível no *PowerCenter*.

As Figura 3.7 e Figura 3.8 apresentam, respectivamente, as abordagens de padronização adotadas para os nomes do Cadastro Único e da RAIS.

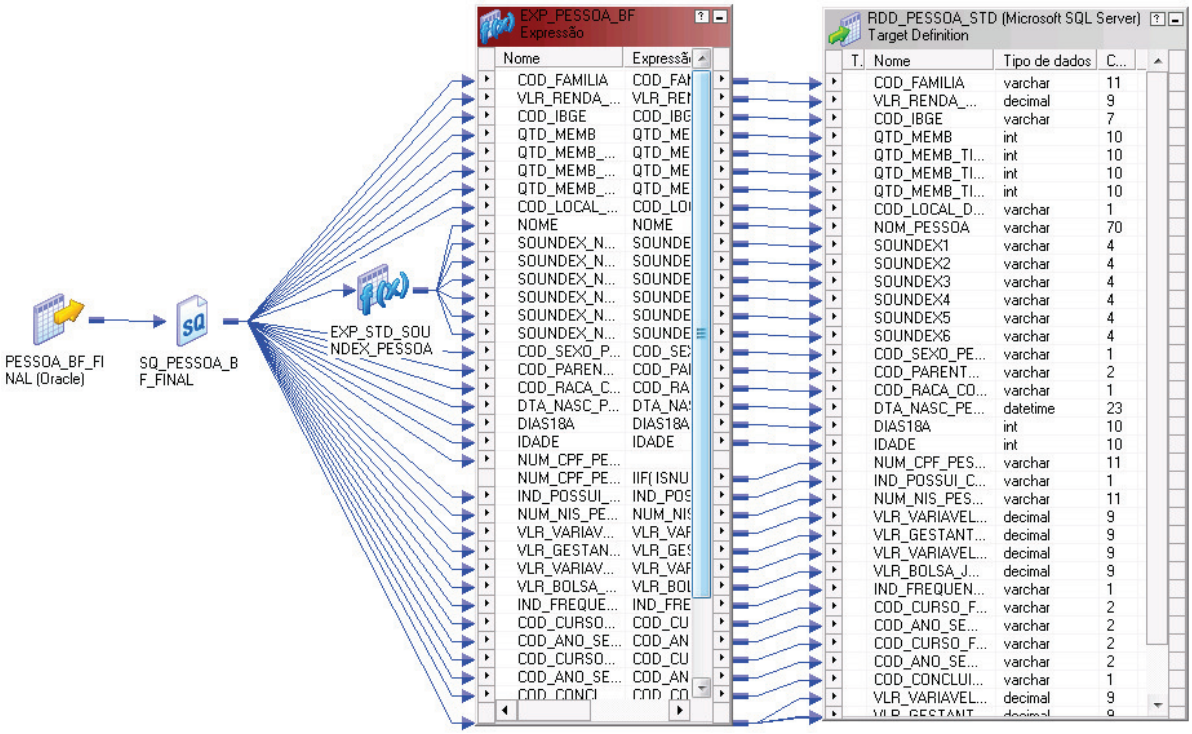

Figura 3.7: Padronização de nomes do Cadastro Único.

Em ambos os casos, os nomes completos das pessoas são divididos em até seis palavras desprezando partículas como do, da, de, dos e das. Cada palavra é então convertida em seu código fonético. Os dados padronizados são armazenados nas tabelas temporárias RDD_PESSOA_STD e RDD_RAIS_STD. Esse passo também resolve a questão da heterogeneidade dos ambientes, pois as tabelas temporárias padronizadas são todas armazenadas no *SQL Server*. Como pode ser verificado no *workflow* (Figura 3.3) as padronizações do Cadastro Único e da RAIS são realizadas em paralelo.

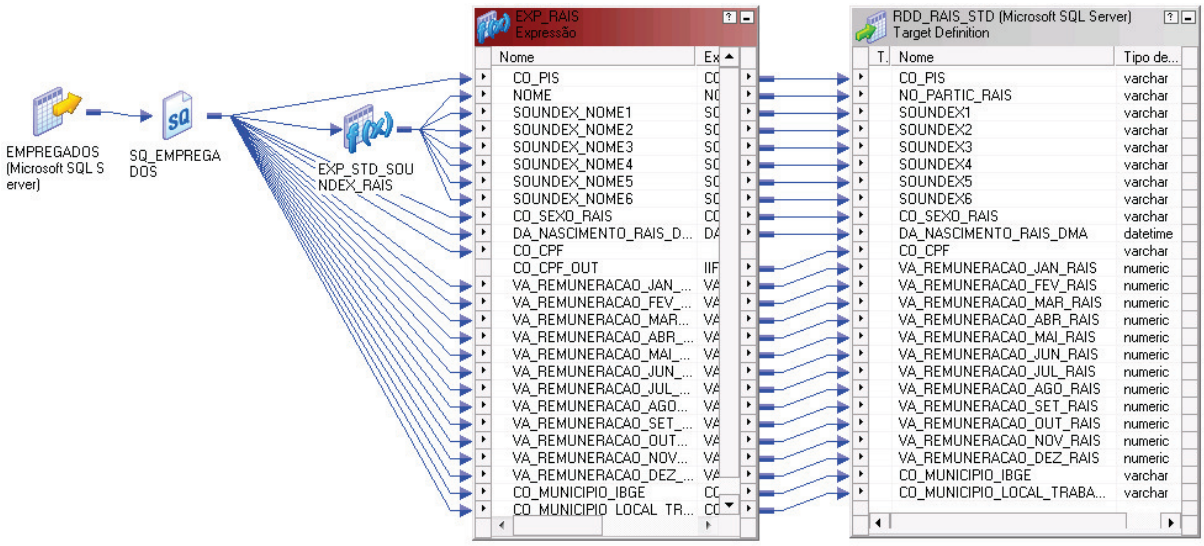

Figura 3.8: Padronização de nomes da Rais.

Estágio 6 - Pareamento

A abordagem adotada no pareamento das informações do Cadastro Único com as da RAIS consiste na atribuição de uma nota (*match score*) para os pares montados a partir dos dados das duas tabelas conforme definido em [75].

Dado o tamanho das bases envolvidas, a formação de pares a partir de todas as combinações possíveis entre os registros das duas tabelas geraria um montante de dados exageradamente grande, demandando uma quantidade muito grande de recursos de máquina para seu processamento. Mesmo considerando que apenas fossem utilizados os dados da RAIS para jovens com menos de 19 anos, isso representaria um montante de aproximadamente 5 milhões de registros na RAIS e 14 milhões de registros oriundos do Cadastro Único, que representaria um total de aproximadamente 70 trilhões de combinações, o que não seria viável.

A solução adotada, apresentada na Figura 3.9 consiste na realização de 3 junções externas independentes, uma por CPF, uma por PIS/PASEP/NIT e uma pelas partículas fonéticas do nome e na posterior união destes três resultados para cálculo da nota (*match score*). Com essa abordagem, apenas não seriam contemplados eventuais *matches* nos quais nem o CPF, nem o PIS nem a versão fonética do nome correspondem. Outras junções podem ser agregadas ao algoritmo proposto para aumentar sua cobertura, por exemplo, as três primeiras partículas fonéticas do nome e a data de nascimento, ao custo de acrescentar alguns milhões de linhas no processamento, entretanto essas alterações serão deixadas para trabalhos futuros.

Os ganhos de cobertura obtidos com o algoritmo de *record linkage* proposto em relação a abordagem de junção simples por meio das chaves são apresentados no Capítulo 4.

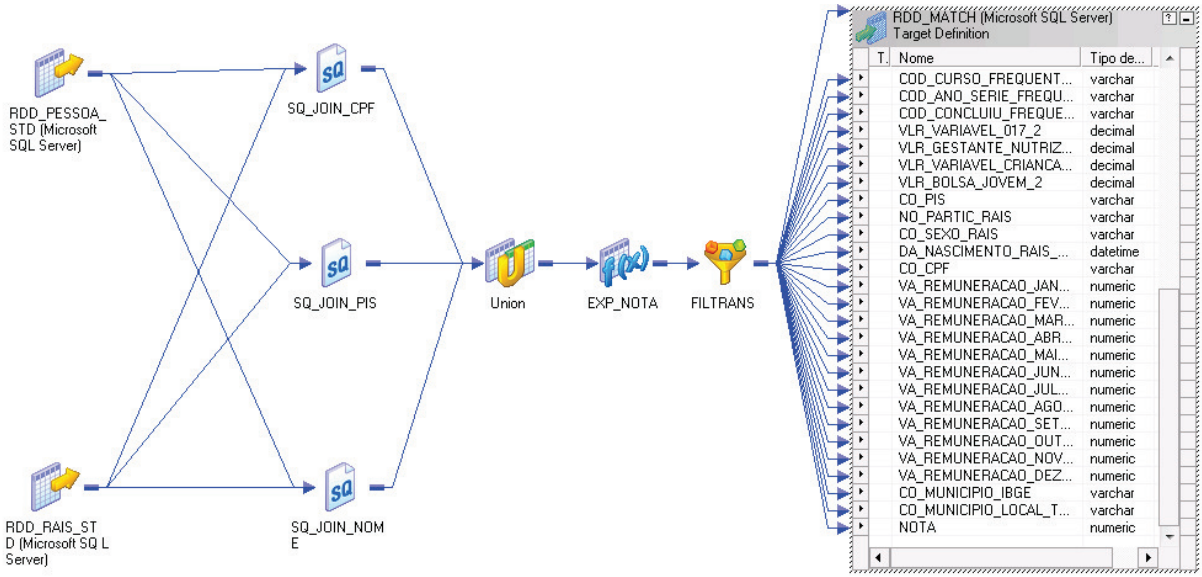

Figura 3.9: Pareamento de Informações.

A atribuição da nota de pareamento ocorre na transformação EXP_NOTA. Os campos disponíveis em ambas as bases que podem ser utilizados na formação da nota de pareamento são CPF, PIS/NIT, nome, data de nascimento, código do município e sexo. O CPF e o PIS/NIT são bons discriminadores, entretanto, o nível de preenchimento dos dois campos não é o mesmo nas duas bases. No caso do CPF ainda existe no no Cadastro Único a ocorrência de filhos que compartilham o CPF de um dos pais e compartilhamento de CPF entre cônjuges. Por sua vez, o PIS/NIT, apresenta a possibilidade de que uma mesma pessoa possua, legitimamente, dois números diferentes. O nome também é um bom descriminador, mas traz a possibilidade de existência de homônimos, além de ser muito suscetível a erros de grafia. A data de nascimento, apesar de sozinha não ser um bom descriminador, quando associada ao CPF ou ao nome, reduz os efeitos da existência de homônimos e do compartilhamento do documento entre membros da família. Dessa forma, optou-se por atribuir a estes 4 campos a mesma pontuação, totalizando 80% da nota de pareamento. Para o campo nome, em razão da possibilidade de erros de grafia, a nota foi dividida entre cada uma das partículas fonéticas, com maior peso para as partículas iniciais e um acréscimo de pontuação em caso de grafia exata. O restante da nota foi distribuída entre o código do município e o sexo. Por possuir maior poder de descriminação, ao código do município foram atribuídos 3/4 da pontuação restante.

Assim sendo, os pesos utilizados para definição da nota foram os seguintes:

- 20 pontos para correspondência de CPF;

- 20 pontos para correspondência de PIS/NIT;
- 20 pontos para correspondência de nome, sendo 3 pontos para a correspondência de cada uma das três primeiras partículas fonéticas, dois pontos para cada uma das três últimas partículas fonéticas e 5 pontos adicionais para o caso de grafia exata;
- 20 pontos para data de nascimento;
- 15 pontos no caso de correspondência exata do código do município ou 5 pontos caso a correspondência ocorra apenas na parte do código do município correspondente à unidade da federação; e
- 5 pontos para a correspondência do sexo.

A nota máxima de pareamento, segundo esse critério, é 100. A próxima decisão a ser tomada no processo de pareamento diz respeito à definição da nota de corte. A transformação responsável pela aplicação da nota de corte é o filtro localizado após a EXP_NOTA (Figura 3.9). Os *matches* encontrados com nota igual ou superior à nota de corte são armazenados na tabela temporária RDD_MATCH, para possibilitar conferência.

Como abordagem para escolha da nota de corte, foi adotado o seguinte procedimento:

- inicialmente, foi adotada uma nota de corte baixa, de modo a admitir, intencionalmente, a ocorrência de falsos *matches*, mas, sem contudo gerar uma base intermediaria muito grande na tabela temporária RDD_MATCH (no caso, utilizou-se como nota de corte inicial 20);
- foi gerada uma amostra aleatória de 3.000 observações dos dados presentes na tabela temporária RDD_MATCH e esta amostra foi ordenada pela nota em ordem decrescente;
- foi feita uma inspeção visual da amostra ordenada e identificado o ponto no qual começavam a ser perceptíveis os falsos *matches*;
- a nota de corte foi definida como sendo a menor nota múltipla de cinco (menor peso de campo) imediatamente anterior ao ponto identificado no passo anterior;
- por fim, o estágio de pareamento foi reexecutado com a nota de corte determinada no passo anterior, que no caso foi 45 pontos.

Apesar de haver um critério para escolha dos pesos, existe certa discricionariedade na sua determinação. Entretanto, como existe a oportunidade de determinação da nota de corte a partir da análise dos resultados obtidos pela aplicação do pesos, o resultado final passa a ter um critério objetivo.

Em trabalhos futuros, outras combinações de pesos podem ser experimentadas. Entretanto, é importante que, para cada combinação de pesos, os passos de determinação do ponto de corte ótimo sejam reexecutados.

Estágio 7 - Montagem do *Dataset* Final

Nesse estágio, apresentado na Figura 3.10, é feita a remoção das duplicidades geradas no processo de *match*, em função das múltipla junções e também das decorrentes de mudança de emprego, nas quais a mesma pessoa possui mais de um registro na RAIS.

A opção por realizar a remoção de duplicidades nesse estágio e não no estágio anterior deve-se ao fato de que, para um mesmo beneficiário, pode haver mais de um registro candidato a *match* na RAIS. Nesse caso, o estágio de *match* é que deve fazer a escolha do registro mais adequado. No estágio anterior apenas poderiam ser realizadas as remoções de duplicidades por múltiplas junções, que geram registros idênticos. Essa situação ocorre, por exemplo, quando existe coincidência de CPF e PIS no mesmo registro. Registros diferentes devem ser submetidos ao *match*. Assim sendo, como a remoção de duplicidades após o *match* não pode ser evitada, devido a múltiplos registros para uma memsma pessoa na RAIS, optou-se por fazer a remoção apenas nesse estágio.

As duplicidades que persistirem após o *match* serão removidas por agregação (transformação AGGTRANS). Na agregação, apenas o maior valor recebido por mês é considerado.

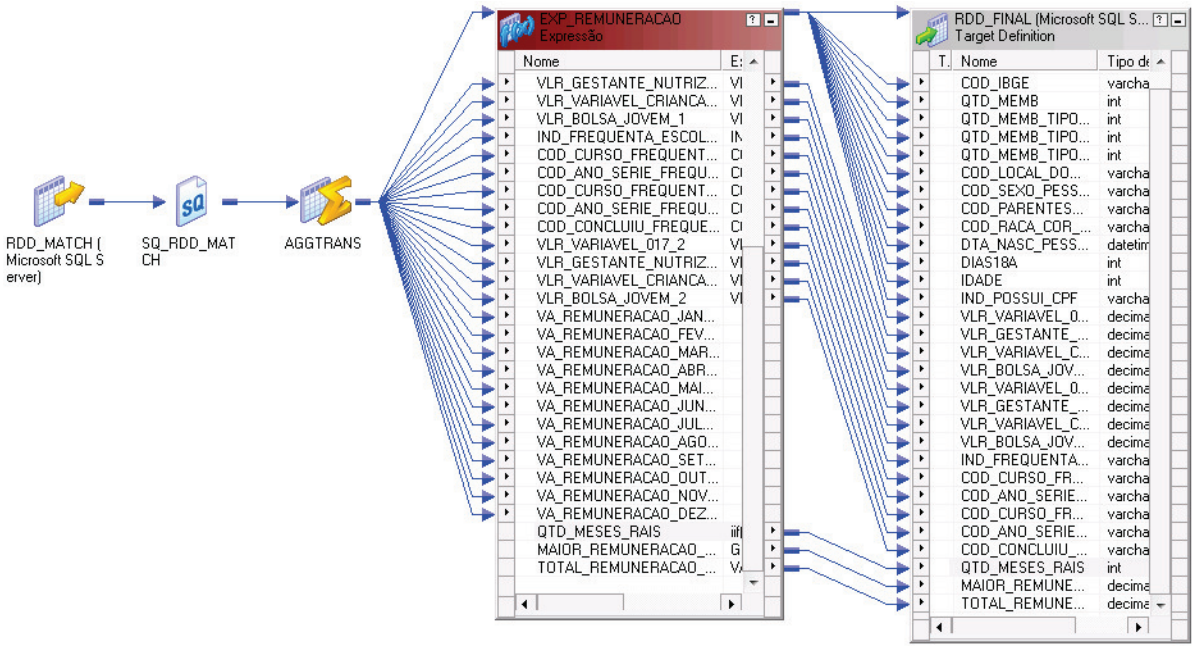

Figura 3.10: Montagem do *Dataset* Final.

Após a agregação, na transformação EXP_REMUNERACAO, os valores mensais são utilizados para determinar a quantidade de meses trabalhados no ano, a maior remuneração recebida e o total de remunerações recebidas.

Ao final, as informações de identificação e de natureza pessoal desnecessárias para a análise RDD como CPF, PIS e nome são removidas e o *dataset* final é gravado na tabela RDD_FINAL. A data de nascimento, o sexo e o código do município são importantes para a análise e são preservados. O nome da tabela é parametrizado para receber como sufixos os anos de referência do Cadastro Único e da RAIS utilizados. Assim sendo, para a análise de impacto do PBF no acesso dos jovens que completam 18 anos na transição entre 2014 e 2015 ao mercado de trabalho formal em 2015, o nome da tabela gerada será RDD_FINAL_2014_2015.

3.2 Avaliação e *Refactoring* de Ferramentas RDD

Antes da realização das análises RDD sobre os dados do Bolsa Família, foi realizada uma avaliação das principais ferramentas de estimação existentes na modalidade *open-source* para verificar seu comportamento quando aplicadas em grandes *datasets* e determinar a necessidade de alguma adaptação. Todos os testes foram executados em uma máquina virtual com 11 núcleos de processador e 25 GB de memória, executando sobre sistema operacional Red Hat Linux versão 3.10.0-327.10.1.el7.x86_64 (64 bits). Como ferramenta de desenvolvimento foi utilizada a plataforma R versão 3.3.0. Os procedimentos descritos nessa seção foram consolidados em um artigo [81] que foi publicado nos anais da conferência *International Joint Conference on Neural Networks(IJCNN 2017), special session Large datasets and big data analytics: Theory, methods, and applications*.

Essa etapa consistiu da avaliação das ferramentas usando dados simulados (Seção 3.2.1), *refactoring* de uma das ferramentas (Seção 3.2.2) e teste da implementação com dados reais (Seção 3.2.3).

3.2.1 Avaliação com Dados Simulados

Conforme descrito nas Seções 1.1.3 e 2.3, de acordo com levantamento e análise comparativa realizada em 2016 [36], existem atualmente três ferramentas *open-source* para estimação RDD escritas em linguagem de programação *R*: *rdd* [37], *rdrobust* [38] e *rddtools* [39]. Conforme pode ser verificado na sua documentação, o pacote *rddtools* estende o pacote *rdd* acrescentando, principalmente, ferramentas de visualização.

Considerando que o *rddtools* utiliza o código central do *rdd* nas estimativas, a avaliação foi realizada apenas sobre os pacotes *rdd* e *rdrobust*. Para avaliar os *softwares*, foram utilizados os dados simulados apresentados na Figura 3.11, que inclui uma descontinuidade

forçada de 5.000 no ponto de corte. O gráfico de dispersão (a) apresenta os dados brutos produzidos pela função de simulação[2]. O gráfico (b), gerado utilizando o ferramental de análise gráfica provido por [38], torna visível a descontinuidade induzida na função de simulação. Já o gráfico (c) mostra a natureza "*Fuzzy*" do modelo, uma vez que a probabilidade do tratamento salta de zero para um valor abaixo de 1 (≈0,5) no ponto de corte. Os dados gerados pela função de simulação apresentam alta variância e contemplam um impacto fixo de 5.000 no ponto de corte, para que seja possível avaliar a qualidade das estimativas.

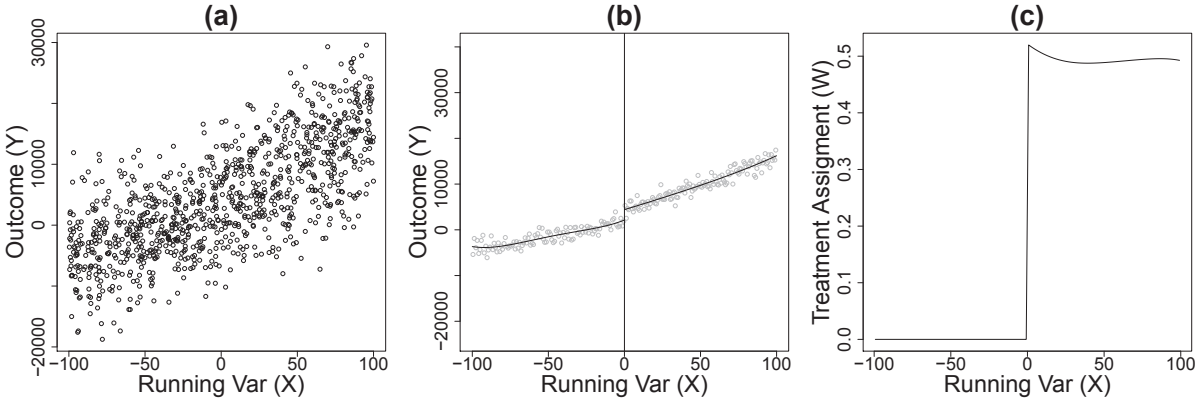

Figura 3.11: Características dos dados simulados. (Fonte: [81]).

Os dados simulados foram aplicados a ambos os *softwares* com 40 diferentes tamanhos de amostra, em passos de 5.000 observações, variando de 5.000 a 200.000 observações. Para cada tamanho de amostra, foram coletados o tempo de execução, o *bandwidth* selecionado pelo algoritmo, a estimativa de impacto e o intervalo de confiança. Os resultados da comparação são apresentados na Figura 3.12. Pelo gráfico (a) é possível perceber que o pacote *rdrobust* executou ligeiramente mais rápido que o *rdd* com o maior tamanho de amostra, mas ambos apresentaram uma complexidade assintótica de tempo quadrática. O *bandwidth*, apresentado no gráfico (b), se mostrou altamente variável entre as amostras, mas amostras maiores tendem a gerar *bandwidth* menores e mais estáveis. O gráfico (c) mostra a importância do tamanho da amostra em situações de alta variância. Amostras maiores geram estimativas mais acuradas e estáveis, com menor intervalo de confiança. É também possível perceber que as estimativas de ambos os softwares se aproximam muito nas amostras maiores.

É importante destacar que neste exemplo de dados simulados com apenas 200.000 observações, o tempo de execução no pior caso foi superior a 500 segundos (oito minutos e vinte segundos), mesmo executando em um servidor profissional. Pesquisadores traba-

[2]Código fonte disponível em https://github.com/githubanonymous001/rdd_parallel

lhando com máquinas *desktop* terão muito mais dificuldades, principalmente considerando que em um estudo, normalmente é necessário repetir as estimativas diversas vezes até a obtenção do resultado final. Pela verificação do consumo de recursos do servidor durante os testes dos dois softwares, ficou evidenciado que nenhum deles utilizou recursos de *multi-thread* da máquina e, portanto, o servidor ficou subutilizado durante todo o período[3]. Este problema serviu de motivação para que este projeto incluísse o *refactoring* do algoritmo RDD visando otimização de desempenho.

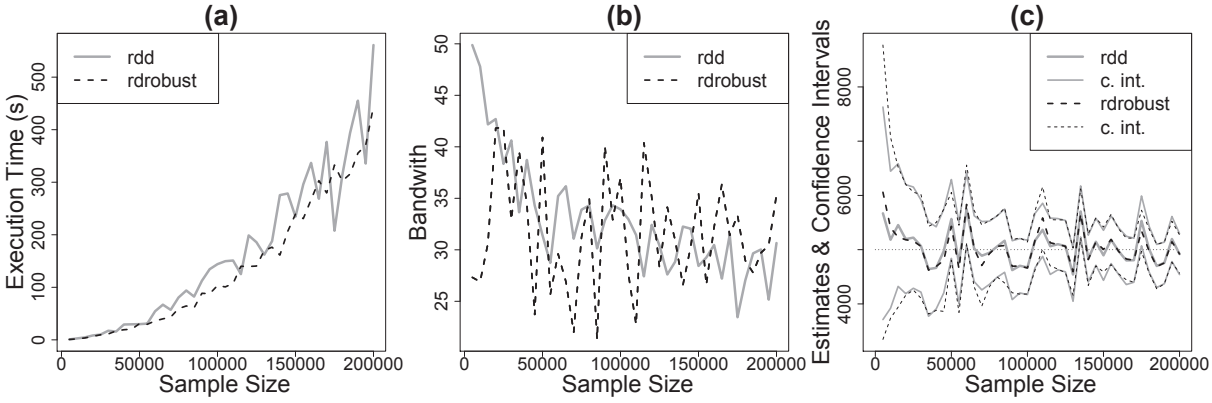

Figura 3.12: Comparação das ferramentas *open-source*. (Fonte: [81]).

3.2.2 *Refactoring*

Como as estimativas dos *softwares* ficaram muito próximas nas amostras maiores, o pacote *rdd* foi escolhido, porque o *rdrobust* apresentou algumas falhas[4] quando submetido aos dados do PBF.

Para aumentar o nível de aproveitamento dos recursos do *hardware* disponível, foi utilizado o pacote *DoMC* [79], uma infraestrutura que possibilita a execução de laços em

[3]O consumo de recursos no servidor foi verificado utilizando o utilitário *top* do Linux, com o comando interativo "1", que possibilita a visualização do consumo de recursos por núcleo de processamento. Durante a execução dos algoritmos RDD originais, apenas um núcleo era designado para o processo *R*. Com *datasets* maiores que 20.000 observações, o núcleo designado para o *R* mantinha-se com consumo variando entre 98 e 100% durante a maior parte do tempo de execução. Os demais núcleos permaneciam com consumo baixo (entre 0 e 5%), sendo designados para tarefas do sistema operacional. Com datasets menores que 20.000 observações, o tempo total de execução não era suficiente para que o consumo de recursos no núcleo se estabilizasse. Percebia-se apenas uma elevação momentânea no consumo do núcleo designado, que rapidamente voltava a zero com o término da execução. Assim como no caso de amostras maiores, o consumo dos demais núcleos permanecia baixo.

[4]Quando executado sobre algumas amostras extraídas do *dataset* do Bolsa Família, após a conclusão da execução, o *rdrobust* apresentava como resultados das estimativas e intervalos de confiança a palavra reservada do *R* NaN, que significa *not a number* (um não número, em português). Esse tipo de resultado é apresentado, por exemplo, quando se tenta obter a raiz quadrada de um número negativo. Não foi possível determinar as causas desse comportamento do *rdrobust*, pois nenhuma mensagem de erro era gerada nessas situações.

paralelo no R em sistemas Linux. Para maximizar os ganhos de desempenho, procurou-se pelos trechos de código com maior consumo de tempo no pacote *rdd* e verificou-se que uma sequência de multiplicação de 6 matrizes utilizada na determinação do intervalo de confiança da estimativa era responsável por aproximadamente 90% do tempo de execução nas amostras maiores. Estas multiplicações estão localizadas na função interna *hatvalues.ivreg*() do pacote *AER* [84] usado pelo pacote *rdd*. O código original simplificado é apresentado na Listagem 3.1. Os nomes das variáveis foram encurtados.

Listagem 3.1: Código original simplificado

```
hatvalues.ivreg <- function (mod, ...) {
  (...)
  diag(A %*% B %*% C %*% D %*% E %*% F)
}
```

Conforme apresentado na Seção 2.3, a multiplicação de matrizes pode ser paralelizada dividindo a primeira matriz por linhas, multiplicando cada fatia da primeira matriz pela segunda matriz e combinando as matrizes resultantes por linha na mesma ordem em que a primeira foi dividida. Entretanto, simplesmente paralelizar cada multiplicação de matriz individualmente conduziria a resultados sub-ótimos em razão do custo da infraestrutura de paralelização. Ao invés disso, foram paralelizadas as 6 multiplicações de uma só vez, usando as propriedades associativas da multiplicação de matrizes: $A * B * C = A * (B * C)$. Outro importante aspecto do problema é que apenas a diagonal da matriz resultante é necessária. Assim, é possível reduzir o custo da combinação de toda a matriz resultante, combinando apenas os elementos apropriados de cada matriz resultante em um vetor. A função refatorada detecta automaticamente o número de núcleos disponíveis usando a função *getDoParWorkers*() do pacote *DoMC* e então fatia a primeira matriz adequadamente, como na listagem 3.2

Listagem 3.2: Código refatorado

```
hatvalues.ivreg_par <- function(mod, ...) {
  (...)
  nc <- getDoParWorkers()
  idx <- splitIndices(nrow(A), nc)
  foreach(i=1:nc, .combine=c) %dopar% {
      diag(A[idx[[i]], ] %*% B %*% C %*%
           D %*% E %*% F[ , idx[[i]]])
  }
}
```

Algumas outras alterações menores foram necessárias para fazer o pacote *rdd* usar a versão otimizada da função *hatvalues.ivreg*(). O código fonte de todas as alterações está disponível, bem como os *scripts* utilizados para avaliar a refatoração[5].

3.2.3 Teste com Dados Reais

Os dados reais utilizados no teste do *refactoring* foram os resultantes do processo descrito na Seção 3.1.2 configurado para utilizar dados do Cadastro Único de 2013 e dados da RAIS de 2014. Ou seja, o impacto avaliado foi o acesso ao mercado de trabalho em 2014 para jovens que completam 18 anos na transição entre 2013 e 2014. A variável de controle utilizada foi a idade em dias e o resultado observado foi o maior salário obtido pelo jovem em 2015[6].

A versão original sequencial e a versão refatorada com processamento paralelo do *software rdd* foram aplicadas a 20 diferentes subamostras do conjunto de dados do *Bolsa Familia*, variando de 1/20 dos dados até todo o conjunto, em passos de 1/20. A versão paralelizada foi avaliada com 2, 5 e 10 núcleos de processadores. Para cada tamanho de amostra foram coletados o tempo de execução, o *bandwidth*, as estimativas e os intervalos de confiança. Os resultados obtidos são apresentados na Figura 3.13.

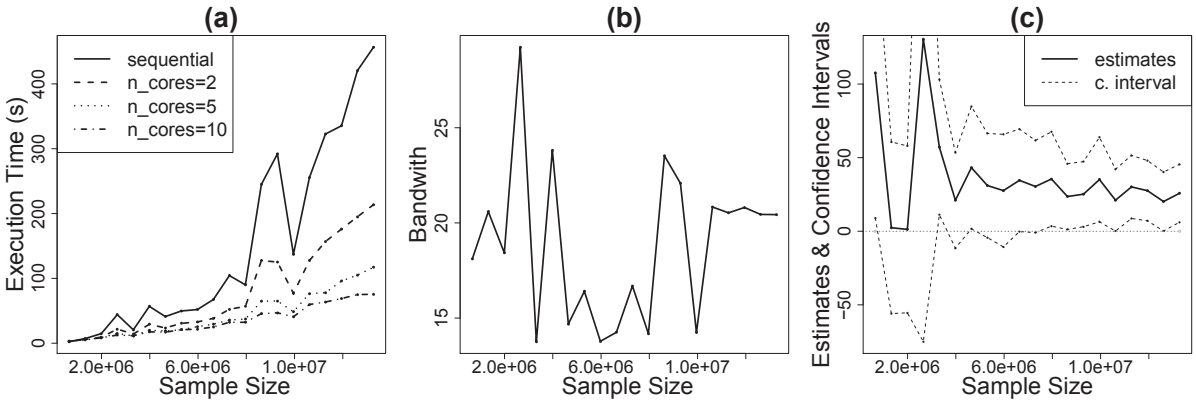

Figura 3.13: Resultados obtidos pelas versões original e modificada do *software* RDD. (Fonte: [81]).

As estimativas de ambas as versões foram exatamente as mesmas, uma vez que a refatoração não afetou a lógica do algoritmo. O gráfico (a) mostra o expressivo ganho de desempenho das versões paralelas nos maiores tamanhos de amostra. É possível observar que o ganho de desempenho não é linear com o aumento do número de processadores,

[5]Código fonte disponível em `https://github.com/githubanonymous001/rdd_parallel`.

[6]A configuração final da análise RDD realizada nesse projeto de pesquisa foi ligeiramente diferente da utilizada no teste do *refactoring*, considerou a comparação entre diferentes períodos e será descrita em detalhes na Seção 3.3

em razão do custo da infraestrutura de processamento paralelo. Os maiores ganhos são observados com os maiores tamanhos de amostra. Com o conjunto de dados completo e usando 10 núcleos de processador, a redução do tempo de execução foi de 88,8% (de 456 segundos para 51 segundos). Como visto anteriormente nos dados simulados, o *bandwidth* (gráfico (b)) é altamente dependente da amostra, mas fica mais estável à medida em que o tamanho da amostra aumenta. O gráfico (c) mostra como o intervalo de confiança reduz e como as estimativas ficam mais estáveis com o aumento do tamanho da amostra.

Usando todo o conjunto de dados, foi possível verificar em um tempo razoável que existe um efeito médio do tratamento no ponto de corte, com valor estimado de R$25,81, sendo que o intervalo de confiança de 95% foi R$6,15 a R$45,47, com um *bandwidth* ótimo de 20,43 dias. A versão paralelizada do software se mostrou adequada para ser utilizada no restante do projeto.

Como não é objetivo do teste de *refactoring*, não será feita nesta seção uma análise detalhada dos resultados das estimativas RDD. Essa análise será feita na Seção 3.3.

3.3 Aplicação da Abordagem RDD

Nesta seção será seguido o processo recomendado no trabalho *Regression Discontinuity Designs: A Guide to Practice*, de Imbens e Lemieux [47].

3.3.1 Análise Gráfica

Segundo Imbens e Lemieux [47], a análise gráfica é um importante instrumento de identificação e validação da estratégia em uma análise por descontinuidade. Nessa seção são apresentadas as análises gráficas utilizadas na escolha das variáveis e na validação empírica do modelo de regressão por descontinuidade.

Densidade da Variável de Controle

Conforme descrito na Seção 2.1.3, é importante que a variável de controle utilizada na análise RDD não sofra manipulação em razão da existência do programa, para assegurar a validade interna do método. Dadas as características do PBF, existem duas variáveis candidatas à variável de controle. A mais óbvia é a renda média declarada da família, uma vez que o benefício apenas é concedido para famílias com renda média inferior a um determinado valor. O ponto de corte relativo à renda pode variar anualmente, em função do reajuste dos valores do programa, concedido por meio de decreto presidencial [23]. Outra variável candidata é a idade dos filhos, uma vez que o benefício variável conhecido como bolsa jovem só é concedido até o ano em que o jovem completa 18 anos [9].

É razoável esperar que haja algum nível de manipulação da variável renda média declarada, uma vez que é auto-declarada pela família e não existem meios objetivos para que os gestores do programa possam validar a informação apresentada. Por outro lado, também é natural imaginar que a idade dos filhos seja mais difícil de manipular, uma vez que pode ser facilmente verificada nos documentos apresentados.

A Figura 3.14 apresenta a densidade da renda média declarada no cadastro único nos anos 2013 e 2014 (mês de referência junho[7]). Os dados foram extraídos diretamente da tabela TB_FAMILIA_1, considerando apenas famílias com dados cadastrais atualizados. Famílias com renda declarada igual a zero foram excluídas do *dataset*, para que a alta concentração de valores nesse ponto não prejudicasse a visualização dos dados. As linhas verticais representam o limite da linha de pobreza (ponto de corte) vigente à época (R$140,00 e R$156,00, respectivamente) com base no decreto presidencial [23]. A análise visual da figura indica a ocorrência de pontos de concentração de renda à esquerda dos pontos de corte. É possível também observar outros pontos de concentração em pontos limites vigentes em anos anteriores e que permanecem no cadastro único. Essas características das curvas de densidade reforçam a percepção empírica de que pode haver manipulação da renda. Para verificar essa hipótese foi utilizado o teste de desidade de McCrary [68], também apresentado na Seção 2.1.3, disponibilizado no pacote *R rdd* [37]. A Figura 3.15 apresenta os resultados dos testes para os anos de 2013 e 2014. Os *p-values* extremamente baixos obtidos em ambos os testes indicam que é possível negar a hipótese nula de que não há manipulação, com um nível de significância de bastante inferior ao limite usual de 5%, confirmando a intuição de que a renda declarada não pode ser usada como variável de controle em análises RDD.

Análise análoga foi feita com a variável idade em dias. A Figura 3.16 apresenta a densidade da idade presente nos *datasets* extraídos do processo de integração descrito no tópico 3.1.2. A densidade foi calculada apenas com dados no intervalo de mais ou menos um ano em relação ao ponto de corte. A análise visual da figura não indica a ocorrência de manipulação e essa intuição pode ser confirmada pelo teste de desidade de McCrary [68]. A Figura 3.17 apresenta os resultados dos testes para a transição entre 2013 e 2014 e entre 2014 e 2015. Em ambos os casos, considerando o *p-value*, não é possível negar a hipótese nula de que não há manipulação, com nível de significância de 5%. Ou seja, por esse aspecto, não há razão para questionar a utilização da variável idade como variável de controle em análises RDD.

Assim sendo, considerando a adoção da variável idade como variável de controle, a

[7]Estão disponíveis no TCU os dados do cadastro único de março, junho, setembro e dezembro a partir do ano de 2013. Como ocorre um processo de revisão cadastral no início de cada ano, o mês de junho foi escolhido por representar um momento mais estável do cadastro, logo após as atualizações terem sido concluídas.

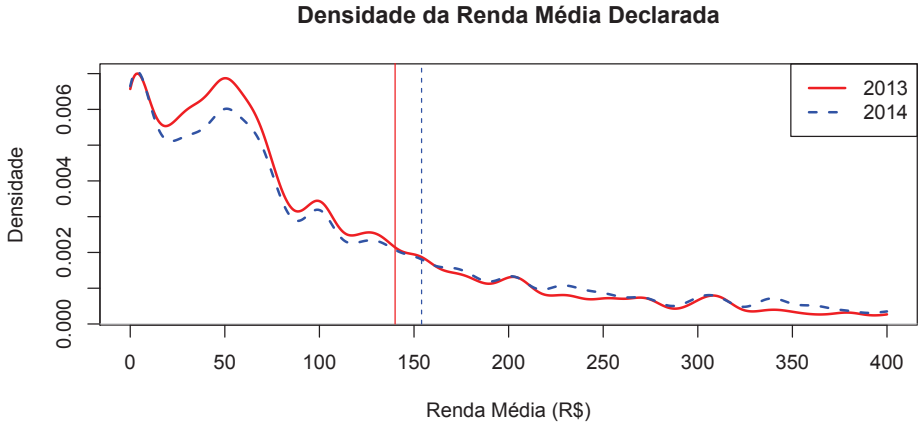

Figura 3.14: Evidência empírica de manipulação da renda declarada.

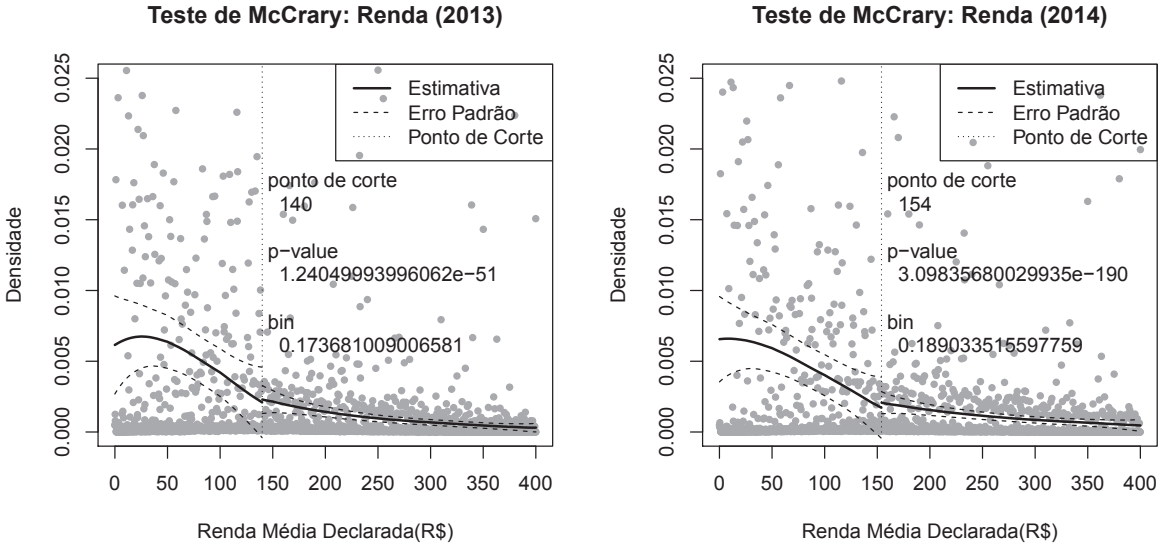

Figura 3.15: Teste de densidade de McCrary para a renda declarada.

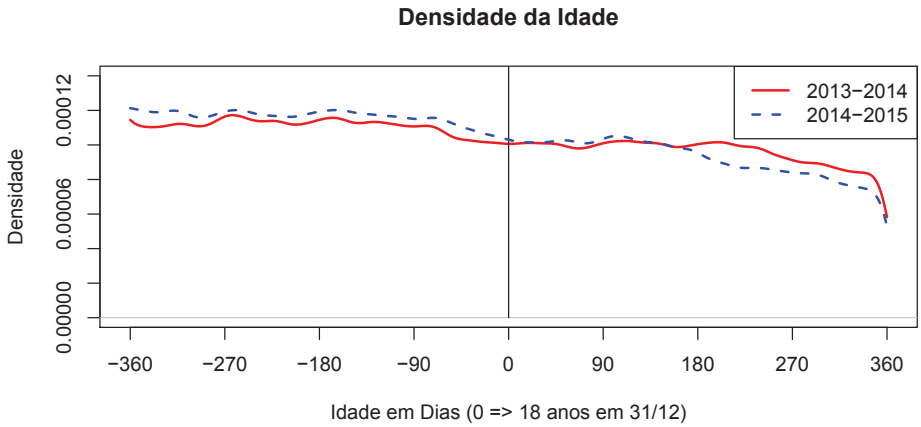

Figura 3.16: Densidade da idade em dias.

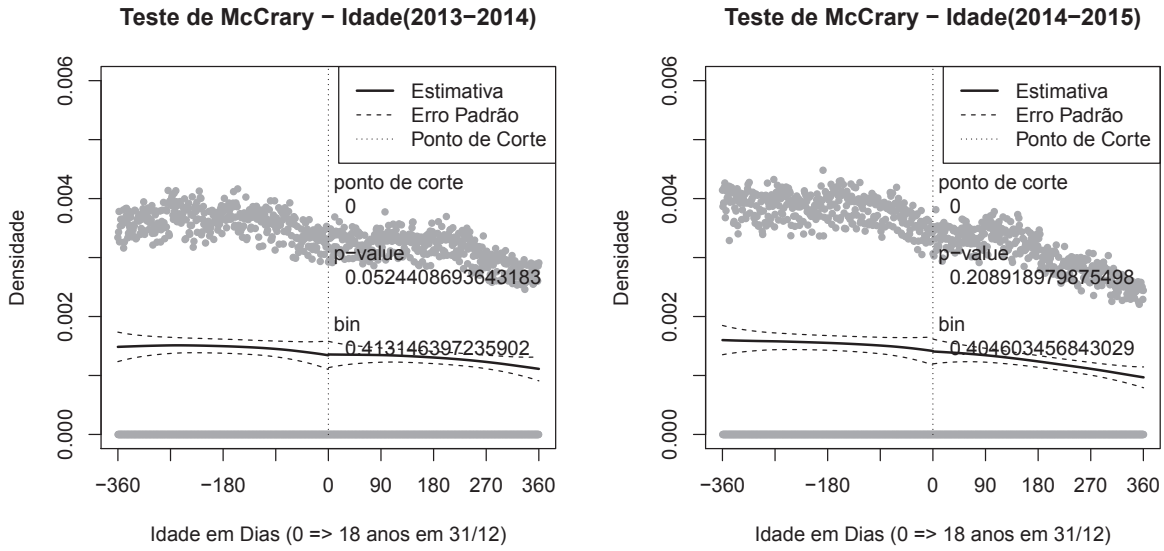

Figura 3.17: Teste de densidade de McCrary para a idade em dias.

Figura 3.18 ilustra o modelo que foi utilizado nas análises. Os jovens posicionados à esquerda do ponto de corte, ou seja, os que completaram 18 anos após 31/12/2013, possuem idade em dias menor que zero e fazem parte do grupo de controle (aqueles que permanecem no programa). Já os jovens posicionados à direita do ponto de corte, possuem idade positiva e fazem parte do grupo de tratamento (aqueles que saíram do programa).

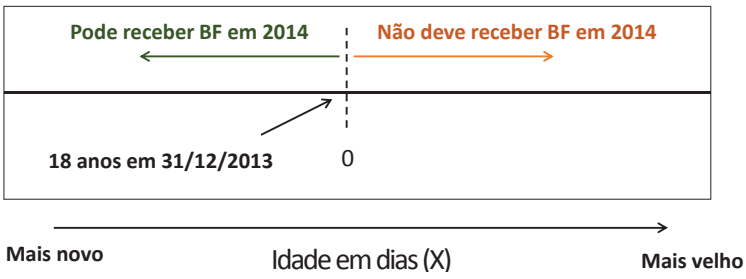

Figura 3.18: Ilustração do Modelo.

Resultados em Relação à Variável de Controle

Imbens e Lemieux [47] propõem que seja analisada graficamente a relação entre a variável de controle e os resultados de interesse, para tentar verificar a existência de algum indicativo visual de impacto no ponto de corte. Segundo os autores, quando não há indicativo visual de descontinuidade, dificilmente os métodos mais avançados de cálculo apontarão para existência de impacto significativo. O método proposto considera divisão da amostra em partes como em um histograma. Cada parte é chamada de *bin*. A divisão em partes

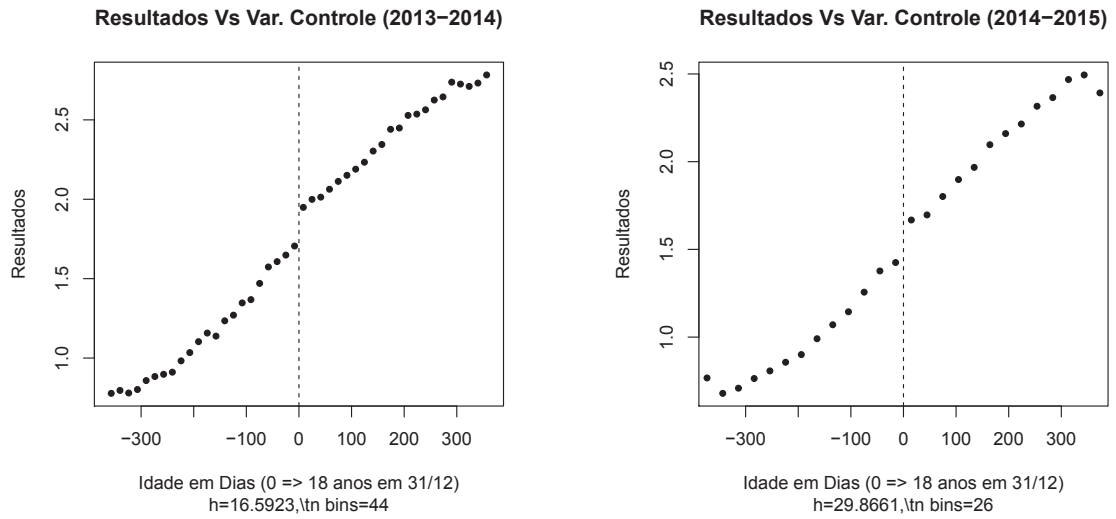

Figura 3.19: Resultados Vs Controle.

deve ser feita de forma que cada um dos *bins* apenas possua dados de um dos lados do ponto de corte.

No presente trabalho, conforme detalhado na Seção 3.1, as variáveis disponíveis no *dataset* final são quantidade de meses trabalhados no ano, a maior remuneração mensal recebida pelo jovem e o total de remunerações recebidas no ano. No presente trabalho optou-se por avaliar o resultado em termos da quantidade de meses trabalhados no mercado formal no ano após a saída do programa. A quantidade de meses trabalhados, além de possuir menor variância, ainda apresenta a vantagem de eliminar diferenças salariais regionais, como por exemplo, os diferentes salários mínimos praticados em alguns estados brasileiros.

A Figura 3.19 apresenta o resultado da análise para os *datasets* correspondentes a 2013-2014 e 2014-2015. Para elaboração dos gráficos foi utilizada a função disponibilizada pelo pacote *rddtools* [39]. A largura dos *bins* foi definida de acordo com a abordagem de seleção de *bandwidth* para Regressão Local por Mínimos Quadrados de Ruppert [85]. Em ambos os *datasets* existe indicativo empírico de existência de impacto no ponto de corte.

Atribuição do Tratamento em Relação à Variável de Controle

Imbens e Lemieux [47] também sugerem que seja verificada a probabilidade de atribuição do tratamento em relação à variável de controle. Para que seja possível a aplicação do RDD é necessário que haja uma descontinuidade dessa probabilidade no ponto de corte. O tipo de descontinuidade determinará a abordagem a ser utilizada (*sharp* ou *fuzzy*) conforme visto no Capítulo 2.

De modo análogo à seção anterior, foi utilizada a função disponibilizada pelo pacote *rddtools* [39] para traçar os gráficos. O tratamento considerado nesse trabalho é a saída

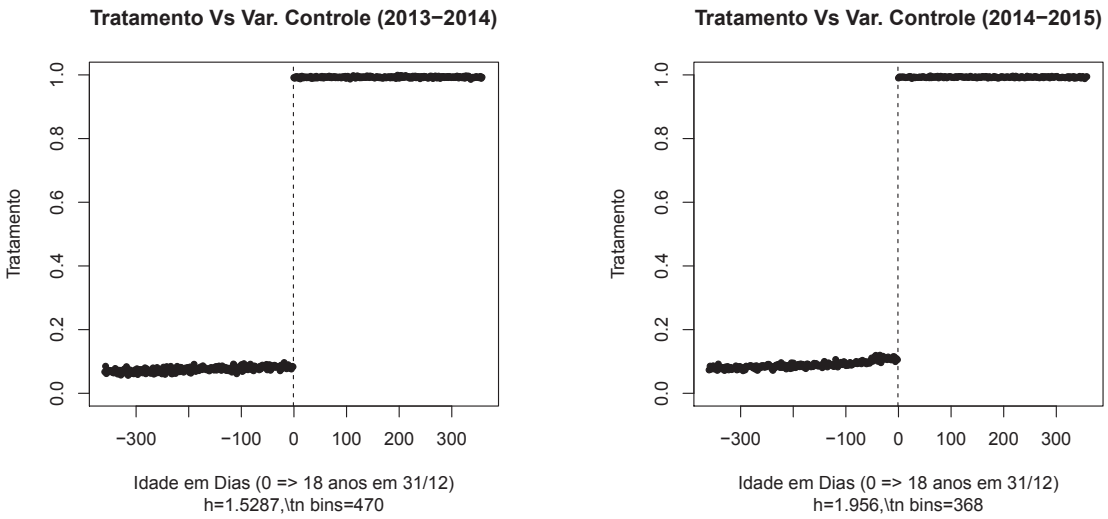

Figura 3.20: Atribuição do Tratamento Vs Controle.

do programa. A Figura 3.20 apresenta o resultado da análise para os *datasets* 2013-2014 e 2014-2015. O comportamento da probabilidade de tratamento é bastante similar em ambos os períodos. Observa-se, como esperado de acordo com as regras do PBF, que todos os jovens que completam 18 anos no ano anterior deixam o programa no próximo ano (p=1). Entretanto, nem todos que poderiam continuar recebendo o benefício, de fato recebem sua parcela (p>0 e p<1). Isso também é esperado, porque existem outros critérios que provocam a saída do programa, como aspectos relacionados ao cumprimento de condicionalidades e à renda. Esse comportamento da probabilidade do tratamento indica a natureza *fuzzy* desse problema.

Análise de Covariáveis

Segundo Imbens e Lemieux [47], covariáveis adicionais podem ser utilizadas para aumentar a precisão das estimativas de análises RDD conforme visto no Capítulo 2. Adicionalmente, os autores afirmam que a existência de descontinuidades em covariáveis no ponto de corte pode levantar questionamentos quanto à plausibilidade do modelo.

Com a abordagem utilizada para montar os *datasets* (Seção 3.1), foi possível extrair as seguintes covariáveis de interesse:

- município - código IBGE do município do domicílio da família do beneficiário;
- gênero - gênero do beneficiário (Masculino, Feminino); e
- local - localização do domicílio (Urbano, Rural).

A variável categórica município, por apresentar milhares valores possíveis, não é interessante para ser utilizada nas análises por segmento, pois representaria uma diminuição

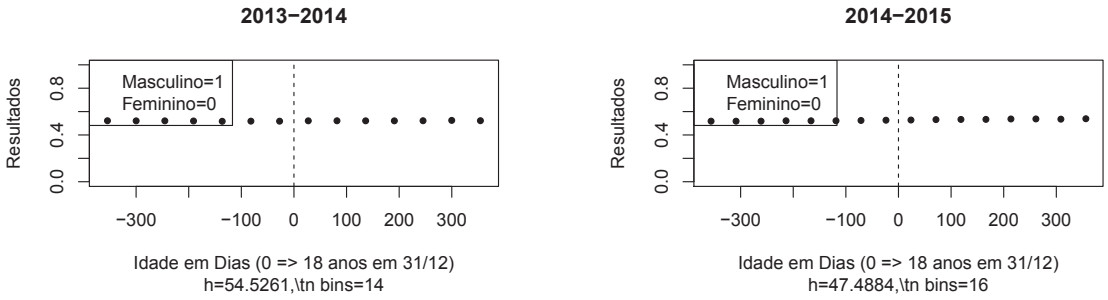

Figura 3.21: Covariável Gênero Vs Controle.

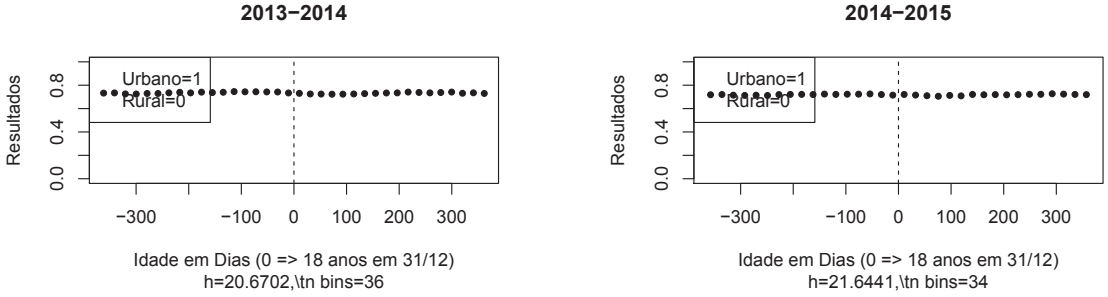

Figura 3.22: Covariável Local Vs Controle.

muito grande do tamanho da amostra, com impactos no intervalo de confiança das estimativas. Entretanto, a partir dela, utilizando base de dados de malhas cartográficas do IBGE [8], é possível obter a informação da região brasileira, na qual se localiza o município (Norte, Nordeste, Centro-Oeste, Sudeste ou Sul). Dessa forma, o conjunto final de covariáveis é o seguinte: região, gênero e local.

Para realização das análises, as variáveis categóricas binárias foram transformadas em variáveis numéricas (0 ou 1). Para a região, foi utilizada a técnica de variáveis *dummies*[9]. Assim como nas subseções anteriores, utilizamos o pacote *rddtools* [39] para inspeção visual do comportamento das covariáveis no ponto de corte. A Figura 3.21 apresenta a análise da covariável gênero, enquanto a Figura 3.22 apresenta a análise da variável local. Por fim, a Figura 3.23 apresenta a análise dos diversos níveis da covariável região.

Por meio da inspeção visual dos gráficos das diversas covariáveis, não é possível verificar indícios de descontinuidades no ponto de corte. Assim, considerando as covariáveis disponíveis, não há motivos para questionar a plausibilidade do modelo.

[8]Disponível em: `ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2015/Brasil/BR/`.

[9]A técnica de variáveis *dummies* é muito utilizada em estatística e na ciência de dados como forma de representar variáveis categóricas por meio de variáveis numéricas. Consiste em criar uma variável binária (0 ou 1) para cada nível da variável categórica, de modo que, para cada observação, apenas a variável binária correspondente ao nível da variável categórica assume o valor um e todas as outras assumem o valor 0.

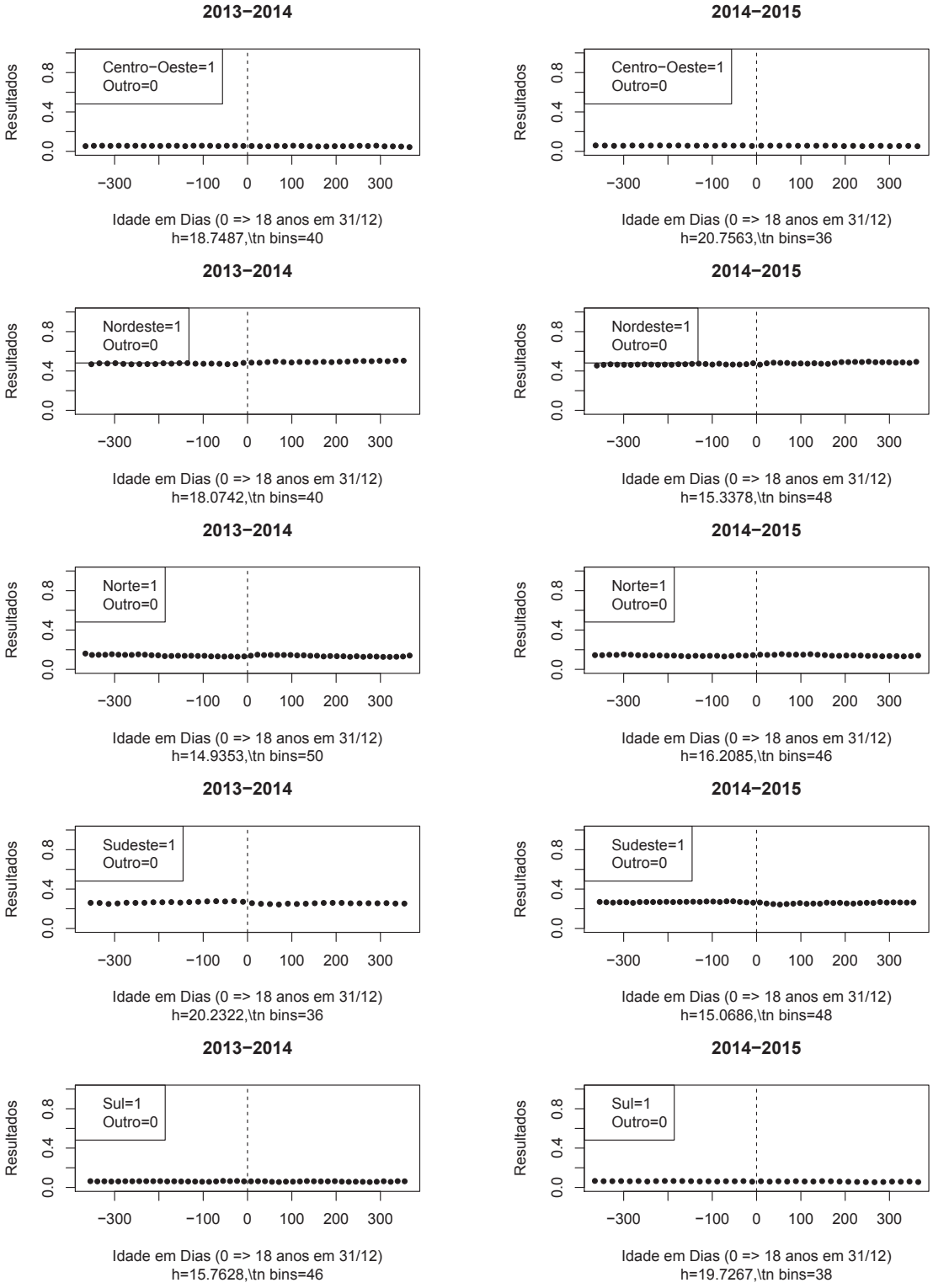

Figura 3.23: Covariável Região Vs Controle.

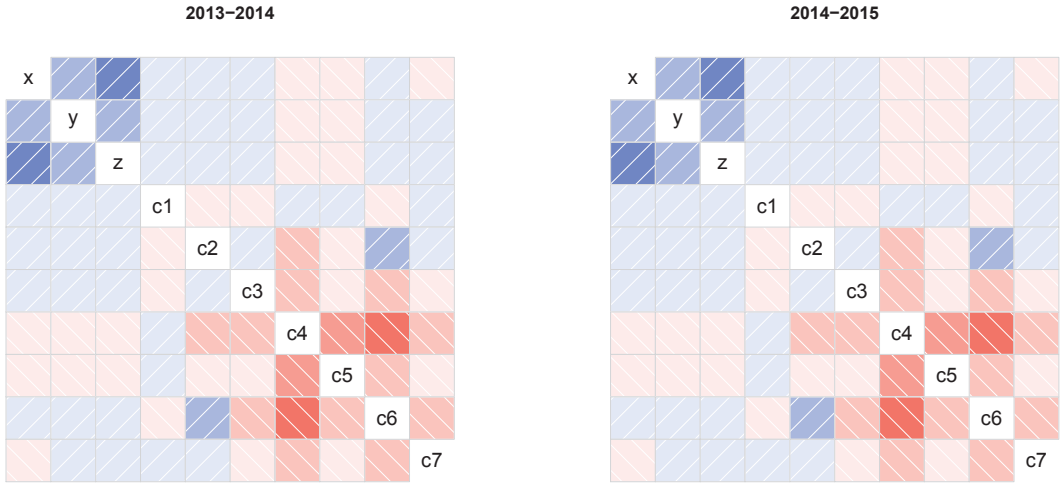

Figura 3.24: Matrizes de Correlação (azul - positiva, vermelho - negativa)

Acerca da possibilidade do uso das covariáveis disponíveis para aumento da precisão das estimativas e redução do intervalo de confiança, foi feita a análise da correlação das diversas variáveis. A Figura 3.24 apresenta a matriz de correlação entre as variáveis, sendo x a variável de controle, y o resultado, z o indicador de tratamento e c_i as covariáveis (1-gênero, 2-local, 3-centro-oeste, 4-nordeste, 5-norte, 6-sudeste, 7-sul). Nas matrizes, os tons azuis representam correlações positivas e os tons vermelhos representam correlações negativas. Quanto mais escuro o tom, maior é o valor absoluto da correlação.

As maiores correlações com o resultado no trabalho formal(y) ocorrem com as variáveis idade(x) e tratamento(z). A correlação entre idade e trabalho formal (0,23) é esperada, visto que, entre os jovens, quanto maior a idade, maiores as possibilidades de se obter um emprego, conforme ficou evidenciado na Figura 3.19. A correlação entre tratamento e trabalho (0,25) é um indicativo empírico de que pode haver algum impacto do programa sobre a variável estudada.

A correlação que ocorre entre a idade e o tratamento (0,30), também é esperada, visto que os jovens são obrigados a deixar o programa no ano subsequente ao que completam 18 anos.

Apesar de existirem algumas correlações importantes entre as demais covariáveis (c_{1-7}), não se percebe o mesmo nível de correlação dessas covariáveis com o resultado. A maior correlação observada entre as covariáveis e o resultado ocorreu com a região nordeste (c_4), com um valor de -0,05, que não é muito forte.

O fato de não haver uma covariável adicional com forte correlação com o resultado indica que, nesse estudo, o uso de covariáveis na estimativa pode não contribuir significativamente para a melhoria das estimativas. Mesmo assim, foi feita uma estimativa RDD usando a covariável c_4, que apresentou maior correlação com o resultado e foi feito

um comparativo com a estimativa sem covariáveis. Os resultados são apresentados e discutidos no Capítulo 4.

Conclusão da análise gráfica

Considerando os testes e análises realizados nesta seção, tendo como variável de controle a idade em dias, não foram encontrados indícios que possam levantar suspeições contra a plausibilidade do modelo de análise RDD proposto.

Adicionalmente, não foram encontradas covariáveis suficientemente correlacionadas com o resultado para justificar sua utilização no modelo com objetivo de melhoria da estimativa e redução do intervalo de confiança.

3.3.2 Estimação do Impacto

A estimação corresponde à aplicação do método propriamente dito e análise dos resultados. A variável de controle considerada foi a idade do jovem em dias e o resultado medido foi o número de meses trabalhados no mercado formal no ano subsequente. Como covariáveis candidatas foram consideradas região, gênero e local, sendo que nenhuma delas apresentou correlação significativa com o resultado. Como verificado na etapa anterior, a abordagem utilizada foi a *Fuzzy Regression Discontinuity Design (FRDD)* e como critério de seleção de *bandwidth* foi utilizado o método IK descrito no Capítulo 2.

As análises foram realizadas para os dois *datasets* obtidos na etapa descrita na Seção 3.1. Inicialmente, foi feita a estimação RDD sem considerar as covariáveis e, em seguida, foi feita nova estimação com a covariável que apresentou maior correlação com a variável de resultado (c_4 - Região Nordeste). Os resultados obtidos nas duas estimativas foram comparados. Adicionalmente, foram realizadas análises complementares sobre diferentes segmentos da amostra para avaliar os impactos do PBF em diferentes sub-populações, como entre homens e mulheres, em áreas urbanas e rurais e em diferentes regiões do país. Todas as estimações foram realizadas usando a versão paralelizada do algoritmo RDD descrito na Seção 3.2.

Os resultados obtidos nessa etapa são apresentados e discutidos no Capítulo 4.

3.3.3 Testes de Robustez do Modelo

Imbens e Lemieux [47] sugerem a realização de testes para verificar a robustez do modelo. Uma das verificações sugeridas é a ocorrência de outras descontinuidades do resultado além do ponto de corte. A ocorrência de tais descontinuidades pode levantar dúvidas sobre a validade do modelo, uma vez que na ausência do tratamento o resultado deve ser

suave e contínuo. Os autores sugerem que sejam realizados testes de descontinuidade na mediana das subamostras do *bandwidth* dos dois lados do ponto de corte.

Outro teste recomendado é a sensibilidade do modelo em relação ao *bandwidth*. Além da estimativa com o *bandwidth* ótimo, os autores recomendam a realização de estimativas com metade e o dobro do *bandwidth* para comparação dos resultados. A ferramenta utilizada para a estimação já contempla a realização destas estimativas complementares automaticamente.

Um teste adicional realizado, específico para este estudo de caso, foi a verificação de uma eventual ocorrência de descontinuidade no nível de preenchimento do CPF no cadastro único. Caso ocorra tal descontinuidade, ou seja, caso os jovens que completam 18 anos no ano corrente possuam nível de preenchimento de CPF maior que os demais, isto pode contribuir para o aumento da taxa de pareamento do Cadastro Único com a RAIS para essas pessoas, introduzindo viés no modelo. A utilização das técnicas de *record linkage* descritas na Seção 3.1.2 contribuem para reduzir esse eventual viés, uma vez que introduzem outros critérios de pareamento além das chaves CPF e PIS, mas, mesmo assim, o teste foi realizado.

Estes testes foram realizados e os resultados são apresentados e discutidos no Capítulo 4.

Capítulo 4

Resultados Obtidos

Nesse capítulo são apresentados e discutidos os resultados obtidos em cada uma das 3 fases previstas na metodologia utilizada para o presente projeto: Obtenção e Preparação de Dados, *Refactoring* da ferramenta RDD e Estimação.

4.1 Obtenção e Preparação de Dados

O processo de obtenção e preparação de dados descrito no Capítulo 3 demonstrou ser eficiente e prático, possibilitando a extração de *datasets* com diferentes configurações de período por meio de simples parametrização. Ao final do processo foram montados dois *datasets*, referentes aos períodos 2013-2014 e 2014-2015 com respectivamente 13.272.022 e 13.403.743 observações. Esses números representam o *dataset* total, incluindo registros do Cadastro Único sem correspondência na RAIS.

A etapa de *record linkage* também funcionou a contento, tendo identificado aproximadamente 16,6% mais registros na RAIS que o *join* (junção) simples por CPF ou PIS. O ganho aqui apresentado foi calculado com base na diferença entre registros do *dataset* final com e sem informação perfeita de CPF ou PIS. Para ilustrar, caso um registro do *dataset* final tenha informação de CPF ou PIS equivalentes no Cadastro Único e na RAIS ele será contado como junção simples. Os registros com *record linkage* correspondem aos anteriores acrescidos daqueles que não possuem relação perfeita nem de PIS nem de CPF. A Tabela 4.1 apresenta esses resultados, em detalhes, para os dois períodos coletados.

Tabela 4.1: Resultados da Etapa de Obtenção e Preparação de Dados

Período	Total	Junção Simples	Com *Record Linkage*	Ganho
2013-2014	13.272.030	307.906	361.404	17,37%
2015-2015	13.403.742	252.345	292.532	15,93%

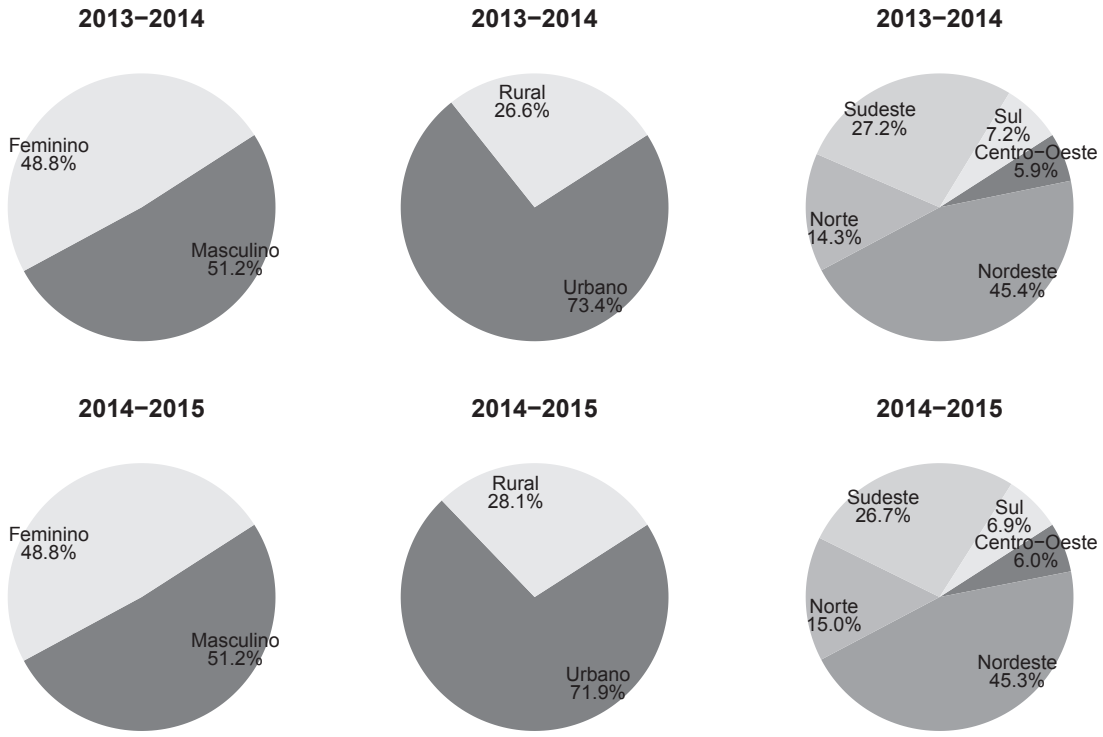

Figura 4.1: Distribuição das observações entre as subpopulações de interesse.

4.1.1 Visão Geral dos Dados Coletados

A Figura 4.1 contém uma série de gráficos que apresentam a distribuição das observações dos *datasets* entre as subpopulações de interesse. Chama a atenção a concentração de observações na área urbana e no nordeste do Brasil.

A média geral da quantidade de meses trabalhados entre os jovens de 17 e 18 anos foi de 1,72 em 2013-2014 e 1,47 em 2014-2015. A Figura 4.2 apresenta a variação dessa média entre as subpopulações de interesse. Como esperado, as menores médias são encontradas nas áreas rurais e nas regiões norte e nordeste, em razão da menor oferta de empregos formais para esses subgrupos. Entretanto, assim como ocorreu no *dataset* como um todo, observa-se significativa redução da média em todas as categorias no segundo período analisado. Cumpre lembrar que o acesso ao trabalho formal é sempre medido no segundo ano de cada período, assim, a redução geral do nível de emprego dos jovens do PBF ocorreu efetivamente no ano de 2015.

Para confirmar a consistência dessas medidas, foi realizada uma análise sobre a base de dados da RAIS disponível no TCU. A Figura 4.3 obtida a partir de dados extraídos diretamente da RAIS mostra a evolução do número de empregados distintos declarados entre os anos de 2012 e 2015. Note-se a expressiva redução no número de empregados declarados que ocorreu em 2015, onde houve um retrocesso para valores inferiores aos

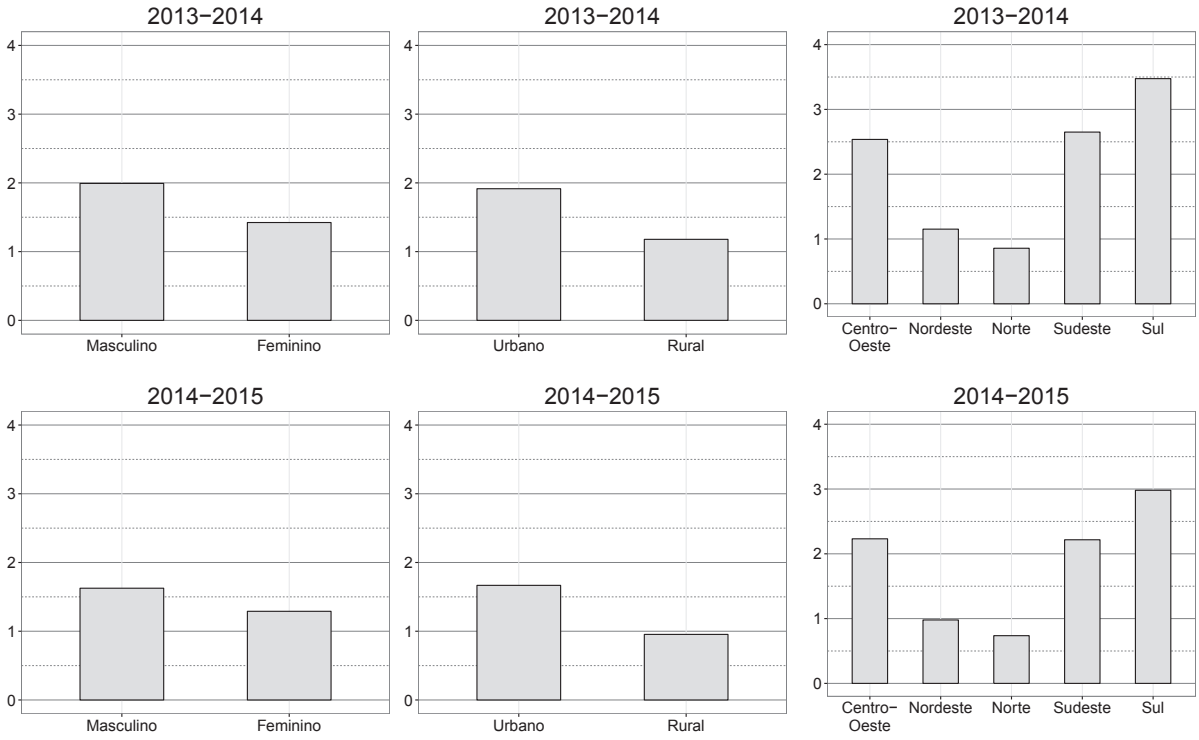

Figura 4.2: Média de meses trabalhados entre as subpopulações de interesse

observados em 2012. Esse fato coincide com o agravamento da crise econômica que ocorreu no Brasil naquele ano, evidenciando os possíveis impactos da crise sobre o mercado de trabalho e pode justificar os efeitos sobre a empregabilidade dos jovens beneficiários do PBF em 2015.

Com objetivo de verificar mais detalhadamente os efeitos regionais da crise econômica sobre o mercado de trabalho em 2014 e 2015, foi feito um levantamento do comportamento da geração de empregos no de dados a partir da base de dados do Cadastro Geral de Empregados e Desempregados (CAGED) disponível no TCU. O CAGED é de

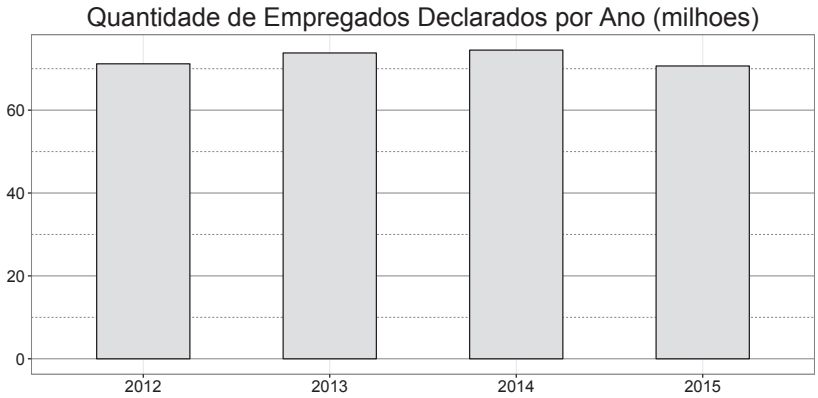

Figura 4.3: Evolução da quantidade de empregados declarados na RAIS.

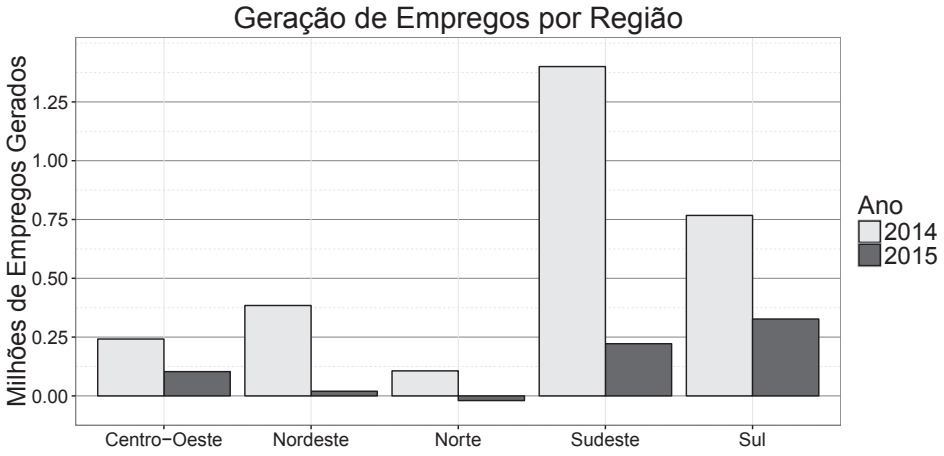

Figura 4.4: Geração de empregos em 2014 e 2015 por região conforme dados do CAGED.

responsabilidade do Ministério do Trabalho e contém as informações sobre admissões, desligamentos e movimentações de empregados que as empresas são obrigadas a informar mensalmente. Como geração de emprego foi considerada a diferença entre as admissões e as demissões ocorridas durante todo o ano na região sob estudo. Aposentadorias e outras movimentações não foram consideradas. A Figura 4.4 apresenta os resultados. Chama a atenção a redução generalizada na geração de empregos que ocorreu em 2015. Também merece destaque a diferença entre o volume de empregos gerados no sul e sudeste e nas demais regiões do Brasil. Note-se que, em 2015, a geração de empregos na região norte foi negativa, ou seja, ocorreram mais demissões que admissões no ano.

Considerando que a oferta de empregos pode influenciar o acesso dos jovens beneficiários do PBF ao mercado de trabalho, foi feita uma análise da relação entre a quantidade de jovens beneficiários de 17 e 18 anos e a geração de empregos obtida no CAGED. Para avaliar o comportamento dessa relação em regiões com diferentes graus de desenvolvimento, foi feita um análise segmentada por mesoregiões, conforme malha geográfica provida pelo IBGE[1]. Os mapas da Figura 4.5 apresentam os resultados. A escala de cores foi montada de tal forma que cada tom representa um decil da relação entre geração de empregos e quantidade de jovens. Quanto maior o valor da relação, maior é o numero de empregos gerados para cada jovem. Tons de vermelho mais escuro representam mesoregiões com menor oferta de empregos. Mais uma vez, percebe-se a redução da oferta de empregos que ocorreu em 2015, entretanto, com a disposição dos dados no mapa é possível perceber que os pontos de menor oferta concentram-se em regiões mais afastadas dos grandes centros urbanos, principalmente no norte e no interior do nordeste.

[1]A malha de mesoregiões do Brasil está disponível em `ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2015/Brasil/BR/`.

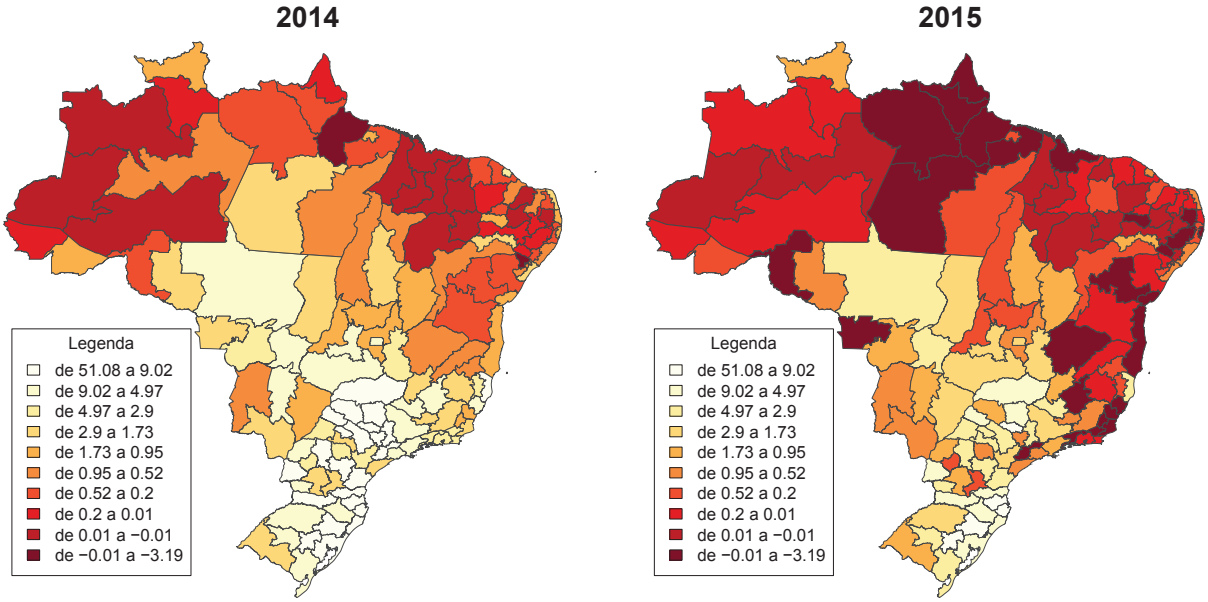

Figura 4.5: Geração de empregos sobre quantidade de jovens beneficiários.

4.2 *Refactoring* da ferramenta RDD

Conforme relatado no Capitulo 3, os testes do *Refactoring* demonstraram que as estimativas da versão original do software e da versão refatorada foram exatamente as mesmas, uma vez que a refatoração não afetou a lógica do algoritmo. A Tabela 4.2 detalha os ganhos de velocidade de execução para diversos tamanhos de amostra. O ganho de velocidade corresponde ao tempo de execução da versão original sobre o tempo de execução da versão paralelizada. Cumpre destacar que a versão paralelizada só apresenta melhor tempo de execução com amostras com mais de um milhão de observações.

É possível observar que o ganho de desempenho não é linear com o aumento do número de processadores, em razão do custo da infraestrutura de processamento paralelo. Os maiores ganhos são observados com os maiores tamanhos de amostra. Com o conjunto de dados completo e usando 10 núcleos de processador a versão paralelizada foi 5,91 vezes mais rápida que a versão sequencial.

4.3 Estimação

Nesta seção são apresentados e discutidos os resultados da aplicação da abordagem RDD sobre os *datasets* obtidos na fase de obtenção e preparação de dados (2013-2014 e 2014-2015). Foram realizadas estimativas sobre todo o conjunto de dados e também sobre subpopulações de interesse. Todas as estimativas foram realizadas utilizando a versão paralelizada do software.

Tabela 4.2: Ganhos de velocidade: tempo de execução sequencial sobre paralelo

Tamanho da Amostra	Ganho de Velocidade		
	2 cores	5 cores	10 cores
663.601	0,87	0,90	0,84
1.327.203	1,18	1,47	1,17
1.990.805	1,16	1,22	1,59
2.654.407	1,20	1,34	1,34
3.318.009	1,71	2,89	3,69
3.981.611	1,53	2,13	2,23
4.645.212	1,52	1,81	1,80
5.308.814	1,68	2,36	2,60
5.972.416	1,87	2,68	3,18
6.636.018	1,54	1,95	1,91
7.299.620	1,43	1,83	2,06
7.963.222	1,84	3,36	4,37
8.626.824	1,90	3,82	4,70
9.290.425	1,52	2,28	2,76
9.954.027	1,74	3,61	5,04
10.617.629	1,91	3,53	4,87
11.281.231	2,02	3,73	5,39
11.944.833	1,58	2,58	3,50
12.608.435	1,99	3,81	5,06
13.272.037	2,00	3,81	5,91

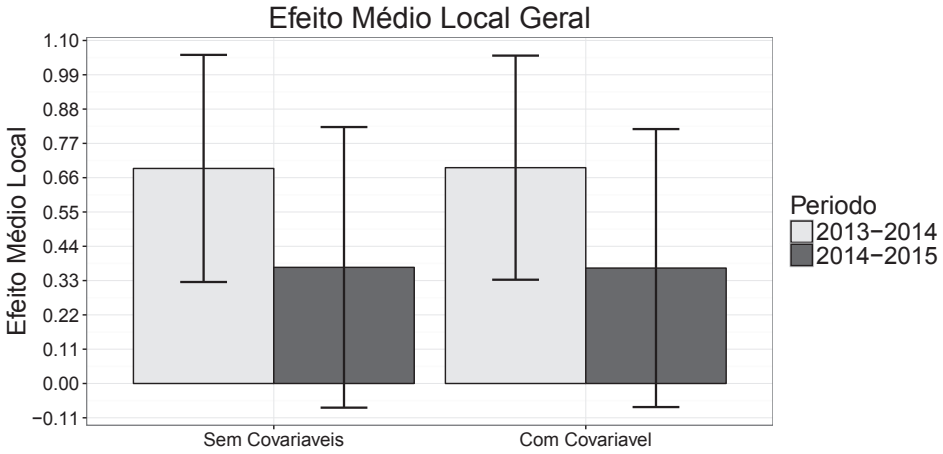

Figura 4.6: Resultados gerais com e sem covariáveis

Tabela 4.3: Detalhe do Resultado Geral

Período	Cov?	Amostra	Estimativa	BW	Ci_1	Ci_2
2013-2014	Não	13.272.030	0,6896386	3,376580	0,32546195	1,0538152
2013-2014	Sim	13.272.030	0,6921296	3,376580	0,33278449	1,0514748
2014-2015	Não	13.403.742	0,3724651	2,996246	-0.07747274	0,8224029
2014-2015	Sim	13.403.742	0,3701491	2,996246	-0,07559233	0,8158905

4.3.1 Resultado Geral

A estimação do efeito da saída do PBF sobre a quantidade de meses trabalhados pelos jovens no mercado formal foi conduzida para ambos os *datasets*, com e sem o auxílio da covariável referente à região nordeste, conforme previsto na metodologia (Capítulo 3). A Figura 4.6 consolida o resultado obtido. Na figura, as barras representam as estimativas de impacto e os segmentos de reta verticais representam os intervalos de confiança das estimativas. Quando o intervalo de confiança cruza o zero, não se pode afirmar a existência de impacto, com confiança de 95%.

A Tabela 4.3 apresenta os mesmos resultados do gráfico de forma tabular. Na tabela, a coluna *Cov*? indica se a covariável foi utilizada na estimativa, BW indica qual foi o bandwidth obtido pelo método IK e Ci_1 e Ci_2 indicam, respectivamente, o início e o fim do intervalo de confiança.

Como esperado, em razão da baixa correlação com o resultado medido, visualmente não é perceptível nenhuma diferença entre as estimativas com e sem o auxílio da covariável (Figura 4.6). Entretanto, ao analisar os resultados apresentados de forma tabular (Tabela 4.3), é possível verificar que houve discreta redução na amplitude dos intervalos de confiança em ambas as estimativas realizadas quando a covariável foi utilizada. A pequena diferença entre as estimativas com e sem a covariável não foi considerada suficiente para justificar a utilização do artifício, de tal forma que nas próximas subseções a covariável

não mais foi utilizada.

A cerca do resultado geral em si, é evidente a existência de impacto da saída antecipada do programa no acesso ao mercado formal no ano de 2014 (o acesso ao mercado é verificado no segundo ano do período). A estimativa de 0,69 mês trabalhado a mais pelos que deixaram o programa, considerando um intervalo de 12 meses, indica que aqueles jovens, em média, trabalharam 5,8% a mais que os que deixaram o PBF. Entretanto, a mesma segurança de avaliação não foi obtida no ano de 2015, no qual houve significativa redução da estimativa e ampliação do intervalo de confiança, que passou a cruzar ligeiramente o zero. Dentre as razões que justificariam essa mudança de comportamento, pode-se citar o agravamento da crise econômica que ocorreu no Brasil em 2015. Nesse ano, observou-se significativa redução na oferta de trabalhos formais, com reflexos no acesso ao mercado de trabalho formal pelos jovens do PBF conforme discutido na Seção 4.1.1. A redução na oferta geral de emprego formal justifica a queda no impacto relativo ao acesso de jovens no mercado de trabalho, visto que uma vez que há menos oferta de emprego, aqueles jovens que estariam em condições ou dispostos a começar trabalhar teriam mais dificuldades de encontrar emprego.

A estimativa de impacto de aproximadamente 0,69 identificada no período 2013-2014 significa que, em média, aqueles jovens que deixaram o programa mais cedo trabalharam no mercado formal, em 2014, 0,69 mês a mais que aqueles que permaneceram no programa, o que equivale a 20,7 dias.

Essa constatação, por si só, não é suficiente para determinar se o programa tem um impacto positivo ou negativo sobre os jovens. Uma discussão sobre esse assunto é apresentada no final dessa seção.

4.3.2 Resultados por Gênero

A Figura 4.7 apresenta os resultados das estimativas de impacto segmentadas por gênero. Observa-se impacto positivo, com confiança de 95% entre jovens do sexo masculino similar nos dois períodos analisados. Entretanto, entre jovens do sexo feminino, o impacto estimado foi positivo apenas no período 2013-2014. De fato, o impacto em 2014-2015 foi praticamente nulo entre as meninas.

Como em 2013-2014, o impacto estimado foi similar entre jovens de ambos os sexos, pode-se supor que a diferença no resultado geral nos dois períodos foi muito concentrada no sexo feminino. É plausível a hipótese de que este fenômeno decorra em razão de uma possível preferência dos empregadores formais por pessoas do sexo masculino em um cenário de escassez de oferta de emprego e excesso de mão de obra, como o que ocorre em momentos de crise econômica.

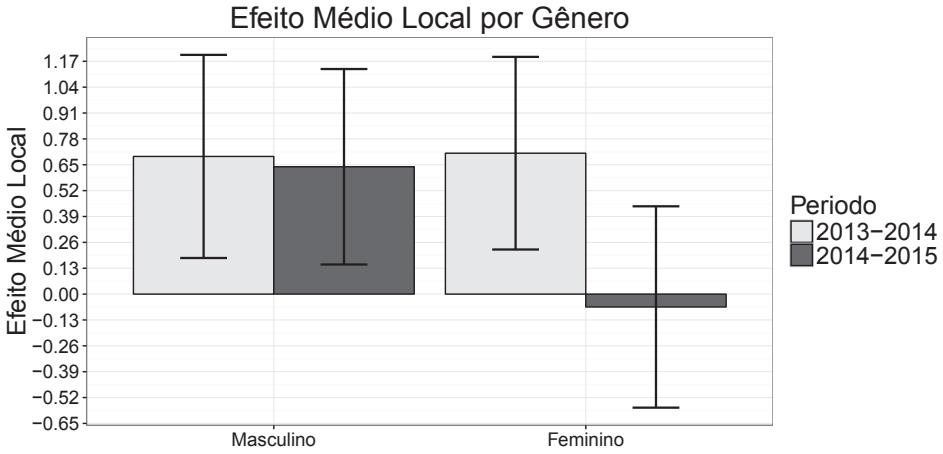

Figura 4.7: Resultados por gênero

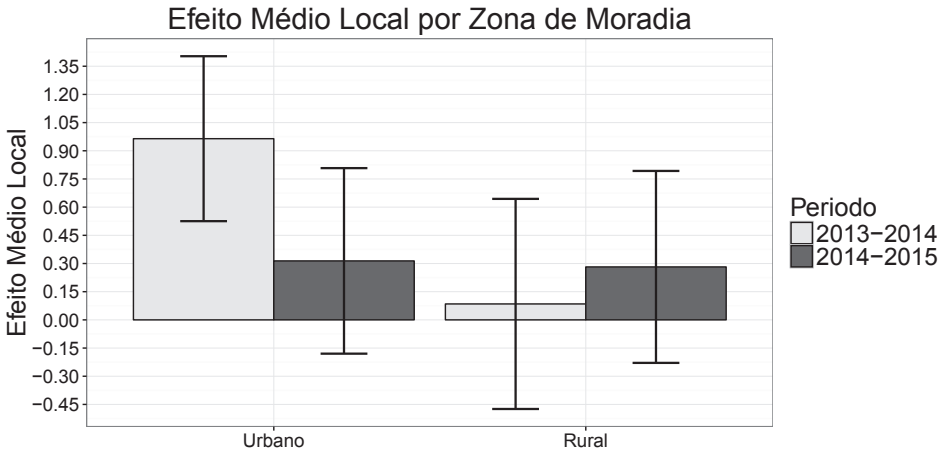

Figura 4.8: Resultados por zona de moradia

Observa-se também, pelos resultados, o aumento da amplitude dos intervalos de confiança em razão da redução do número de observações com a divisão da amostra, já que, no caso da segmentação por gênero, a redução da amostra é de aproximadamente 50%.

4.3.3 Resultados por Zona de Moradia

A Figura 4.8 apresenta os resultados das estimativas de impacto segmentadas pela zona do domicílio da família do jovem. Observa-se impacto positivo, com confiança de 95%, apenas em jovens que moram em áreas urbanas, no período 2013-2014. A redução do impacto que ocorreu no período de 2014-2015 para os jovens da zona urbana, também pode ser justificada pela crise de 2015.

Já entre os jovens da zona rural, nada se pode afirmar sobre o impacto, em razão do intervalo de confiança das estimativas. A ampliação do intervalo de confiança na zona rural é justificada pela menor quantidade de observações, como visto na Seção 4.1.1.

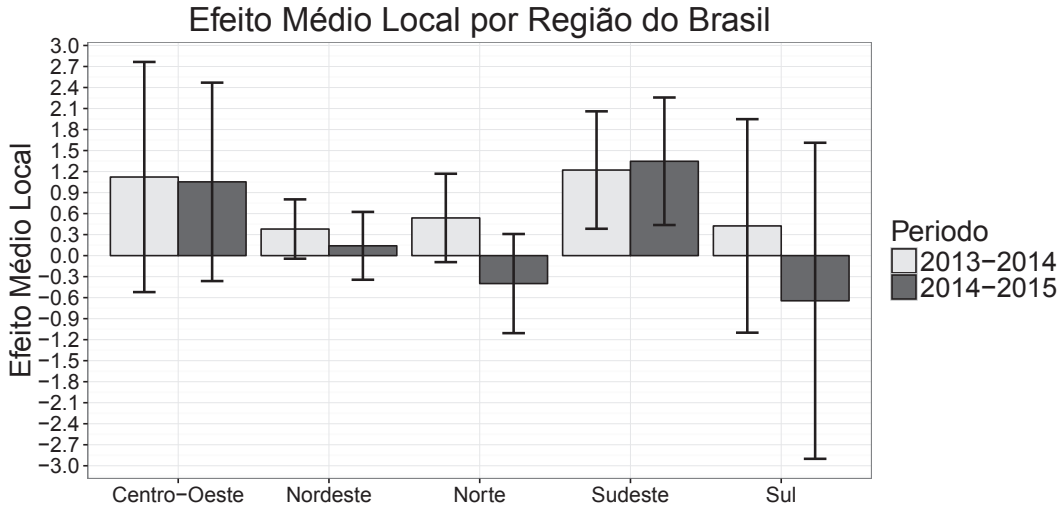

Figura 4.9: Resultados por região geográfica

4.3.4 Resultados por Região Geográfica

A Figura 4.9 apresenta os resultados das estimativas de impacto segmentadas pela região geográfica do Brasil. Em função da redução do tamanho da amostra e consequente aumento da amplitude dos intervalos de confiança, pouco se pode afirmar para a maioria das categorias. De maneira geral, com exceção do sudeste, percebe-se redução dos impactos estimados em 2014-2015 quando comparado com 2013-2014.

As regiões sul e centro-oeste, que possuem menos observações, são as que apresentam os intervalos de confiança mais amplos, de modo que não se pode extrair conclusões sobre o impacto estimado, especialmente no sul, onde o intervalo de confiança está praticamente centrado no zero.

No norte e no nordeste, as regiões com mais observações, são percebidos os menores intervalos de confiança. Para essas regiões, no período 2013-2014, existe tendência de haver um leve impacto positivo, uma vez que o intervalo de confiança apenas cruza o zero em sua extremidade inferior, mas essa afirmação não tem confiança de 95%. Já em 2014-2015, uma vez que a estimativa central é consideravelmente menor, nada pode se afirmar sobre a existência de impacto.

Apenas no sudeste observa-se um impacto consistente, da ordem de 1,3 meses, com 95% de confiança, nos dois períodos, que, considerando o período de 12 meses, representa um impacto de 10,8%. Esse comportamento pode indicar que, nessa região, não houve reflexo da crise no acesso dos jovens beneficiários ao mercado formal de trabalho.

Tabela 4.4: Teste de sensibilidade ao *bandwidth*

$BWdesc$	$BWval$	Est	Ci_1	Ci_2
Optimal	3,376580	0,6896386	0,3254620	1,0538152
Half	2,025948	0,6677582	0,1988862	1,1366302
Double	6,753160	0,5475404	0,3013327	0,7937482

4.3.5 Testes de Robustez do Modelo

Os testes de robustez tem por objetivo eliminar suspeitas que possam levantar questionamentos sobre o modelo RDD. Nessa seção serão apresentados os testes de robustez realizados, conforme previsto na metodologia proposta.

Teste de sensibilidade ao *bandwidth*

Conforme detalhado no Capítulo 3, a ferramenta utilizada para estimativa RDD já realiza os cálculos referentes à adoção da metade e do dobro do *bandwidth* ótimo. Entretanto, como o *bandwidth* ótimo calculado pelo método IK já é bastante pequeno, da ordem de 3,37 dias, o algoritmo não foi capaz de realizar a estimativa com a metade do *bandwidth*, que seria equivalente a apenas 1,18 dias, em função do baixo número de observações resultante. Assim, no lugar da metade do *bandwidth* (bw), foi utilizado $bw * 0,6$. O teste foi realizado sobre a integralidade do *dataset* 2013-2014, no qual houve indicação de impacto. Na Tabela 4.4 que apresenta os resultados, $BWdesc$ indica o tipo de *bandwidth* utilizado na estimativa, $BWval$ contém a largura do *bandwidth*, Est representa a estimativa obtida e Ci_1 e Ci_2 indicam, respectivamente, o início e o fim do intervalo de confiança..

Como esperado, a redução do *bandwidth* provoca a ampliação do intervalo de confiança (Ci), mas a estimativa (Est) não se altera muito. Também como esperado, o aumento do *bandwidth* provoca a redução do intervalo de confiança, mas introduz viés. Entretanto, as diferentes estimativas se mantêm consistentes em relação à indicação de descontinuidade no ponto de corte, com confiança de 95%.

Assim sendo, o modelo se mostrou robusto ao *bandwidth*.

Teste de descontinuidades adicionais (*placebo test*)

O teste de descontinuidades adicionais, também conhecido como *placebo test*, foi realizado, conforme previsto na metodologia, nos pontos correspondentes à mediana do *bandwidth* dos dois lados do ponto de corte. No lado esquerdo, a mediana calculada foi -2 e no lado direito, 2. Os resultados são apresentados na Tabela 4.5, onde os pontos de corte são os pontos de teste de descontinuidade, correspondentes as medianas do *bandwidth* calculadas

Tabela 4.5: *Placebo test*

Ponto de Corte	Est	Ci_1	Ci_2
-2	-0,185	-0,6182925	0,2483015
2	0,04781	-0,4386258	0,5342403

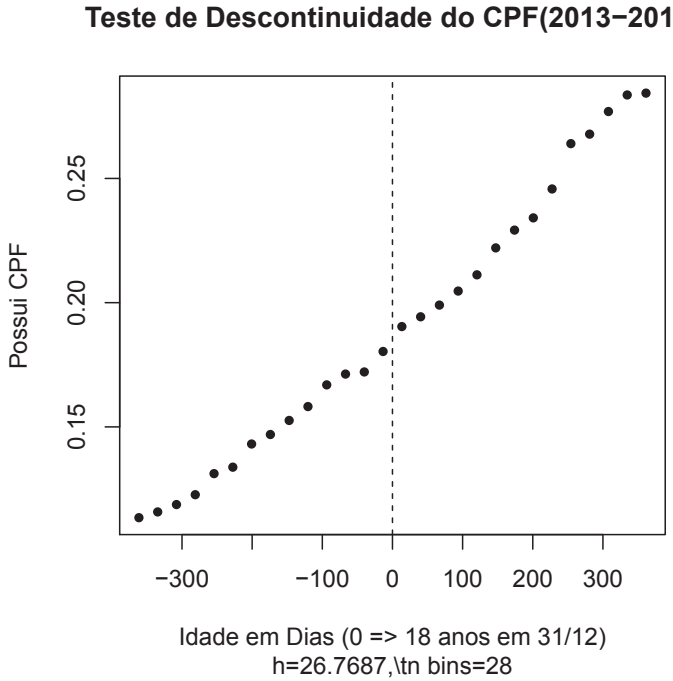

Figura 4.10: Verificação de descontinuidade no nível de informação do CPF

à esquerda e a direita do ponto de corte original, *Est* representa a estimativa obtida e Ci_1 e Ci_2 indicam, respectivamente, o início e o fim do intervalo de confiança.

Como, em ambas as estimativas, os intervalos de confiança cruzam o zero, não se pode afirmar que exista descontinuidade nos pontos do teste, confirmando a plausibilidade do modelo.

Teste de descontinuidade na informação do CPF

Para verificação de descontinuidade no nível de preenchimento do CPF no ponto de corte, foi utilizada a ferramenta visual fornecida no pacote *rdtools*. O *dataset* utilizado foi o 2013-2014. A Figura 4.10 apresenta os resultados. O gráfico não indica existência de descontinuidade aparente que possa comprometer as estimativas.

4.3.6 Discussão sobre os resultados

Nesse tópico são apresentadas algumas constatações, hipóteses e considerações sobre os resultados obtidos.

Principais Constatações e Hipóteses

A análise dos resultados leva a crer que o impacto da saída antecipada do PBF sobre o acesso dos jovens ao mercado formal de trabalho possui relação estreita com a disponibilidade de empregos, sendo maior o impacto quanto maior for a oferta de trabalho. Duas constatações levam a esta conclusão: a redução do impacto com o agravamento da crise em 2015 e a existência de maior impacto em zonas urbanas e na região sudeste. O agravamento da crise em 2015, e a consequente redução da oferta de empregos, contribuiu para uma redução praticamente generalizada do impacto naquele período, conforme os resultados indicam. Adicionalmente, os maiores níveis de impacto serem percebidos em zonas urbanas e na região sudeste, áreas que, conforme apresentado na Seção 4.1.1 (ver Figuras 4.4 e 4.5), têm maior oferta de empregos formais, também contribuem para esta hipótese.

Duas outras constatações também chamam a atenção nesse estudo. A primeira diz respeito à concentração dos efeitos da crise econômica sobre jovens do sexo feminino, que pode ser atribuída a algum tipo de preconceito contra a contratação de força de trabalho feminina e jovem em situação de pouca oferta de emprego e grande número de pessoas à procura de trabalho. A segunda é relativa ao comportamento do impacto no sudeste, que não sofreu alterações diante da crise.

Considerações e Ressalvas

Inicialmente, cumpre destacar que os efeitos aqui estimados são locais, ou seja, são aplicáveis apenas àqueles jovens beneficiários de aproximadamente 18 anos. Dadas as características do método, os impactos não podem ser generalizados para todos os beneficiários.

Adicionalmente, apesar de ter sido possível identificar, quantitativamente, a existência de impacto nos dados gerais de 2013-2014 e em algumas subpopulações dos dados de 2013-2014 e de 2014-2015, apenas com esse estudo não é possível concluir se o PBF tem efeito positivo ou negativo amplo sobre os jovens. O alcance da autossuficiência da família por meio do trabalho é, de fato, uma das mais importantes portas de saída do programa. O acesso ao trabalho é tão importante que o programa possui proteções que visem evitar que a obtenção de renda por parte da família impeça, de imediato, a perda do benefício. Por exemplo, quando existe um acréscimo na renda per capita da família, até o limite de meio salário mínimo, os beneficiários contam com uma carência de 2 anos nos quais

a manutenção do benefício é assegurada. Depois desse período a família entra em uma fila para revisão que pode durar até um ano. Apenas se a renda per capita da família permanecer superior ao limite após esse período, que pode variar ente 2 e 3 anos, é que o benefício será cessado. Mas, apesar do trabalho ser uma importante porta de saída, existem outras portas de saída e outros aspectos que precisam ser avaliados para uma conclusão mais ampla. Dentre os estudos que ainda precisam ser realizados, destaca-se, principalmente, a avaliação de impactos do programa sobre a educação.

Os jovens beneficiários são estimulados a permanecer na escola, pelo menos até a conclusão ensino médio [9]. De acordo com o PNAD 2013 [86], no Brasil, apenas 54,3% dos jovens até 19 anos completaram o ensino médio, de forma que é possível que muitos dos jovens no ponto de corte desse estudo ainda estejam na escola. Embora não seja impossível trabalhar e estudar ao mesmo tempo, é razoável supor que jovens que ainda estudem tenham mais dificuldade de se estabelecer em um emprego que os jovens que já tenham deixado a escola. Assim, o fato da saída antecipada do programa estar levando mais jovens ao trabalho pode significar que os que ficam no programa estejam estudando mais. Sob esse prisma, o impacto aqui estimado estaria indicando uma característica positiva do programa. Entretanto, essa conclusão só é valida se, de fato, os jovens que ficarem no programa estiverem realmente na escola, e, preferencialmente, com bons resultados acadêmicos. Caso contrário, o impacto identificado por esse trabalho representaria um efeito negativo do programa sobre os jovens, pois o programa não estaria contribuindo para o alcance da porta de saída via trabalho formal, nem tampouco estaria contribuindo para a continuidade dos estudos dos jovens para aumentaria suas chances de obter melhores empregos no médio prazo.

Para a avaliação dos impactos do programa sobre a permanência na escola e sobre os resultados acadêmicos dos jovens, um estudo similar a este pode ser conduzido, desta feita usando dados do Censo da Educação Básica e do Exame Nacional do Ensino Médio (ENEM) no lugar da RAIS. Convém mencionar que a divulgação no âmbito do TCU de resultados intermediários desse trabalho motivou a SecexPrevi a iniciar ações no sentido de obter os dados necessários para esta avaliação complementar junto ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão responsável pelo Censo da Educação e pelo ENEM. As técnicas de integração e pareamento de dados aqui propostas bem como a ferramenta desenvolvida e a abordagem de aplicação do RDD podem ser totalmente aplicadas nesse trabalho complementar. Com a avaliação conjunta dos impactos sobre o trabalho e a educação seria possível realizar uma análise qualitativa ampla dos efeitos causados pelo programa sobre os jovens. Essa ação do Tribunal, motivada pelo presente trabalho, confirma, desde já, o atingimento dos objetivo de contribuir para a ampliação do volume de auditorias operacionais baseadas em dados e para

a melhoria das sistemática de avaliação e fiscalização dos programas sociais no Brasil.

Capítulo 5

Conclusões e Trabalhos Futuros

5.1 Conclusões

O presente projeto de pesquisa apresentou uma abordagem para avaliação de impacto de políticas públicas utilizando grandes bases governamentais que uniu técnicas de integração de dados e processamento paralelo da área da ciência da computação com a abordagem econométrica quasi-experimental de avaliação de impacto RDD.

A abordagem proposta foi aplicada na avaliação de impacto do PBF no que tange à sua contribuição para o acesso de jovens beneficiários ao mercado formal de trabalho e incluiu as seguintes etapas:

- integração das bases de dados governamentais Cadastro Único, Folha de Pagamentos do PBF e RAIS disponíveis no TCU;
- desenvolvimento de uma versão paralelizada de um *software open-source* de estimação RDD;
- estimação do impacto do PBF no acesso de jovens ao mercado formal de trabalho por meio da abordagem RDD e análise da variação dos impactos entre subpopulações de interesse.

O processo de integração de dados proposto se mostrou eficiente e flexível, com possibilidade de extração de diferentes *datasets* por meio de simples parametrização. O algoritmo paralelo desenvolvido apresentou desempenho quase 6 vezes superior à versão original no *hardware* e tamanho de amostra utilizados. A estimação RDD permitiu identificar, quantitativamente, a existência de impacto nos dados gerais de 2013-2014 e em algumas subpopulações dos dados de 2013-2014 e de 2014-2015. Testes de robustez foram realizados não tendo sido encontrados motivos para questionar a plausibilidade do modelo. Os resultados também demonstraram a importância da utilização de grandes amostras

para a obtenção de estimativas precisas, ao passo que processo de integração proposto e a versão paralela desenvolvida se mostraram aptos a lidar com grandes volumes de dados.

De acordo com os resultados obtidos, ficou claro que é possível realizar avaliações de impactos de políticas públicas usando bases de dados governamentais disponíveis nos órgão gestores. A abordagem aqui proposta apresentou vantagens frente a abordagem tradicional baseada em pesquisas, por não requerer investimentos adicionais e por apresentar resultados mais precisos, devido à maior quantidade de dados que pode ser utilizada.

Conforme apresentado e discutido no Capítulo 4, as análises indicaram a existência de relação entre a oferta de empregos e o tipo de impacto do PBF sobre o acesso de jovens ao mercado formal de trabalho. Em regiões com maior oferta de empregos, os jovens que deixam de receber o benefício mais cedo, tendem a ter mais sucesso no acesso ao mercado formal. Esse tipo de conclusão proporcionada por este trabalho pode ser usado pelo gestor da política pública como subsídio para a tomada de decisão acerca da realização de ajustes no programa. Por exemplo, visando maximizar seus resultados, o gestor pode optar por direcionar do mais recursos para regiões onde os efeitos do programa podem ser mais positivos, ou ajustar o volume de recursos destinado a determinadas faixas etárias.

Também conforme os resultados obtidos, ficou evidenciado que o impacto do programa pode ser influenciado pelo momento econômico vivido pelo país. A grande diferença observada entre os resultados de 2014 e 2015 apontam para a necessidade de que a avaliação da política pública seja realizada de forma sistemática e continuada. A abordagem aqui proposta, baseada em bases de dados existentes e que não depende de pesquisas adicionais, representa um facilitador para a sistematização da avaliação. O Tribunal de Contas da União (TCU), por intermédio de sua Secretaria de Controle Externo de Previdência e Assistência Social (SecexPrevi), no exercício da auditoria operacional da administração pública, já vislumbrou a oportunidade de incorporar a abordagem de avaliação de impacto do PBF aqui proposta em seu processo de fiscalização contínua, inclusive, com a ampliação do escopo para incluir a avaliação de impactos sobre a educação. Além disso, preocupado com seu objetivo de induzir melhorias nos processos internos de seus jurisdicionados, o TCU realizou uma apresentação dos resultados intermediários obtidos por esse trabalho para a Secretaria de Avaliação e Gestão da Informação (SAGI) do Ministério do Desenvolvimento Social e Agrário (MDSA)[1]. Na oportunidade a SAGI informou que já vinha estudando RDD e demonstrou interesse no trabalho a na possibilidade de absorver as técnicas de manipulação de grandes bases de dados aqui apresentadas.

É importante destacar que a abordagem aqui proposta não é restrita ao PBF. Outras políticas públicas que disponham de bases de dados de suporte e que possuam algum

[1]O MDSA é o ministério responsável pela gestão do PBF e a SAGI é a unidade responsável pelas ações de gestão da informação, monitoramento, avaliação e capacitação de agentes sociais. Mais informações sobre a SAGI podem ser obtidas em `https://aplicacoes.mds.gov.br/sagi/portal/`

critério de elegibilidade descontínuo são candidatas a serem objeto de avaliação por RDD pelo método apresentado. Por exemplo, a abordagem aqui proposta poderia ser aplicada ao Fundo de Financiamento Estudantil (FIES)[2], que possui um critério de elegibilidade descontínuo baseado na renda per capita da família do estudante.

Conforme apresentado, a contribuição do presente trabalho no que se refere ao processo de fiscalização de programas sociais no âmbito do TCU já é um fato concreto. A possibilidade de aplicação como instrumento de avaliação e de identificação de oportunidades de melhoria pelo gestor também é clara. Além disso, a abordagem proposta pode ser utilizada em outras políticas, ampliando ainda mais a sua possibilidade de aplicação. Dessa forma, entende-se que os objetivos estabelecidos para o presente trabalho foram cumpridos.

5.2 Trabalhos Futuros

Já se encontrava em andamento no TCU, à época da conclusão do presente trabalho, as tratativas para obter, junto ao INEP, os dados necessários para uma avaliação complementar dos impactos do PBF sobre a educação dos jovens beneficiários. Nesse trabalho complementar pretende-se utilizar as técnicas de integração e pareamento de dados aqui propostas, bem como a ferramenta RDD desenvolvida. Os resultados dessa avaliação de impactos sobre a educação associados ao resultados do presente trabalho permitirão realizar uma avaliação qualitativa ampla dos impactos do principal programa de transferência de renda do Brasil sobre os jovens de baixa renda. Além disso o processo flexível de integração de dados desenvolvido será utilizado para realização das análises relativas ao período 2015-2016, tão logo os dados da RAIS2016 estejam disponíveis.

Acerca do processamento RDD em paralelo, pretende-se avaliar a possibilidade de implementação de paralelismo por meio de *Graphic Processing Unit* (GPU) utilizando o pacote gpuR[3].

[2]O FIES é o programa do Ministério da Educação que financia cursos superiores não gratuitos e com avaliação positiva no Sistema Nacional de Avaliação da Educação Superior. Mais informações sobre o FIES podem ser obtidas em `http://fiesselecao.mec.gov.br/`

[3]O pacote gpuR está disponível em `https://cran.r-project.org/web/packages/gpuR/gpuR.pdf`

Referências

[1] Brasil: *L8443 DE 16 DE JULHO DE 1992 - Dispõe sobre a Lei Orgânica do Tribunal de Contas da União e dá outras providências*, 1992. http://www.planalto.gov.br/ccivil_03/Leis/L8443.htm, acesso em 2016-02-02. 1

[2] Bugarin, Bento José: *Avaliacao de programas públicos orientada para resultados : o papel dos órgaos de controle externo no Brasil*. Congreso Interamericano del CLAD sobre la Reforma del Estado y de la Administración Pública, (3):562, 1997. 2

[3] TCU: *Plano Estratégico do TCU 2015-2012*. 2015. http://goo.gl/ZIUQyZ. 2

[4] TCU: *Acordão 718/2016 - Plenario*, 2016. https://goo.gl/z7BXAs, acesso em 2016-02-02. 2

[5] TCU: *Acórdão 1009/2016 - Plenário*, 2016. https://goo.gl/0UOP2t, acesso em 2016-02-02. 2

[6] TCU: *Acórdão 1181/2016 - Plenário*, 2016. https://goo.gl/UtDcG4, acesso em 2016-02-02. 2

[7] Brasil: *Constituição da Rebública Federativa do Brasil de 1988*, 1988. http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm, acesso em 2016-02-02. 3

[8] Jannuzzi, Paulo: *Avaliação de programas sociais no Brasil: repensando práticas e metodologias das pesquisas avaliativas*. Planejamento e Políticas Públicas, (36):252–275, 2011. 3

[9] Brasil: *L10836 DE 9 DE JANEIRO DE 2004 - Cria o Programa Bolsa Familia e dá Outras Providências*, 2004. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l10.836.htm, acesso em 2016-02-02. 3, 4, 48, 72

[10] Brasil: *L10219 DE 11 DE ABRIL DE 2001 - Cria o Programa Nacional de Renda Mínima vinculada à educação - Bolsa Escola, e dá outras providências.*, 2001. http://www.planalto.gov.br/ccivil_03/leis/LEIS_2001/L10219.htm, acesso em 2016-02-02. 3

[11] Brasil: *L10689 DE 13 DE JUNHO DE 2003 - Cria o Programa Nacional de Acesso à Alimentação – PNAA.*, 2003. http://www.planalto.gov.br/ccivil_03/Leis/2003/L10.689.htm, acesso em 2016-02-02. 3

[12] Brasil: *MP2206-1 DE 6 DE SETEMBRO DE 2001 - Cria o Programa Nacional de Renda Mínima vinculado à saúde: Bolsa-Alimentação e dá outras providências.*, 2001. http://www.planalto.gov.br/ccivil_03/mpv/Antigas_2001/2206-1.htm, acesso em 2016-02-02. 3

[13] Brasil: *D6392 DE 12 DE MARÇO DE 2008 - Altera o Decreto no 5.209, de 17 de setembro de 2004, que regulamenta a Lei no 10.836, de 9 de janeiro de 2004, que cria o Programa Bolsa Família.*, 2008. http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/decreto/d6392.htm, acesso em 2016-02-02. 3

[14] MDSA: *Bolsa Familia - O que é*, 2017. https://mds.gov.br/assuntos/bolsa-familia/o-que-e, acesso em 2016-02-02. 3

[15] CGU: *Portal da Transparência - Bolsa Familia - Pagamentos*, 2017. http://www.portaltransparencia.gov.br/downloads/mensal.asp?c=BolsaFamiliaFolhaPagamento, acesso em 2016-02-02. 3

[16] Marinho, Alexandre e Luís Otávio Façanha: *Programas sociais: efetividade, eficiência e eficácia como dimensões operacionais da avaliação.* Textos para Discussão - IPEA, 2001, ISSN 1415-4765. http://repositorio.ipea.gov.br/handle/11058/2328, acesso em 2016-02-02. 3

[17] Jannuzzi, Paulo: *Indicadores para diagnóstico, monitoramento e avaliação de programas sociais no Brasil.* Revista do Serviço Público, 2(56), 2005. 4

[18] Mourão, Luciana e Jacob A. Laros: *Avaliação de programas sociais: comparando estratégias de análise de dados.* Psicologia: Teoria e Pesquisa, 24(4):545–558, 2008. https://www.researchgate.net/profile/Jacob_Laros/publication/233388205_Avaliao_de_Programas_Sociais_Comparando_Estratgias_de_Anlise_de_Dados/links/0fcfd50a0b203b543e000000.pdf, acesso em 2016-02-02. 4

[19] Rubin, Donald B.: *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.* Journal of Educational Psychology, 66(5):688 – 701, 1974. 4, 5, 11, 12

[20] Campbell, Donald Thomas e Julian Cecil Stanley: *Experimental and quasi-experimental designs for research.* Houghton Mifflin Comp, Boston, 2. print edição, 1967, ISBN 978-0-395-30787-8. OCLC: 247359300. 4, 11, 12

[21] Buddelmeyer, Hielke e Emmanuel Skoufias: *An evaluation of the performance of regression discontinuity design on PROGRESA*, volume 827. World Bank Publications, 2004. https://goo.gl/OA8Oyp, acesso em 2016-02-10. 4, 5, 12, 27

[22] Thistlethwaite, Donald L. e Donald T. Campbell: *Regression-Discontinuity Analysis - An alternative to ex post facto experiment.pdf.* The Journal of Education Psychology, 51(6):309–317, 1960. 4, 15

[23] Brasil: *D5209 DE 17 DE SETEMBRO DE 2004 - Regulamenta a Lei no 10.836, de 9 de janeiro de 2004, que cria o Programa Bolsa Família, e dá outras providências.*, 2004. http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/decreto/d5209.htm, acesso em 2016-02-02. 4, 48, 49

[24] Sousa, Darcon e Alisson Felipe de Melo Brito: *Os mecanismos de portas de saida do programa Bolsa Familia e as perspectivas dos beneficiários no municipio de Caturié, Paraíba.* 2015. http://www.joinpp.ufma.br/jornadas/joinpp2015/pdfs/eixo4/os-mecanismos-de-portas-de-saida-do-programa-bolsa-familia-e-as-persp pdf, acesso em 2017-02-11. 4

[25] Brauw, Alan de, Daniel O. Gilligan, John Hoddinott e Shalini Roy: *The Impact of Bolsa Família on Schooling.* World Development, 70:303–316, junho 2015, ISSN 0305750X. http://linkinghub.elsevier.com/retrieve/pii/S0305750X1500025X, acesso em 2016-02-02. 5, 14

[26] Reynolds, Sarah Anne: *Brazil's Bolsa Familia: Does it work for adolescents and do they work less for it?* Economics of Education Review, 46:23–38, junho 2015, ISSN 02727757. http://linkinghub.elsevier.com/retrieve/pii/S0272775715000229, acesso em 2016-02-02. 5, 14

[27] Barbosa, Ana Luiza Neves de Holanda e Carlos Henrique Leite Corseuil: *Conditional cash transfer and informality in Brazil.* IZA Journal of Labor & Development, 3(1), dezembro 2014, ISSN 2193-9020. http://www.izajold.com/content/3/1/37, acesso em 2016-02-10. 5, 27, 28

[28] Barrientos, Armando e Juan Miguel Villa: *Antipoverty transfers and labour force participation effects.* Brooks World Poverty Institute, Manchester, 2013, ISBN 978-1-909336-03-2. http://www.bwpi.manchester.ac.uk/resources/Working-Papers/wp_18513.html, acesso em 2016-06-10, OCLC: 931311200. 5, 27

[29] Filmer, Deon e Norbert Schady: *Does more cash in conditional cash transfer programs always lead to larger impacts on school attendance?* Journal of Development Economics, 96(1):150–157, setembro 2011, ISSN 03043878. http://linkinghub.elsevier.com/retrieve/pii/S0304387810000507, acesso em 2016-06-10. 5, 27

[30] Klaauw, Wilbert Van der: *Estimating the effect of financial aid offers on college enrollment: A Regression–Discontinuity Approach*.* International Economic Review, 43(4):1249–1287, 2002. http://onlinelibrary.wiley.com/doi/10.1111/1468-2354.t01-1-00055/full, acesso em 2016-02-16. 5

[31] Hahn, Jinyong, Todd e Wilbert Van der Klaauw: *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design.* Econometrica, 69(1), 2001. 5, 15, 18, 25

[32] Card, David, Nicole Maestas e Carlos Dobkin: *The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare*. The American Economic Review, 98(5):2242–2258, 2008, ISSN 00028282. http://www.jstor.org/stable/29730170. 5

[33] Varian, Hal R.: *Big Data: New Tricks for Econometrics* †. Journal of Economic Perspectives, 28(2):3–28, maio 2014, ISSN 0895-3309. http://pubs.aeaweb.org/doi/abs/10.1257/jep.28.2.3, acesso em 2016-05-18. 5, 6, 8

[34] Einav, Liran e Jonathan Levin: *The data revolution and economic analysis*. Innovation Policy and the Economy, 14(1):1–24, 2014. http://www.journals.uchicago.edu/doi/abs/10.1086/674019, acesso em 2017-02-11. 5

[35] Foundation, R: *The R Project for Statistical Computing*, 2017. https://www.r-project.org/, acesso em 2016-02-02. 6, 25

[36] Thoemmes, F., W. Liao e Z. Jin: *The Analysis of the Regression-Discontinuity Design in R*. Journal of Educational and Behavioral Statistics, novembro 2016, ISSN 1076-9986, 1935-1054. http://jeb.sagepub.com/cgi/doi/10.3102/1076998616680587, acesso em 2017-02-05. 6, 9, 25, 43

[37] Dimmery, Drew: *Package 'rdd'*. 2013. https://cran.r-project.org/web/packages/rdd/rdd.pdf, acesso em 2016-06-11. 6, 25, 43, 49

[38] Calonico, Sebastian, Matias D. Cattaneo e Rocio Titiunik: *rdrobust : An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs*. R Journal, 7(1):38–51, 2015. http://goo.gl/B2pVWX, acesso em 2016-05-18. 6, 20, 25, 43, 44

[39] Stigler, Matthieu e Bastiaan Quast: *Package 'rddtools'*. 2016. http://cran.cnr.berkeley.edu/web/packages/rddtools/rddtools.pdf, acesso em 2017-02-05. 6, 25, 43, 52, 54

[40] Aruoba, S. Borağan e Jesús Fernández-Villaverde: *A comparison of programming languages in economics*. Relatório Técnico, National Bureau of Economic Research, 2014. http://www.nber.org/papers/w20263.pdf, acesso em 2017-02-05. 6, 9, 25

[41] MathWorks: *MATLAB - The Language of Technical Computing*, 2017. https://www.mathworks.com/products/matlab.html, acesso em 2016-02-02. 6

[42] Wolfram: *Mathematica - The world's definitive system for modern technical computing*, 2017. https://www.wolfram.com/mathematica/, acesso em 2016-02-02. 6

[43] Grimmer, Justin: *We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together*. PS: Political Science & Politics, 48(01):80–83, janeiro 2015, ISSN 1049-0965, 1537-5935. http://www.journals.cambridge.org/abstract_S1049096514001784, acesso em 2016-05-18. 6, 8

[44] Alles, Michael G., Edson Luiz Riccio, Miklos A. Vasarhelyi e Fernando Tostes: *Continuous Auditing: the USA Experience and Considerations for its Implementation in Brazil*. Journal of Information Systems and Technology Management, páginas 211–224, agosto 2006, ISSN 18071775. http://www.jistem.fea.usp.br/index.php/jistem/article/view/10.4301%252FS1807-17752006000200007, acesso em 2016-07-20. 8

[45] Figueiredo, Marcus Faria e Argelina Maria Cheibub Figueiredo: *Avaliação política e avaliação de políticas: um quadro de referência teórica*. Número 15. Instituto de Estudos Econômicos, Sociais e Políticos de São Paulo São Paulo, 1986. http://www.josenorberto.com.br/josenorberto/AC-2007-38.pdf, acesso em 2017-02-12. 10

[46] Arretche, Marta T. S.: *Tendências no estudo sobre avaliação*. Em *Avaliação de Políticas Sociais: Uma Questão em Debate*. Cortez, São Paulo, 2000. http://www.teses.usp.br/teses/disponiveis/48/48134/tde-17122014-113247/en.php, acesso em 2017-02-12. 10

[47] Imbens, Guido W. e Thomas Lemieux: *Regression discontinuity designs: A guide to practice*. Journal of econometrics, 142(2):615–635, 2008. http://www.sciencedirect.com/science/article/pii/S0304407607001091, acesso em 2016-02-10. 10, 15, 16, 17, 18, 20, 21, 29, 48, 51, 52, 53, 57

[48] Holland, Paul W.: *Statistics and Causal Inference*. Journal of the American Statistical Association, 81(396):945, dezembro 1986, ISSN 01621459. http://www.jstor.org/stable/2289064?origin=crossref, acesso em 2017-02-12. 11

[49] Diaz, Juan Jose e Sudhanshu Handa: *An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator Evidence from Mexicos PROGRESA Program.pdf*, 2005. http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=862049, acesso em 2016-10-02. 12, 14

[50] Abadie, Alberto: *Semiparametric difference-in-differences estimators*. The Review of Economic Studies, 72(1):1–19, 2005. http://restud.oxfordjournals.org/content/72/1/1.short, acesso em 2016-07-04. 12, 13

[51] Card, David e Alan B Krueger: *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*. The American Economic Review, 84(4):772–793, 1994. 12

[52] Gruber, Jonathan: *The Incidence of Mandated Maternity Benefits*. The American Economic Review, 84(3):622–641, 1994. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.333.8024, acesso em 2016-07-04. 13

[53] Rosenbaum, Paul R. e Donald B. Rubin: *The Central Role of the Propensity Score in Observational Studies for Causal Effects*. Biometrika, 70(1):41, abril 1983, ISSN 00063444. http://www.jstor.org/stable/2335942?origin=crossref, acesso em 2016-07-04. 14

[54] Rosenbaum, Paul R. e Donald B. Rubin: *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score*. Journal of the American Statistical Association, 79(387):516, setembro 1984, ISSN 01621459. http://www.jstor.org/stable/2288398?origin=crossref, acesso em 2016-07-04. 14

[55] Hirano, Keisuke, Guido W. Imbens e Geert Ridder: *Efficient estimation of average treatment effects using the estimated propensity score*. Econometrica, 71(4):1161–1189, 2003. http://onlinelibrary.wiley.com/doi/10.1111/1468-0262.00442/abstract, acesso em 2016-07-04. 14, 15

[56] Abadie, Alberto e Guido Imbens: *Simple and bias-corrected matching estimators for average treatment effects*. National Bureau of Economic Research Cambridge, Mass., USA, 2002. http://www.nber.org/papers/t0283, acesso em 2016-07-04. 14

[57] Hirano, Keisuke e Guido W. Imbens: *Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization*. Health Services and Outcomes research methodology, 2(3-4):259–278, 2001. http://link.springer.com/article/10.1023/A:1020371312283, acesso em 2016-07-04. 14

[58] Heckman, J. J., H. Ichimura e P. E. Todd: *Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme*. The Review of Economic Studies, 64(4):605–654, outubro 1997, ISSN 0034-6527, 1467-937X. https://academic.oup.com/restud/article-lookup/doi/10.2307/2971733, acesso em 2017-02-12. 14

[59] Caliendo, Marco e Sabine Kopeinig: *Some practical guidance for the implementation of propensity score matching*. Journal of economic surveys, 22(1):31–72, 2008. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6419.2007.00527.x/full, acesso em 2016-07-05. 14

[60] Angrist, Joshua D. e Victor Lavy: *Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achieviment*. The Quarterly Journal of Economics, (144):533–575, 1999. 15

[61] Black, Sandra E.: *Do better schools matter? Parental valuation of elementary education*. The Quarterly Journal of Economics, 114(2):577–599, 1999. http://qje.oxfordjournals.org/content/114/2/577.short, acesso em 2017-02-17. 15

[62] Jacob, Robin Tepper, Pei Zhu, Marie Andrée Somers e Howard S. Bloom: *A practical guide to regression discontinuity*. Citeseer, 2012. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.671.4723&rep=rep1&type=pdf, acesso em 2016-02-16. 16, 18, 21, 26

[63] Gelman, Andrew e Guido Imbens: *Why high-order polynomials should not be used in regression discontinuity designs*. Relatório Técnico, National Bureau of Economic Research, 2014. http://www.nber.org/papers/w20405.pdf, acesso em 2016-05-18. 18

[64] Imbens, Guido e Karthik Kalyanaraman: *Optimal bandwidth choice for the regression discontinuity estimator*. The Review of Economic Studies, página rdr043, 2011. http://goo.gl/yM3veq, acesso em 2016-05-18. 18, 19, 20

[65] Card, David, David Lee, Zhuan Pei e Andrea Weber: *Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design*. Relatório Técnico, National Bureau of Economic Research, 2012. http://www.nber.org/papers/w18564, acesso em 2016-05-28. 19, 20

[66] Calonico, Sebastian, Matias D. Cattaneo e Rocio Titiunik: *Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs: Robust Nonparametric Confidence Intervals*. Econometrica, 82(6):2295–2326, novembro 2014, ISSN 00129682. http://doi.wiley.com/10.3982/ECTA11757, acesso em 2016-05-18. 20

[67] Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell e Rocıo Titiunik: *Regression Discontinuity Designs Using Covariates*. Relatório Técnico, working paper, University of Michigan, 2016. http://faculty.chicagobooth.edu/max.farrell/research/RD-covbc.pdf, acesso em 2016-05-18. 20

[68] McCrary, Justin: *Manipulation of the running variable in the regression discontinuity design: A density test*. Journal of Econometrics, 142(2):698–714, fevereiro 2008, ISSN 03044076. http://linkinghub.elsevier.com/retrieve/pii/S0304407607001133, acesso em 2016-05-18. 21, 49

[69] Paiva, Eduardo Soares, Kate Cerqueira Revoredo e Fernanda Araujo Baião: *DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro*. iSys-Revista Brasileira de Sistemas de Informação, 9(1):6–32, 2016. http://www.seer.unirio.br/index.php/isys/article/view/5350, acesso em 2017-04-09. 22

[70] Bernstein, Philip A. e Laura M. Haas: *Information integration in the enterprise*. Communications of the ACM, 51(9):72–79, 2008. http://dl.acm.org/citation.cfm?id=1378745, acesso em 2017-02-19. 22, 23, 33

[71] Chaudhuri, Surajit e Umeshwar Dayal: *An overview of data warehousing and OLAP technology*. ACM Sigmod record, 26(1):65–74, 1997. http://dl.acm.org/citation.cfm?id=248616, acesso em 2017-04-09. 23

[72] Kabiri, Ahmed e Dalila ChiadmiI: *Survey on ETL Processes*. Journal of Theoretical & Applied Information Technology, 54(2), 2013. goo.gl/8LEzZi, acesso em 2017-02-19. 23

[73] Wiederhold, Gio: *Mediators in the architecture of future information systems*. Computer, 25(3):38–49, 1992. http://ieeexplore.ieee.org/abstract/document/121508/, acesso em 2017-04-09. 23

[74] Alonso, G., F. Casati, H. Kuno e V. Machiraju: *Web Services—Concepts, Architectures and Applications*. Springer, 2004. 23

[75] Winkler, William E.: *Matching and record linkage*. Wiley Interdisciplinary Reviews: Computational Statistics, 6(5):313–325, setembro 2014, ISSN 19395108. http://doi.wiley.com/10.1002/wics.1317, acesso em 2017-02-20. 24, 25, 37, 38, 39

[76] Sayers, Adrian, Yoav Ben-Shlomo, Ashley W Blom e Fiona Steele: *Probabilistic record linkage*. International Journal of Epidemiology, 45(3):954–964, junho 2016, ISSN 0300-5771, 1464-3685. https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyv322, acesso em 2017-04-23. 24, 38

[77] Fellegi, Ivan P. e Alan B. Sunter: *A Theory for Record Linkage*. Journal of the American Statistical Association, 64(328):1183, dezembro 1969, ISSN 01621459. http://www.jstor.org/stable/2286061?origin=crossref, acesso em 2017-04-09. 24

[78] Cerotti, D., M. Gribaudo, M. Iacono e P. Piazzolla: *Modeling and analysis of performances for concurrent multithread applications on multicore and graphics processing unit systems: ANALYSIS OF PERFORMANCES ON MULTITHREAD APPLICATIONS*. Concurrency and Computation: Practice and Experience, 28(2):438–452, fevereiro 2016, ISSN 15320626. http://doi.wiley.com/10.1002/cpe.3504, acesso em 2017-02-20. 25

[79] Weston, Steve: *Getting started with doMC and foreach*. https://cran.r-project.org/web/packages/doMC/vignettes/gettingstartedMC.pdf, acesso em 2017-06-24. 25, 45

[80] Hasanov, Khalid, Jean Noël Quintin e Alexey Lastovetsky: *Hierarchical approach to optimization of parallel matrix multiplication on large-scale platforms*. The Journal of Supercomputing, 71(11):3991–4014, novembro 2015, ISSN 0920-8542, 1573-0484. http://link.springer.com/10.1007/s11227-014-1133-x, acesso em 2017-02-20. 25

[81] Dourado, Aloisio, Rommel N. Carvalho, Gustavo van Erven e Donald Pianto: *Brazil's bolsa familia and young adult workers: A parallel RDD approach to large datasets*. Em *Neural Networks (IJCNN), 2017 30th International Joint Conference on*, páginas 17–24. INNS - IEEE. 26, 43, 44, 45, 47

[82] Scarlato, Margherita, Giorgio d'Agostino e Francesca Capparucci: *Evaluating CCTs from a Gender Perspective: The Impact of Chile Solidario on Women's Employment Prospect: Evaluating CCTs from a Gender Perspective*. Journal of International Development, 28(2):177–197, março 2016, ISSN 09541748. http://doi.wiley.com/10.1002/jid.3124, acesso em 2016-06-10. 27

[83] Brasil: *D76900, DE 23 DE DEZEMBRO DE 1975 - Institui a Relação Anual de Informações Sociais – RAIS e dá outras providências.*, 1975. http://www.planalto.gov.br/ccivil_03/decreto/antigos/d76900.htm, acesso em 2016-02-02. 28

[84] Kleiber, Christian e Achim Zeileis: *Applied Econometrics with R*. Springer-Verlag, New York, 2008. https://CRAN.R-project.org/package=AER, ISBN 978-0-387-77316-2. 46

[85] Ruppert, David, Simon J. Sheather e Matthew P. Wand: *An effective bandwidth selector for local least squares regression.* 90(432):1257–1270. http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476630, acesso em 2017-05-27. 52

[86] IBGE: *Pesquisa nacional por amostra de domicílios - 2013.* http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2013/, acesso em 2017-09-06. 72