Julio Alfredo Racchumi Romero

UTILIZANDO O RELACIONAMENTO DE BASES DE DADOS PARA AVALIAÇÃO DE POLÍTICAS PÚBLICAS: UMA APLICAÇÃO PARA O PROGRAMA BOLSA FAMÍLIA

Belo Horizonte, MG
UFMG/Cedeplar
2008

Julio Alfredo Racchumi Romero

UTILIZANDO O RELACIONAMENTO DE BASES DE DADOS PARA AVALIAÇÃO DE POLÍTICAS PÚBLICAS: UMA APLICAÇÃO PARA O PROGRAMA BOLSA FAMÍLIA

Tese apresentada ao curso de doutorado em Demografia do Centro de Desenvolvimento e Planejamento Regional da Faculdade de Ciências Econômicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de doutor em Demografia.

Orientador: Profª . Drª. Ana Maria Hermeto Camilo de Oliveira
Co-orientador: Profª. Drª. Diana Reiko Tutiya Oya Sawyer

Belo Horizonte, MG
Centro de Desenvolvimento e Planejamento Regional
Faculdade de Ciências Econômicas - UFMG
2008

Folha de Aprovação

A meus pais Maria e Julio e minhas irmãs Betty e Norma.

AGRADECIMENTOS

Para que o trabalho fosse concluído foi imprescindível contar com apoio de várias pessoas. Em particular registro aqui meu agradecimento:

A professora Ana Maria Hermeto Camilo de Oliveira, pela orientação, paciência e atenção dispensada, que me permitiu trilhar o caminho da conclusão desta tese. Também agradeço à professora Diana Reiko Tutiya Oya Sawyer com quem iniciei o caminho deste trabalho.

Aos professores e funcionários do programa da pós-graduação do CEDEPLAR, pela ajuda dispensada durante o curso. Especialmente gostaria de agradecer à professora Laura Lídia Rodríguez Wong pela disponibilidade nos momentos que precisei de conselhos. Aos professores Rômulo Paes de Sousa, Adriana Miranda Ribeiro, Carla Jorge Machado e Eduardo Luiz Gonçalves Rios Neto, pelas colocações oportunas durante a defesa e que contribuíram para enriquecer este trabalho.

Aos companheiros da turma 2004 (Cíntia, Clarissa, Denise, Elisangela, Geovane, Izabel, Laetícia, Gilberto, Edwan, Juliana, Mário, Nelson e Rofília). Especialmente gostaria de agradecer, também, a Marisol, Elisenda e Cláudia. Companheiros que compartilharam desde minha chegada ao CEDEPLAR, momentos bons para minha permanência.

Aos meus amigos Mirela, Marcos e Luiza pessoas que me ajudaram desinteressadamente desde as primeiras semanas em Belo Horizonte, para que a minha passagem por aqui fosse mais fácil e confortável.

Aos meus amigos Almada, Regiane e Gláucia, pessoas que não só contribuíram para melhorar a redação do texto da tese, mas também, pela amizade sincera que me ofereceram. Da mesma forma gostaria agradecer ao Leonardo que também contribuiu no processamento das informações utilizadas na tese.

A todos os demais amigos que fiz durante o período de estudo, cujos nomes omitirei para não cometer a injustiça de esquecer algum.

A todos os meus familiares, em especial as meus pais Maria e Julio, minhas irmãs Betty e Norma, minha tia Julia e minha prima Cinthya, porque apesar da distância sempre estiveram ao meu lado.

Ao CNPq, pelo apoio financeiro.

E finalmente gostaria de agradecer ao bom Deus, por ter colocado estas pessoas em meu caminho.

LISTA DE ABREVIATURAS E SIGLAS

AFDC: Homemaker-House Health Aide Demonstration.

AIBF: Avaliação de Impacto do Bolsa Família.

AIH: Assessoria e Informatização Hospitalar.

AOD: Serviço de álcool / drogas.

AVE: Acidente Vascular Encefálico.

BPC: Beneficio de Prestação Continuada.

CadÚnico: Cadastro Único.

CAGED: Cadastro General de Emprego y Desemprego.

CBDB: Base de Dados de Nascimentos Canadense

CEDEPLAR – Centro de Desenvolvimento e Planejamento Regional

CEPAL: Comissão Econômica para América Latina e o Caribe.

CHI :Índice de Saúde de Comunidade.

CMHS: Centro para Saúde Mental Conserta.

CPF: Cadastro de Pessoas Física.

CSAT: Tratamento de Abuso de Substância Proibido.

DSE: Sistema de Estimação Dual.

FIP: Fondo de Inversión para la Paz.

FRD: Regressão Descontínua Fuzzy.

GISES/CT: Gerência de Filial de Serviços Sociais

GSF: Gastos Sociais Federais.

IBGE: Instituto Brasileiro de Geografia e Estatística.

IDB:Base de Dados Integrada

INSS: Instituto Nacional do Seguro Social.

JTPA: The National Job Training Partnership Act Study.

LEHD: Projeto Longitudinal da Dinâmica Empregador-Empregado.

MAS: Estudo de Relacionamento Automatizado

MDS: Ministério de Desenvolvimento Social.

MH: Serviço de Saúde Mental.

NDI: Índice de Morte Nacional.

NHS: Serviço Nacional de Saúde.

NHSCR: Registro Central de Serviços de Saúde Nacional.

NNM: Nearest Neighbor Matching.

NSW: National Supported Work Demostration.

NYSIIS: Sistema de Informação de Inteligência Estatal de Nova Iorque.

NZCMS: Registros Civis de Mortalidade de Nova Zelândia.

ONC: One Number Censu

ONGs: Organismos não Governamentais.

PAMS: Pesquisa de Assistência Médico-Sanitária

PBF: Programa Bolsa Família.

PES: Pesquisa de pós-Enumeração.

PES-NZ: Pesquisa pós-Enumeração de Nova Zelândia

PETI: Programa de Erradicação do Trabalho Infantil.

PIA: Pesquisa Industrial Anual.

PÍB: Produto Interno bruto.

PME: Pesquisa Mensal de Emprego.

PNAD: Pesquisa Nacional por Amostra de Domicílios.

PNSB: Pesquisa Nacional de Saneamento Básico.

PPM: Pesquisa da Pecuária Municipal

Project STAR: Outside labor programs Tennessee's Student Teacher Achievement Ratio

PSM: Pareamento por Escore de Propensão.

PSU: Unidade Primaria de Amostragem.

RAIS: Relação Anual de Informação Social.

RCT: Revenue Canadá.

RD: Regressão Descontínua.

RDS: Regressão descontínua Sharp.

RPICC: Registros do Centro de Cuidados Intensivos Regional das Crianças Pré-Natais.

RPS: Red de Protección Social.

RRC: Reverse Record Check

SAMHSA: Centro de Serviços Administrativos de Abusos de Sustâncias e Saúde Mental.

SETP: Secretaria Estadual de Trabalho, Emprego e Promoção Social

SIH: Sistema de Informação Hospitalaria.

SIM: Sistema de Informações sobre Mortalidade.

SINASC: Sistema de Informações sobre Nascidos Vivos.

SISBEN: Sistema de Identificación de Potenciales Beneficiários de Programas Sociales

SM: Stratification Matching.

TCR: Transferências Condicionadas de Renda.

TE: Titulo de Eleitor.

UFMG: Universidade Federal de Minas Gerais

SUMÁRIO

LISTA DE ILUSTRAÇÕES

RESUMO

Os programas sociais constituem, desde a última década, uma das respostas mais freqüentes aos problemas de desigualdade social. No Brasil, o Programa Bolsa Família (PBF) tem adquirido ampla relevância nacional porque objetiva reduzir a pobreza e desigualdade de hoje e de amanhã. A eficácia e a qualidade do PBF só podem ser medidas por meio de mecanismos de avaliação. Para garantir uma apropriada avaliação de impacto do PBF é crucial dispor de informação confiável e oportuna que identifique visivelmente os grupos de tratamento e comparação, com viés de seleção amostral, o menos possível que sejam semelhantes em todos os aspectos, diferenciando-se unicamente pela participação no programa. Considerando as características dos beneficiários do PBF, a pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF), realizada em 2005, não conseguiu efetuar uma avaliação experimental do programa, optando-se pela elaboração de uma pesquisa de linha de base domiciliar, executando-se previamente uma operação de *screening* ou varredura, para categorizar os domicílios segundo benefício recebido. Embora a informação obtida na varredura seja considerada adequada para análise na AIBF, é possível que as respostas estejam influenciadas por aspetos subjetivos. No entanto, é importante ressaltar que utilizar registros administrativos do CadÚnico possibilita conferir e avaliar as classificações dos domicílios alvo, porque são informações utilizadas pelos encarregados do monitoramento do PBF. Diante da importância da avaliação e da metodologia que abrange o processo de uma avaliação para estabelecer os limites da análise e da descrição dos resultados, esta tese explora as possibilidades únicas que são abertas pelo relacionamento de bases de dados para analisar a sensibilidade dos resultados de impacto dos programas sociais de transferência de renda, quando se utiliza dois tipos de fontes de informação para a alocação das famílias nos grupos de comparações. Para realizar a comparação dos resultados foram utilizadas duas fontes de informação: as bases de dados obtidas da pesquisa de campo AIBF e a dos registros administrativos do CadÚnico. Segundo as características destas bases de dados, duas estratégias de relacionamento foram utilizadas: a determinística e a probabilística. Como resultados destes relacionamentos foi possível mensurar os efeitos do impacto sobre a educação do Programa Bolsa Família (PBF) para a população entre 7 e 14 anos, quando as famílias são alocadas nos grupos de comparação, segundo a pesquisa de campo AIBF e segundo os registros administrativos do CadÚnico. Para encontrar os resultados de avaliação de impacto foi utilizado o método de pareamento por escore de propensão não-experimental. Além disso, recorrendo a uma forma particular de identificar os grupos potencialmente beneficiários e não beneficiários do PBF, utilizou-se o método da regressão descontínua, exercício que não seria viável usando apenas uma única fonte de informação. Os resultados do trabalho sugerem que, com o relacionamento de base de dados, o número de famílias relacionadas foi considerado satisfatório para analisar as variações ou sensibilidades dos resultados de impacto com as duas fontes de informação. Por sua parte, os resultados da análise comparativa evidenciam diferenciais que não são relevantes se considerado a alocação das famílias pela pesquisa de campo AIBF, mas que se tornam significativos com a alocação decorrente do CadÚnico. Além disso, há indícios de que estes resultados dependem em maior medida das características educativas próprias das crianças pertencentes às famílias que foram alocadas em outros grupos.

Palavras-chave: relacionamento; avaliação de impacto; escore de propensão; regressão descontínua.

ABSTRACT

The social programs form, since the last decade, one of the most frequent responses to the problems of social inequality. In Brazil, the "Bolsa Familia" Program (BFP) has acquired a broad national relevance because at aims at reducing the poverty and inequality of today and of tomorrow. The efficacy and the quality of the BFP can only be measured via evaluation mechanisms. To assure an appropriate evaluation of the impact of the BFP it is most important to have available a reliable and opportune information that identifies in a visible manner the treatment and comparison groups, with a small fold of simple selection, which are similar in all aspects, being solely different as to the participation in the program. Taking into consideration the characteristics of those benefited by the BFP, the research of the Impact Evaluation of Bolsa Família (IEBF) – carried out in 2005 – was not able to effect an experimental evaluation of the program, making an option for the elaboration of a domicile base line research, previously executing a screening operation, to categorize the domiciles as per the benefit received. Although the information obtained in the screening is considered adequate for analysis in the IEBF, it is possible that the responses be influenced by the subjective aspects. However, it is important to emphasize that using the administrative records of the CadÚnico enables to check and evaluate the targeted classifications, because these are information used by the supervisors of the monitoring of the BFP. In view of the importance of the evaluation and of the methodology that cover the process of an evaluation to establish the limits of the analysis and of the description of the results, this dissertation has as objective to explore the sole possibilities that are generated with the record linkage the bases to analyze the sensibility of the results of impact of the social programs for transfer of income, when are analyzed two types of sources of information for allocation of the families in the groups of treatment and comparison. To effect the comparison of the results, two sources of information were used: the data bases obtained from the field research of the IEBF and the one of the administrative records of the CadÚnico. In accordance with the characteristics of these data bases, two record linkage strategies were used: the deterministic and the probabilistic. As result of this record linkage it was possible to measure the effects of the impact upon the education in the BFP for the population between age 7 and age 14, when the families are allocated in the comparison groups, as per the field research of the EIBF and as per the administrative records of the CadÚnico. To find the results of the evaluation of impact it was used the non-experimental method propensity score matching (PSM); besides, using a special form to identify the potentially beneficiary and non-beneficiary groups of the BFP, it was used the Regression-Discontinuity method, an exercise that would not be feasible using one only source of information. The results of the work suggest that – with the record linkage of the bases – the number of listed families was considered satisfactory to analyze the variations or sensibilities of the results of impact with the two sources of information. On their turn, the results of the comparative analysis evidence differentials that are not relevant if it is considered the allocation of the families by the field research of the IEBF, but that become significant with the allocation derived from the CadÚnico. Besides, there are indications that these results depend largely on the educative characteristics proper of the children belonging to the families that were allocated in other groups.

Keywords: record linkage; impact evaluation; propensity score; regression-discontinuity.

1 INTRODUÇÃO

Reduzir os níveis de pobreza e desigualdade social, que prevalecem na maioria dos países da América Latina e Caribe, constitui uma das metas prioritárias dos governos dos países da região. Avaliar os diferentes aspectos da gestão pública em termos de política social tem adquirido maior importância nos últimos anos porque, através destes, conhece-se a eficiência e resultados dos esforços que se vêm realizando para melhorar as condições de vida da população (CEPAL, 2004).

Os programas sociais destinados à proporção da população considerada como pobre (ou menos favorecidos) constituem, desde a última década, uma das respostas mais freqüentes dos governos federais aos problemas de desigualdade social, porque pressupõe-se que por meio de subsídios alimentícios, transferência de renda, investimento em infra-estrutura e emprego por conta própria seria possível reduzir a vulnerabilidade das famílias frente a eventos negativos como a desigualdade social, recessão econômica e desastres naturais (BANCO MUNDIAL, 2003b).

No Brasil, dentre os principais programas sociais coordenados e fiscalizados pelo Governo Federal que visam aliviar ou combater a pobreza, destacamos o Programa Bolsa Família (PBF), que apresenta relevância nacional e será objeto neste trabalho. O PBF objetiva, primeiro, reduzir a pobreza e a desigualdade de hoje, fornecendo transferências em dinheiro para famílias pobres; e, segundo, reduzir a pobreza e a desigualdade de amanhã, provendo incentivos ao investimento em capital humano das famílias beneficiárias, tornando possível que essas famílias possam sair da pobreza. O PBF condiciona que as famílias mantenham as crianças e adolescentes em idade escolar freqüentando a escola e que cumpram os cuidados básicos na saúde (BRASIL, 200-?c).

A eficácia e a qualidade dos programas sociais como o PBF só pode ser medida por meio de mecanismos de avaliação. Para aplicar os métodos de avaliação tornam-se cada vez mais necessário dispor de informação confiável e oportuna que permita caracterizar, avaliar e conhecer as áreas e os grupos populacionais destinatários dos benefícios dos programas sociais. A avaliação de um programa é importante porque unicamente através

deste se poderá conhecer se o programa teve resultados positivos ou não, e se este deve continuar ou se modificar (RAVALLION, 2001; BUDDELMEYER e SKOUFIAS, 2004).

Além disso, o trabalho da avaliação pode ser considerado flexível para combinar diferentes métodos, instrumentos e fontes de informações que estão relacionados com as características e o contexto da intervenção (NAVARRO, 2005).

O processo de avaliação de impacto, em seu rigor metodológico, estima o cenário contrafactual ou simulado alternativo. Para determinar o cenário contrafactual, precisa-se separar o efeito das intervenções de outros fatores, uma tarefa um tanto complexa. Isto é conseguido com a ajuda de grupos de controle[1] (aqueles que não participam em um programa nem recebem benefícios), que se comparam com o grupo de tratamento (pessoas que recebem a intervenção). Além disso, considera que o grupo de controle (ou comparação) deve ser semelhante ao grupo de tratamento em todos aspectos e a única diferença entre os grupos é a participação no programa. A determinação do cenário contrafactual, que é essencial para o desenho da avaliação, e, por conseguinte, para a determinação dos grupos de comparação (tratamento e controle) está ligada estreitamente à implementação do método de avaliação e da coleta de dados. A configuração da coleta dos dados é uma das atividades importantes na avaliação, devido à alta incidência da qualidade nos resultados (NAVARRO, 2005).

Nas avaliações dos programas como o PBF, a coleta dos dados é particularmente complexa devido às características dos beneficiários e geralmente pela existência de restrições de tempo e orçamentos. Entre as fontes de informação mais utilizadas para obter os dados necessários para a avaliação, ressaltam-se as pesquisas domiciliares, que coletam informação sobre as características demográficas e socioeconômicas das famílias, e em alguns casos sobre a participação das pessoas nos programas. No entanto, é importante utilizar registros administrativos ou fontes de informação secundarias úteis para conferir e avaliar as classificações dos domicílios alvos, segundo critérios de elegibilidades entre os beneficiários e não beneficiários de determinado programa social (COADY et al., 2004).

---

[1] Os grupos de domicílios foram reclassificados em termos de elegibilidade: tratamento e controle; denominando-se daqui para frente, grupo de comparação 1 e comparação 2 para descrever os dois grupos de controles definidos na pesquisa AIBF e que serão utilizados nesta tese.

São eventos importantes na implementação da avaliação a medição adequada do impacto dos programas sociais e a focalização destes, o que induziu alguns países da América Latina e Caribe a desenvolverem seus próprios critérios e índices, principalmente com o objetivo de focalizar apropriadamente as intervenções sociais classificando adequadamente as famílias beneficiárias dos programas sociais. Dois casos específicos da América Latina são mencionados, na Colômbia, o índice SISBEN[2] e, no México, um índice de elegibilidade multidimensional. Tais índices são importantes na avaliação de impacto, uma vez que por meio destes pode-se ordenar as famílias ou domicílios alvos, permitindo a pré-classificação de beneficiários e não beneficiários. No entanto, em outros contextos, com o objetivo de não incorrer na exclusão de alguma família pobre, classifica-se as famílias segundo a renda, verificando se elas satisfazem determinadas necessidades que são consideradas essenciais pela sociedade. Assim, considera-se como população alvo todas aquelas pessoas com renda inferior a linha de pobreza; a participação desse grupo de pessoas no total da população indica a magnitude dos beneficiários (NAVARRO, 2005). Como resultado dessas formas de focalizar e classificar as famílias ou pessoas, gera-se um cadastro de famílias ou pessoas para a seleção de beneficiários, considerando que algumas variáveis podem perder o poder de prever a pobreza ou de discriminar beneficiário e não beneficiário (COADY et al, 2004).

No caso do Brasil, como o PBF foi estruturado para ser um programa universal, cuja elegibilidade está baseada na renda autodeclarada das famílias (*unverified means testing*) e uma vez que a informalização da pobreza brasileira não permite outra forma de mensuração, o público alvo foi constituído pela população abaixo da linha da indigência e da linha de pobreza (BRASIL, 200-?c).

Depois de definir a informação coletada, o cenário contrafactual, a focalização apropriada e as variáveis a serem consideradas para mensurar o impacto, um método de avaliação quantitativa deve-se expressar numericamente utilizando ferramentas estatísticas para a sua análise. O objetivo de utilizar estas ferramentas é estimar o impacto médio do programa e o seu nível de significância. Baker (2000), considerando as características na construção do cenário contrafactual, define dois tipos de desenhos para avaliar programas sociais: os desenhos experimentais e os não-experimentais (chamados também quase-experimentais).

[2] O índice SISBEN foi utilizado como instrumento para a seleção de beneficiários de subsídios de gasto social na saúde, educação, moradia, bem-estar familiar.

A avaliação com desenho experimental é dada quando a seleção do tratamento (ou beneficiários) e controle (ou comparação ou não beneficiários) do programa em estudo é realizada aleatoriamente; enquanto no desenho não experimental, não se seleciona aleatoriamente os grupos de tratamento (ou beneficiários) e controle (ou comparação ou não beneficiários) (EZEMINARI, RUDQVIST e SUBBARAO, 2002; DIAZ e HANDA, 2004).

Considerando os objetivos previstos no PBF e a implementação desse, se fez necessário mensurar os diferenciais atingidos pelo programa nos grupos de beneficiários do PBF. Uma avaliação do programa permitiria determinar os avanços conseguidos desde a sua implementação, no ano de 2003. Assim, a pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF), realizada em 2005, teve como objetivo avaliar o impacto do PBF, nas dimensões decorrentes das restrições orçamentárias e da operação de aspectos comportamentais ligados às condicionalidades do programa, tendo sido analisados os seguintes aspectos: Estrutura Relativa de Gastos, Antropometria, Saúde, Educação, Trabalho Infantil. Esta pesquisa ganhou importância pela abrangência que o Programa Bolsa Família (PBF) tem atingido na população brasileira (OLIVEIRA et al, 2007).

Na implementação do AIBF não foi possível efetuar uma avaliação experimental do programa. Em primeiro lugar, porque o programa foi criado a partir da migração e integração de vários programas prévios[3], impossibilitando a definição de um momento "antes" para realizar o experimento (ou aleatorização). Em segundo lugar, o Governo Federal estabeleceu uma meta de universalização do programa entre o público alvo, considerando a população abaixo da linha de miséria e da linha de pobreza, evitando determinar um grupo de controle aleatório, porque criaria um problema ético de negação do benefício a um determinado número de famílias (OLIVEIRA et al, 2007).

Como a avaliação dos registros administrativos do Cadastro Único (CadÚnico)[4] na época indicou um nível de inconsistência de informações que poderia comprometer o processo amostral, optou-se pela elaboração de uma pesquisa de linha de base domiciliar, de cunho

---

[3] Programa tais como: Bolsa Escola, Auxilio Gás, Bolsa Alimentação, Cartão Alimentação, e recentemente BPC e PETI.

[4] Base constituída por informações dos membros da família potencial que se inscreveu para receber algum benefício dos programas de transferência de renda do Governo Federal, disponibilizada pelo MDS, 2006.

observacional, denominada Pesquisa Linha de Base e que foi desenhada para servir como base para outras pesquisas, dentro do mesmo plano amostral. O desenho da amostra que determinou os procedimentos adotados na pesquisa de linha de base foi a partição da amostra em três grupos diferentes. O primeiro grupo foi formado pelas famílias beneficiárias do PBF em novembro de 2005 (casos). O segundo grupo, constituído por famílias, cadastradas no Cadastro Único dos Programas Sociais do Governo Federal, mas que ainda não eram beneficiárias do programa (controle tipo 1). O último grupo congregou as famílias que não eram nem beneficiárias do Programa Bolsa Família, nem eram cadastradas no Cadastro Único (controle tipo 2). Dessa forma, tornou-se possível considerar toda a população de famílias do país, inclusive inserindo uma pequena amostra de famílias não elegíveis para o programa. Previamente à pesquisa domiciliar com a finalidade de conseguir amostra com famílias de cada um desses três grupos, executou-se uma operação chamada de *screening* ou varredura. Através dessa operação categorizou-se os domicílios que apresentaram características que interessaram à investigação levando em conta os setores selecionados e o status em relação ao benefício do programa e em relação ao cadastramento.

Com a informação coletada e considerado o desenho não-experimental aplicado, realizou-se pelo AIBF as análises decorrentes das restrições orçamentárias e da operação de aspectos comportamentais ligados às condicionalidades do programa, por meio da técnica de Pareamento por Escore de Propensão (PSM)[5], o qual compara resultados de famílias similares do grupo de tratamento com o grupo de controle. Essa técnica possui o pressuposto de independência condicional com os atributos observáveis dos grupos de tratamento e controle, o que significa que se somente os atributos observáveis causam viés nas medidas de impacto, então a estimativa não-experimental dará uma boa medida de impacto (OLIVEIRA et al, 2007).

Em relação à analise realizada pela pesquisa AIBF, deve-se enfatizar que esta foi baseada na declaração dos domicílios acerca do recebimento dos benefícios de programas sociais, isto é, as famílias foram alocadas no grupos de tratamento e comparação 1 e 2, segundo a pesquisa de campo da AIBF. A razão disto decorre do fato do PBF ordenar as famílias para

[5] O termo "Pareamento" será referido para a técnica utilizada na avaliação de impacto dos programas sociais com o escore de propensão.

a seleção de beneficiários (renda familiar segundo linha de pobreza) conforme à insuficiência de renda, e de que os registros administrativos CadÚnico que continham informação dos beneficiários, na ocasião da pesquisa, indicavam um nível de inconsistência de informações com as ferramentas com as quais se contavam nesse momento.

Embora a informação declarada do recebimento do benefício por parte das famílias entrevistadas seja considerada adequada para análise na pesquisa AIBF, é possível que as respostas estejam influenciadas por aspetos subjetivos, como opiniões ou atitudes das pessoas, ainda que na pesquisa de campo a coleta de dados siga um conjunto de regras. No entanto, a realização prévia do *screening* impede que os aspetos subjetivos invalidem os resultados da pesquisa AIBF. Ainda assim, algumas variações ou diferenças de informação podem alterar a significância estatística dos impactos ou diferenciais entre os grupos de comparação e, conseqüentemente, os resultados da avaliação.

Considerando esta última reflexão, compete indagar sobre a possibilidade de utilizar os registros administrativo do CadÚnico, considerando alguma ferramentas estatísticas, para alocar às famílias ao grupo de tratamento e controle, segundo estes registros. A utilização dos registros administrativo, de forma geral, são vantajosos porque obtêm-se dados a baixo custo, com cobertura completa da população alvo, não contêm erro de amostragem e permitem separação específicas de sub-população (CEPAL, 2003b). Especificamente, o registro administrativo do CadÚnico, caracteriza-se por ser desenhado para registrar informações socioeconômicas das famílias com renda per capita mensal até meio salário mínimo por mês, por permitir a identificação das necessidades e características da família e seus membros, utilizar para selecionar beneficiários dos diversos programas sociais e possibilitar a geração de um número único nacional de identificação para os programas sociais, denominado "Número de Identificação Social" (NIS)[6], evitando duplicidades. Ponderando estas características, cabe saber se utilizando o CadÚnico, na alocação das famílias nos grupos de comparação, algumas variações ou diferenças com esta informação alterariam a significância estatística dos impactos ou diferenciais dos resultados da avaliação.

---

[6] NIS: Número de identificação social, por meio do qual o operador do Cadastro Único poderá localizar as pessoas cadastradas, atualizar dados do cadastro e verificar a situação do benefício.

Diante dos argumentos e reflexões feitas sobre a relação entre alocação das famílias nos grupos de tratamento e controle com os dados da pesquisa de campo e os registros administrativos, e, conseqüentemente, sobre as presumíveis variações ou diferenças dos resultados de impactos ou diferenciais da avaliação, surge a possibilidade de estudar e analisar uma comparação dos resultados de impacto da avaliação utilizando ambas as fontes de informação, para alocar às famílias no grupo de comparação. Além disso, ressalta-se que outros trabalhos de avaliação de impacto sugerem utilizar várias configurações de informações disponíveis, com vistas a realizar a avaliação de um programa, porque os procedimentos de seleção dos beneficiários podem enfrentar uma série de dificuldades e limitações (financeiras e políticas) no momento da implementação do programa e da avaliação (SKOUFIAS, 2006).

Para realizar a comparação dos resultados utilizando as duas fontes de informação na alocação das famílias nos grupos de tratamento e controle, tornou-se necessário utilizar o relacionamento das bases de dados obtidas da pesquisa de campo AIBF e as dos registros administrativos CadÚnico. Como essas bases procedem de diferentes fontes, foram tomadas algun cuidados para resolver os problemas de conciliação, sobretudo porque a informação combinada com o resultado do relacionamento deveria identificar a mesma entidade, que nesse caso corresponde a cada uma das famílias e seus respectivos membros. Assim, o processo de relacionamento de dados utilizados neste trabalho define-se como a comparação de dois ou mais registros das bases, que contêm informações de identificação para determinar se estes registros referem-se à mesma entidade (HOWE, 1988). Nesse ponto, vale ponderar que para os trabalhos que se valem de banco de dados, quando existe algum número identificador único comum dos registros, o problema é facilitado; mas, caso contrário, ao buscar relacionar os dados há que se considerar outras variáveis, tais como nome, sexo, data de nascimento, código de município, dentre outras (CAMARGO e COELI, 2002a). Estas características das bases de dados tornam-se importantes já que, na área social, com freqüência nas bases de dados disponíveis, a informação com códigos ou identificadores unívocos do indivíduo ou eventos não estão presentes requerendo uma estratégia onde se considere mais de uma variável identificadora da entidade ou do indivíduo que se está relacionando.

Duas principais estratégias de relacionamento de bases de dados foram utilizadas nesta tese: determinística e o probabilística. A estratégia de relacionamento determinístico utiliza

um identificador único[7] e classifica os registros comparados como pares ou não pares. Esta estratégia é de fácil entendimento e implementação, embora possa ser laboriosa e consumir muito tempo em algumas situações, envolvendo decisões subjetivas. (COPAS e MILTON, 1990). O relacionamento probabilístico se baseia na teoria estatística desenvolvida por Fellegi e Sunter (1969), e é apropriado quando as bases de dados a relacionar não contenham ao menos um identificador único, comum nas bases a serem relacionadas.

O relacionamento determinístico foi aplicado quando, em ambas as bases a serem relacionadas, a informação do NIS das pessoas esteve presente. No caso em que esta informação estivesse incompleta ou contivesse erros na declaração aplicou-se o relacionamento probabilístico, utilizando informações comuns em ambas as bases, tais como: nome, sexo, data de nascimento e município de residência.

Como resultados deste relacionamento[8] das bases de dados, foi possível contar com uma base contendo informações conjuntas, e conseqüentemente, as mesmas famílias conseguiram ser alocadas nos grupo de tratamento e controle, segundo cada fonte de informação. Com estas informações, o passo seguinte foi analisar as mudanças e variações encontradas nos resultados de impacto, quando as famílias são alocadas nos grupos de comparação, segundo a pesquisa de campo AIBF e alocados segundo o relacionamento, com os registros administrativos CadÚnico. Neste caso as mudanças e variações dos resultados de impacto serão expressas com a sensibilidade que apresentam os resultados quando se utilizam as duas alocações de famílias referidas na tese. Além disso, como se está comparando resultado de impacto, decidiu-se analisar a sensibilidade dos resultados da seção de educação da pesquisa AIBF que retratam a situação educacional da população entre 7 e 14 anos, cujas variáveis resultados foram: freqüência à escola, evasão da escola, progressão na escola, repetência escolar, e alocação entre trabalho e estudo.

Para encontrar os resultados de avaliação de impacto de educação foi utilizado o método de Pareamento (*matching)* não-experimental, dado que o desenho do programa não foi conduzido aleatoriamente e seguindo o trabalho realizado pela pesquisa AIBF, nesta tese

---

[7] Código ou identificador de um indivíduo ou entidade que permite distinguir univocamente o ente (Indivíduo, família, empresa, entre outros).

[8] O termo "relacionamento" será utilizado quando nós referimos a relacionamento das bases de dados realizado entre a base da pesquisa AIBF como os registros Administrativos do CadÚnico.

também foi utilizado a metodologia de pareamento por escore de propensão (PSM). Calculou-se, então, o efeito médio do tratamento sobre o tratado através de distintos algoritmos de *matching* não paramétricos. Para completar esta ressalva metodológica, o diferencial obtido na linha de base não é uma medida de impacto, ou seja, uma medida que possa ser consideradas como tal, assim, duas condições de cautela devem ser mencionadas: primeira: que o viés variável de seletividade (não observável) está presente e não será corrigida pelo método da diferença nas diferenças, e a segunda condição, que não há um controle sobre o tempo de exposição dos beneficiários ao programa (efeito duração) e nem sobre o valor do benefício recebido durante a totalidade do período (efeito dose) (Oliveira et al, 2007). Além disso, com o enriquecimento das informações obtidas com o produto do relacionamento das bases, esta tese propõe uma análise das famílias que recebem o beneficio do PBF, utilizando uma opção para realizar a análise dos diferenciais do programa, denominada Desenho de Regressão Descontínua (RD), modelo que recorre em uma forma particular de identificação das variáveis instrumentais. Neste caso será utilizada a descontinuidade no processo de alocação ao PBF, para identificar o efeito causal dos beneficiários do programa.

Neste contexto o trabalho da tese é uns exercícios metodológicos, orientado pela seguinte questão: A aplicação do relacionamento entre bases de dados de uma pesquisa de campo e registros administrativos para alocar as famílias nos grupos de comparação capta em forma diferente, os resultados de avaliação de impacto dos programas sociais?

O procedimento utilizado neste estudo deverá fornecer uma boa alternativa para o aperfeiçoamento dos métodos não-experimentais utilizados na avaliação dos programas sociais, desta forma será possível analisar o efeito da alocação das famílias nos grupos de comparação para avaliação do impacto dos resultados, porque acrescenta uma nova configuração para alocar estas famílias. Além disso, do ponto de vista metodológico os estudos longitudinais têm sido um desafio para os estudos de população. Com o procedimento de relacionamento de bases de dados aplicado neste estudo, pretende-se obter um acompanhamento ou seguimento dos domicílios imersos no estudo da avaliação do impacto dos programas sociais ao longo do tempo.

Assim, a finalidade desta tese é explorar as possibilidades únicas que são abertas pelo relacionamento de bases de dados para analisar a sensibilidade dos resultados de impacto dos programas sociais de transferência de renda, quando se utilizam dois tipos de fontes de

informação para a alocação das famílias nos grupos de comparações. Para tal análise, será realizada a aplicação específica da avaliação de impacto do Programa Bolsa Família nos indicadores da educação, utilizando a alocação das famílias nos grupos de comparação, segundo a pesquisa de campo AIBF e os registros administrativos CadÚnico. São os seguintes os objetivos específicos:

• Construir uma base de dados com informações combinadas por família, a partir das bases da pesquisa de campo de domicílios AIBF e dos registros administrativos do CadÚnico, através do relacionamento de base de dados determinístico e probabilístico.

• Adicionar à base da pesquisa de campo domiciliar AIBF a informação obtida como produto do relacionamento de dados, substituindo-se os dados declarados com alguns viés pelas famílias na pesquisa de campo domiciliar AIBF, pelos encontrados nos registros administrativos CadÚnico.

• Medir os diferenciais da educação do programa de transferência de renda Bolsa Família, a partir de um conjunto de indicadores e do modelo econométrico escolhido, que procuram retratar a situação educacional das crianças.

• Comparar os resultados dos diferenciais na educação obtidos com alocação das famílias nos grupos de comparação segundo a pesquisa de campo AIBF e registros administrativos CadÚnico.

• Empregar a informação de renda familiar dos registros administrativos, para avaliar os diferenciais na educação, recorrendo a uma forma particular de identificar os grupos potencialmente beneficiários e não beneficiários do PBF, utilizando o método da regressão descontínua.

• Identificar as vantagens e desvantagens da aplicação do relacionamento de bases de dados para alocar as famílias segundo o registros administrativos CadÚnico, avaliando a sensibilidade dos resultados dos diferenciais na educação do PBF.

A presente tese, além desta introdução, está organizada como segue. O capítulo 2 apresenta a concepção básica de relacionamento de base de dados. O capítulo 3 aborda os métodos de implementação e avaliação dos programas sociais, destacando o PBF e o AIBF. O capítulo 4 apresenta a aplicação do relacionamento de bases de dados para os dados coletados da pesquisa de campo AIBF e registros do CadÚnico. O capítulo 5 discute as

aplicações dos métodos de avaliação e os resultados encontrados. Por fim, o capítulo 6 apresenta as considerações finais da tese.

2 RELACIONAMENTOS PROBABILÍSTICO E DETERMINÍSTICO DE BASES DE DADOS

Este capítulo está composto em cinco seções. Tem-se, inicialmente, um breve histórico do relacionamento de bases de dados. Em seguida formaliza-se a definição de relacionamento determinístico. Na terceira seção, introduz-se o fundamento matemático da teoria do relacionamento probabilístico e aborda-se, na seqüência, os conceitos da teoria estatística que permitem colocar em prática o relacionamento de bases de dados. Na quarta seção são descritos os avanços computacionais utilizados no relacionamento probabilístico, exemplificando-se a utilização do relacionamento de bases de dados no Brasil e em outros países. Finalmente, discutem-se as pesquisas de campo e dos registros administrativos e a integração dessas informações em relacionamento nas bases de dados.

2.1. Relacionamento de Dados

Para a construção de relacionamentos das bases de dados, informações que combinam indivíduos ou entidades a partir de várias fontes de dados, são freqüentemente necessárias e crescentemente possíveis. Em estudos médicos, por exemplo, uma coorte ou grupo de indivíduos é seguido para averiguar uma situação de morbidade. Uma forma que pode ser utilizada em tais estudos longitudinais é seguir o grupo de interesse fisicamente, porém tal método é limitado pelos recursos econômicos, restringindo o tamanho e tipo dos grupos que podem ser seguidos. Outro modo de seguir coortes de indivíduos é através da supervisão de bases de dados que contêm resultados contínuos (ex. registros civis, certificados de morte, bases de dados de escola pública) e a utilização de relacionamento de dados ou *record linkage* (GOMATAM e CARTER, 1999).

A partir de uma perspectiva global, relacionar bases de dados deveria ser familiar, já que este é constantemente aplicado em atividades cotidianas, como por exemplo, sempre que se busca um número na lista telefônica, um serviço nas páginas amarelas ou um produto em um catálogo. Para buscar estas informações pode-se exemplificar com a seguinte preceituação do procedimento, inicialmente introduz-se certas informações como o nome e sobrenome, nome da organização, ou o logradouro (embora esta procura esteja limitada

pelos grupos e ordem utilizados na compilação do diretório). Assim, para procurar um número de telefone, examina-se o diretório pela área geográfica apropriada e, usando o mais recente diretório provido pela empresa de telefonia e comunicação, seleciona-se a seção para indivíduos ou para negócio e organizações profissionais. A seguir, busca-se o item procurado segundo o índice alfabético. Em alguns casos, quando há variações de grafia nos nomes e sobrenomes do subscritor ou logradouros, utilizam-se decisões subjetivas para identificar o número de telefone procurado (GILL, 2001)

A partir da idéia básica do relacionamento de dados, pode-se formalizar o termo de "relacionamento de dados" como o processo de comparação de dois ou mais registros, que contêm informações de identificação para determinar se estes registros referem-se à mesma entidade (HOWE, 1988). Embora, o conceito sugira ser uma simples extensão da idéia básica, existem muitos interessantes e desafiantes problemas técnicos que devem ser resolvidos para empreender o relacionamento de dados em grande escala.

Existem duas principais estratégias de relacionamento de dados, o determinístico e o probabilístico. A estratégia de relacionamento determinístico utiliza um identificador único que permite distinguir univocamente ao ente (indivíduo, família, empresa, entre outros) e classifica os registros comparados como pares ou não pares. Esta estratégia é comumente de simples entendimento e implementação, embora, em alguns casos envolvendo decisões subjetivas, possa ser laboriosa e consumir muito tempo. O relacionamento probabilístico se baseia na teoria estatística desenvolvida por Fellegi e Sunter (1969), e é apropriado quando as bases de dados a relacionar não contenham ao menos um identificador único, comum às bases a serem relacionadas, bem como quando os resultados puderam variar entre a total concordância (exato) à total discordância ou com vários níveis de concordância entre eles (CHRISTEN e CHUCHES, 2006?).

O processo de relacionar registros tem adquirido vários nomes em diferentes comunidades de usuários. Enquanto os epidemiologistas e estatísticos falam de relacionamento de dados – *record linkage*, o mesmo processo é freqüentemente chamado como emparelhamento de dados – *matching data* ou como problemas de identidade de objeto por cientistas da computação, sendo também conhecido como processo de combinar/remover (ou *merge/purge*) e como limpeza de listas em processo comercial de bases de dados de cliente ou listas de clientes (*mailing lists*). Historicamente, os estatísticos e cientistas informáticos

desenvolveram as próprias técnicas, e até recentemente poucas referências cruzadas poderiam ser achadas (CHRISTEN e CHUCHES, 2006?).

Um aspecto importante nesta metodologia é que se ressalta nos trabalhos em que se aplicam o relacionamento de bases de dados, a sua utilidade para a melhoria da quantidade e qualidade das informações nas áreas de pesquisas correspondentes. Além disso, em muitos estudos o relacionamento de dados é utilizado como uma ferramenta importante quando se precisa conhecer informação adicional diferente daquelas que se contam inicialmente (GILL, 2001).

2.2. O Relacionamento determinístico ou exato (*Deterministic record linkage*)

A técnica ou procedimento mais adequado para ser utilizado é o relacionamento determinístico ou exato, quando o identificador único[9] permite distinguir univocamente o ente, sendo útil para unir ou relacionar conjunto de bases que contêm diferentes informações. Se o identificador único de indivíduo ou entidade está disponível em todas as bases de dados a serem relacionadas, então o problema é trivial. Dessa forma, com uma simples rotina ou operação em algum sistema de administração e manipulação de bases de dados pode ser realizado um relacionamento de bases de dados (CAMARGO e COELI, 2000; WHALEN et al, 2001)

Tal relacionamento é geralmente fácil para a implementação e o entendimento, sobretudo pelas praticidades não estatísticas utilizadas. No entanto, quando o processo envolve tratar na implementação questões subjetivas, ele pode ser laborioso e consumir muito tempo.

Para autores que discutem o método determinístico, a existência da pouca literatura é um indício que faz considerá-lo como uma estratégia simples de ser utilizada. Roos e Wajda (1991) sugerem utilizar uma medida chamada de "número médio de casos por bloco", para estimar a quantidade de informação relacionada em qualquer base de dado ou arquivo. Boussy e Scott (1993) apresentam uma visão geral dos métodos de relacionamento incluindo alguma discussão do método determinístico. Neste tipo de relacionamento, ao

[9] São exemplos deste identificador único: número de registros nacional, número de identificador nacional, número de seguro social, número de cadastros de pessoais físicas, entre outros.

comparar dois registros, por exemplo, o primeiro e último nome, os registros só são considerados pares se os nomes nos dois registros concordarem em todos os caracteres. No RD os registros podem ser relacionados através de uma sucessão de passos, e em cada um deles decide-se o estado de relacionamento do par de registros (par ou não par), considerando uma concordância exata em um subconjunto particular de identificadores. Neste caso, em cada passo, os pares únicos são extraídos do procedimento; os duplicados e as observações restantes que não forem relacionadas em cada uma das duas bases de dados (os resíduos) formam parte dos dados para o próximo passo no processo de relacionamento que continua com um subconjunto diferente de identificadores. Os passos implementados subseqüentemente serão menos restritivos que os dos passos anteriores. Desta forma a sucessão de passos que se pode implementar depende muito da quantidade de conhecimento que se tem dos dados a serem relacionados, já que, por se tratar do pareamento exato[10], existem apenas dois resultados: par verdadeiro ou não par verdadeiro (GOMATAM e CARTER, 1999).

Em relação a esta metodologia, enfatizam-se alguns trabalhos que discutem o desenvolvimento integrado de um projeto de relacionamento de bases de dados. São eles: o projeto dos Estados Unidos, que relaciona registros do Centro de Cuidados Intensivos Regional das crianças pré-natais (RPICC) com os resultados educacionais subseqüentes destas crianças no Departamento de Educação do Estado da Flórida (1999); o relacionamento de informações do Censo de Nova Zelândia, com os dados reportados dos registros civis de mortalidade (NZCMS), cujo objetivo é determinar a associação de fatores socioeconômicos coletados no censo com as causas da morte (1991).

Nesta investigação, será considerada como uma etapa prévia do todo o procedimento do relacionamento das bases de dados, tratada como uma das estratégias a utilizar no relacionamento de dados. O relacionamento determinístico ou exato considera par somente "todos ou nada" (*"all or nothing"*), isto é, concordância única de todos os algarismos ou caracteres do identificador chamado de "variável identificadora de relacionamento" (*match key*) (GOMATAM e CARTER, 1999).

---

[10] Neste caso o Pareamento é considerado como o relacionamento exato de bases de dados. este termo só será utilizado neste capitulo como este significado.

2.3. Relacionamento probabilístico de dados (*Probabilistic record linkage*).

2.3.1. Desenvolvimento no tempo do relacionamento probabilístico.

A primeira referência que cita o termo de relacionamento de dados – *record linkage* é encontrado no trabalho do Dr. Halbert Dunn, chefe de *the U.S. National Office of Vital Statistics* (DUNN, 1946). Dunn (1946) declarou a necessidade de relacionar registros no Canadá, promovendo a utilização do número de certidão de nascimento como um identificador eficiente e único para relacionar os dados dos registros do sistema estatístico vital (WEBER, 1995).

Métodos computacionais de relacionamento de dados emergiram como uma ferramenta importante nos anos 40 e 50, quando despertava o interesse de pesquisadores pela criação da árvore genealógica de indivíduos para pesquisas genéticas; até então, muitos dos projetos de relacionamento de bases de dados estavam baseadas em métodos heurísticos *ad-hoc*. (NEWCOMBE et al, 1959).

A primeira aplicação prática do relacionamento de dados por meios computacionais foi feita nos anos 50, utilizando registros vitais civis para localizar doenças hereditárias. Em 1959 foi proposto utilizar relacionamento de bases de dados para combinar informações diferentes de dois registros que representam o mesmo indivíduo (NEWCOMBE et al, 1959).

Usando técnicas computacionais, a idéia básica do relacionamento de dados probabilístico foi introduzida por Newcombe e Kennedy em 1962. Adicionalmente, com a criação em 1960 da fundação do relacionamento de bases de dados probabilístico, pesquisadores como DuBois (1969), Nathan (1967), Tepping (1968), e Fellegei e Sunter [1969] desenvolveram várias aproximações matemáticas para o relacionamento de bases de dados probabilístico. Embora cada aproximação fosse diferente, os conceitos fundamentais estavam baseados na mesma teoria. Para todo par de registros comparado, cada variável ou campo (i.e. determinado nome, sobrenome, sexo e idade) era comparado, e o registro classificado como par, não par, ou indeterminado. A realização de cada comparação era usada para calcular os pesos para os respectivos campos utilizados. Logo, considerando a adição dos pesos poder-se-ia obter uma estatística de teste, resultado utilizado na determinação das classificações dos registros pareados (KIRKENDALL, 1995).

A aproximação de DuBois (1969) sobre o relacionamento de dados baseou-se em combinações da distribuição binomial. Nathan (1967) focalizou seus trabalhos no relacionamento de novos registros a uma base de dados mestre completa e sem erros. Já Tepping (1968) utilizou regras de otimização para minimizar o custo de registros pareados erroneamente. Fellegi e Sunter (1969) foram os que avançaram mais na aproximação matemática do relacionamento probabilístico, desenvolvendo o Método Probabilístico Bayesiano com base nas idéias de Newcombe. A teoria proposta por Fellegi e Sunter tentou limitar o número de registros indeterminados (não classificados), embora o grau ótimo deste método dependa do conhecimento prévio das probabilidades utilizadas no cálculo dos pesos.

2.3.2. Teoria estatística do relacionamento probabilístico

i) Termos utilizados no relacionamento probabilístico

1. Pareamento ou relacionamento exato, utilizado freqüentemente no relacionamento determinístico. Quando duas bases de dados contêm o mesmo identificador único seus registros podem ser relacionados por meio desse identificador. O relacionamento baseado nesse identificador único é denominado de “pareamento exato ou relacionamento exato”. O identificador único pode ser uma só variável ou uma combinação de variáveis, dependendo da suficiente qualidade da variável a ser utilizada na combinação, para definir um registro único.

2. Dois registros são considerados como “pares” quando ao relacionar-se pertencem à mesma pessoa/entidade ou evento. Considerando que a função do relacionamento de bases de dados é determinar quais registros relacionados ou pareados são considerados como pares, o termo utilizando como “par” para os registros que pertencem à mesma entidade, pode ser diferenciado, quando utilizamos a palavra "par verdadeiro" referendo-se à mesma entidade.

3. Dois registros são considerados “*link* - relacionados”, se por algum procedimento precisa-se determinar se dois registros se referem à mesma unidade (seja uma pessoa, agência, entidade ou evento). Quando se produz relações de registros (*links*) o procedimento de relacionamento de dados, indica que não todo “par verdadeiro” é uma

relação ou link, e não toda relação ou link é um "par verdadeiro", como se mostra no quadro seguinte:

Quadro 2.1 – Comparação e decisão de registros a relacionar ou linkar.

Tipo de relação	Par verdadeiro	Par não verdadeiro.
Relação ou *link*	Resultado correto	Relações ou *links* falsos positivos
Não relação ou *Non-link*	Relações ou *links* falsos negativos	Resultado correto

4. "O pareamento ou relacionamento" é o processo de comparação de registros e decisão onde esses são relacionados ou linkados. As variáveis utilizadas no processo de relacionamento são denominadas 'variáveis do relacionamento', 'campos de relacionamento' ou 'variáveis de comparação'. Este procedimento é bastante utilizado em relação à manipulação de dados que tem como objetivo comparar registros de duas ou mais bases de dados, e se refere propriamente ao processo conhecido como *"record linkage"* ou "relacionamento de dados".

5. Arquivos de relacionamento - Sejam dois arquivos, A e B, o objetivo é comparar um registro de cada um dos arquivos, e logo decidir se os registros a serem relacionados devem ser unidos ou não como um "par verdadeiro". Ilustramos este conceito por meio de um exemplo:

FIGURA 2.1 – Registros a serem comparados de dois arquivos ou bases de dados: A x B (Exemplo hipotético).

Arquivo ou Base	A
Nome	Maria Souza
Data de nascimento	15/07/1975
Sexo	Feminino
Endereço	Rua Três 125, Minas Gerais.

Arquivo ou Base	B
Nome	Maria Sousa
Data de nascimento	15/07/1977
Sexo	Feminino
Endereço	Rua Rios 125,

Na FIG. 2.1 observam-se dois registros, no qual o sobrenome da pessoa varia apenas numa letra, o ano de nascimento no último digito e o endereço no nome da rua, no entanto, o sexo da pessoa é igual. Neste caso, cabe decidir se a informação trata-se da mesma pessoa ou não.

ii) Parâmetros do relacionamento probabilísticos.

Embora a teoria do relacionamento probabilístico tenha sido desenvolvida por vários matemáticos, tais como Newcombe *et al* (1959), Howe e Lindsay (1981), Newcombe (1988), couberam a Fellegi e Sunter (1969) as primeiras apresentações do modelo matemático e dos fundamentos teóricos rigorosos para o relacionamento probabilístico considerando a aproximação computacional. A teoria foi desenvolvida ao longo da linha de hipótese clássica que testa e proporciona orientação para a o tratamento do problema de relacionamento, e torna as bases fundamentais para a teoria do relacionamento de bases de dados.

Os fundamentos básicos considerados nessa teoria começam definindo dois arquivos de registros ou conjunto de dados, A e B, contendo n_A e n_B registros respectivamente. Assumindo-se que dois arquivos ou conjuntos são relacionados, o conjunto de pares possíveis será dado por:

$$AxB = \{(a,b);\ a\ \varepsilon\ A,\ b\ \varepsilon\ B\}$$

que é a união de dois conjuntos disjuntos, representados por:

$$M = \{(a,b);\ a = b,\ a\ \varepsilon\ A,\ b\ \varepsilon\ B\} \quad e \quad U = \{(a,b);\ a \neq b,\ a\ \varepsilon\ A,\ b\ \varepsilon\ B\},$$

designando como pares considerados "verdadeiros", "não pares verdadeiros", respectivamente.

Como cada conjunto contem n_A e n_B registros, estes possuem também diversas variáveis, que descrevem informações pertencentes a um individuo específico, como de sobrenome, nome, idade, sexo, raça, entre outros.

Para um registro $a \in A$ e registro $b \in B$, a informação disponível sobre o registro é denotado por $\alpha(a)$ e $\alpha(b)$ respectivamente. Quando comparamos o par de registros, um de A e um de B, a comparação ou vetor de concordância, γ, é denotado por,

$$\gamma\ [\alpha(a) \text{ e } \alpha(b)] = \{\gamma^1[\alpha(a) \text{ e } \alpha(b)], \ldots, \gamma^k[\alpha(a) \text{ e } \alpha(b)]\}$$

que é uma função sobre o conjunto de todos os n_A x n_B registros pareados. Na qual γ é uma função sobre AxB, γ^i é uma vetor de comparação sobre uma só variável e K variáveis são

apresentados em cada γ. Cada γ^i considera diferentes valores quando diversas variáveis concordam.

Uma concordância ocorre quando as variáveis de comparação da população são equivalentes. Uma concordância parcial existe quando uma parte das variáveis de comparação é a mesma ou existe evidência significante para manter a concordância A discordância apresenta-se quando as variáveis de comparação diferem sem grau definido de semelhança.

O conjunto de todas as possíveis realizações de γ observado é denominado de Γ, o espaço de todos os possíveis vetores de comparações. Com base neste vetor de comparação γ a decisão pode se realizar para um par de registros, e definem-se três possíveis resultados para o par (*a*, *b*).

1) (*a*,*b*) é um par verdadeiro, tal que $(a,b) \in M$, denominando-se como relações ou enlaces ou *links* positivos, denotado por A_1.

2) (*a*,*b*) é um não par verdadeiro, tal que $(a,b) \in U$, chamado relações ou enlaces ou *links* negativos, denotado por A_3.

3) (*a*,*b*) é um possível par (ou enlaçado ou *link*) ou par indeterminado, denotado por A_2.

Logo a regra de relacionamento ou *link L* é definida agora como a distribuição de Γ, sobre um conjunto funções de decisão aleatória $D = \{d(\gamma)\}$, onde:

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\}; \gamma \in \Gamma$$

e

$$\sum_{\gamma \varepsilon \Gamma} P(A_1 \mid \gamma) = 1$$

A regra de relacionamento ou *linkage* considera uma probabilidade para cada uma das três possíveis ações.

Para alguns, ou até mesmo todos os possíveis valores de γ , a função de decisão pode degenerar-se, assinalando para uma das ações uma probabilidade de um (FELLEGI e SUNTER, 1969).

Além das ações mencionadas, também se deve considerar que nem todas estejam corretas (isto é, a dois registros pode ser atribuída a probabilidade de ser um par verdadeiro quando, ele realmente não é um par), evento que é causado pela probabilidade de unidades mal – classificadas, os quais são taxas de erro que precisam ser consideradas para a regra de relacionamento ou *linkage*. (GU, 1983)

Assim, para um par de registros (*a*,*b*) aleatoriamente selecionada para a comparação de duas populações A x B, γ é considerado como uma variável aleatória. A probabilidade condicional de γ observada, dado o registro pareado (*a*, *b*) é um par verdadeiro definido por,

$$m(\gamma) = P(\gamma \mid (a,b)\, \varepsilon\, M) = \sum_{\gamma\, \varepsilon\, M} P(\gamma).P[(a,b) \mid M]$$

similarmente

$$u(\gamma) = P(\gamma \mid (a,b)\, \varepsilon\, U) = \sum_{\gamma\, \varepsilon\, U} P(\gamma) P[(a,b) \mid U],$$

denota a probabilidade condicional de γ observado, dado que o registro pareado (*a*, *b*) é um não-par verdadeiro.

Logo há duas classes de possíveis erros mal-classificados: falsos pares e falsos não pares. A probabilidade de um par verdadeiro ser falso é:

$$u = P(A_1 \mid U) = \sum_{\gamma\, \varepsilon\, \Gamma} u(\gamma) P(A_1 \mid \gamma)$$

e a probabilidade de um não par verdadeiro ser falso é

$$m = P(A_3 \mid M) = \sum_{\gamma\, \varepsilon\, \Gamma} m(\gamma) P(A_3 \mid \gamma)$$

Para um valor fixo da taxa de pares falsos (μ) e taxa de não pares falsos (λ), Fellegi e Sunter (1969) definem uma regra ótima de enlace, sobre Γ nos níveis μ e λ, denotando L(μ, λ, Γ) como a regra pelo qual,

$$P(A_1 \mid U) = \mu, \quad P(A_3 \mid M) = \lambda \quad e \quad P(A_2 \mid L) \leq P(A_2 \mid L')$$

para todas outras regras L'.

A regra de relacionamento ótima maximiza a probabilidade de classificar um par em A_1 e A_3, sujeito aos níveis fixos de erro definidos na regra de relacionamento. Esta metodologia é desejável porque atenua a probabilidade de classificar um par no conjunto A_2 (pares não conclusivos) que requerem revisão manual. Deste modo, quando existe um número grande de pares não conclusivos (A_2), o tempo e esforço que se precisa realizar para definir estes pares como conclusivos, desacreditam o uso de métodos probabilísticos computadorizados (JENSEN, 2004).

Sobre o espaço Γ, define-se a regra de relacionamento L_o, seguidamente, um único ordenamento de o conjunto finito de possíveis realizações de γ é realizado. Se para qualquer valor γ, o valor de ambos $m(\gamma)$ e $u(\gamma)$ é igual a zero, então a probabilidade (incondicional) de realizações de γ é igual a zero, e não precisa-se ser incluída em Γ. Logo ordenando todas as restantes de realizações γ, de tal um modo que a sucessão de relações de probabilidade, $R = m(\gamma) / u(\gamma)$ é qualquer função monotonamente crescente e associado a um λ arbitrariamente.

Para melhor entendimento, ordenam-se o conjunto de $\{\gamma\}$ e indexa-se por sub-índices i; (γ=1, 2, …, N_Γ) e $u_i = u(\gamma_i)$, e $m_i = m(\gamma_i)$.Seja (μ, λ) um par aceitável de níveis de erros e escolhendo, n e n' tal que

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i \quad e \quad \sum_{i=n'}^{N_\Gamma} m_i < \lambda \leq \sum_{i=n'+1}^{N_\Gamma} m_i$$

na qual N_Γ é o número de pontos do espaço Γ, e além disso, assume que se esta condição é satisfatória então $1 < n \leq n' - 1 < N_\Gamma$. Assim a regra de relacionamento $L_0(\mu,\lambda,\Gamma)$ pode definir o seguinte:

Para um vetor de comparação observado, γ_i, que se encontra no conjunto A_1 (relações ou *link* positivos), se $i \leq n$-1, encontra-se em A_2 (status não conclusivos) se $n< i\leq n'$-1; e encontra-se em A_3 (relaciones ou *links* negativas) se $i \geq n'$+1. Quando $i = n$ ou $i = n'$, então uma decisão aleatória é exigida para achar os níveis de erros μ e λ exatamente.

Isto pode ser representado formalmente como

$$d(\gamma_i) = \begin{cases} (1,0,0) & i \leq n-1 \\ (P_\mu, 1-P_\mu, 0) & i = n \\ (0,1,0) & n < i \leq n'-1 \\ (0,1-P_\lambda, P_\lambda) & i = n' \\ (0,0,1) & i > n'+1 \end{cases}$$

onde, P_μ e P_λ são definidos como as soluções para as equações

$$u_n P_\mu = \mu - \sum_{i=1}^{n-1} u_i \quad e \quad m_{n'} P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i$$

Teorema. se $L_0(\mu,\lambda,\Gamma)$ é a regra de relacionamento definido por $d(\gamma_i)$, então L é a melhor regra de relacionamento sobre o espaço Γ nos níveis (μ,λ).

Corolário 1: Se $\mu = \sum_{i=1}^{n} u_i$, $\lambda = \sum_{i=n}^{N_\Gamma} m_i$, $n < n'$, então $L_0(\mu,\lambda,\Gamma)$, a melhor regra de relacionamento nos níveis de erros (μ,λ), transforma-se

$$d(\gamma_i) = \begin{cases} (1,0,0) & se \quad i \leq i < n \\ (0,1,0) & se \quad n < i \leq n' \\ (0,0,1) & se \quad n' \leq i < N_\Gamma \end{cases}$$

Se definem dois limiares tal que

$$t_\mu = \frac{m(\gamma_n)}{u(\gamma_n)} \quad e \quad t_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

Então, a regra de relacionamento, $d(\gamma_i)$, pode ser escrita equivalentemente como,

$$d(\gamma_i) = \begin{cases} (1,0,0) & se \quad t_\mu \leq \dfrac{m(\gamma)}{u(\gamma)} \\ (0,1,0) & se \quad t_\lambda < \dfrac{m(\gamma)}{u(\gamma)} < t_\mu \\ (0,0,1) & se \quad \dfrac{m(\gamma)}{u(\gamma)} \leq t_\mu \end{cases}$$

Portanto, a decisão da regra de relacionamento está baseada nos valores limiares da razão de verossimilhança *R*.

iii) Pressuposto simplificado para o vetor de concordância γ

Na prática, os diferentes valores de γ podem ser tão grandes que a estimação das probabilidades de $m(\gamma)$ e $u(\gamma)$ tornam-se impraticável. Nestes casos é conveniente realizar algumas suposições simplificando sobre a distribuição γ.

Assumindo que as componentes do vetor γ podem ser reordenadas e agrupadas tal que

$$\gamma = \{\gamma^1, \gamma^2, \ldots, \gamma^k\}$$

e que os componentes são mutuamente independente com respeito à distribuição condicional. Assim:

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \ldots \cdot m_k(\gamma^k)$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \ldots \cdot u_k(\gamma^k)$$

onde

$$m(\gamma^i) = P(\gamma^i \mid (a,b)\, \varepsilon\, M)$$

$$u(\gamma^i) = P(\gamma^i \mid (a,b)\, \varepsilon\, U),$$

Tal suposição permite a conclusão que, $\gamma^1, \gamma^2, \ldots, \gamma^k$ são distribuídos condicionalmente independentemente. Esse suposto de independência associada com os erros dos campos, refere-se a que, se existe erros de um determinado campo tal como o nome, estes são independentes dos erros encontrados em outro campo tal como a idade (FELLEGI e SUNTER, 1969).

iv) Os pesos

Utilizando as componentes das probabilidades associados à decisão da regra de relacionamento, o peso para um campo ou variável pode ser calculado. O cálculo usado depende se os valores no campo concordam ou não. Se eles concordam, um peso positivo será gerado, e se eles discordam será gerado um peso negativo. Assim, o tamanho do peso mede a evidência de que os valores provêem sobre o par de registros comparado ser um par verdadeiro.

Seja qualquer função monotonamente crescente de $m(\gamma)/u(\gamma)$ que pode ser utilizada como um teste estatístico para definir a regra de comparação. O algoritmo desta razão é particularmente utilizado e é definido como o vetor de pesos

$$w^{k}\left(\gamma^{k}\right) = \log\left[m\left(\gamma^{k}\right)\right] - \log\left[u\left(\gamma^{k}\right)\right]$$

Onde, $k = 1, 2, ..., K$ é o número total de campos ou variáveis a serem comparadas. Então os pesos podem ser somados sobre todos os campos dados para os valores dos dois registros de comparação, ou estatística de teste, de

$$w(\gamma) = w^{1} + w^{2} + \ldots + w^{k}\ .$$

Logo o teste estatístico $w(\gamma)$ é utilizado para facilitar o entendimento no caso que, $u(\gamma)=0$ ou $m(\gamma)=0$, então $w(\gamma) = +\infty$ (ou $w(\gamma) = -\infty$) no sentido que $w(\gamma)$ é grande (ou pequeno) do que, qualquer número finito dado.

Assumindo que γ^{k} pode tomar sobre n_k diferentes configurações, $\gamma_1^{k}, \gamma_2^{k}, \ldots, \gamma_{nk}^{k}$. Então

$$w_{j}^{k} = \log\left[m\left(\gamma_{j}^{k}\right)\right] - \log\left[u\left(\gamma_{j}^{k}\right)\right]$$

Assim, os pesos são definidos positivos quando $m(\gamma_j^{k}) > u(\gamma_j^{k})$ e negativos quando $m(\gamma_j^{k}) < u(\gamma_j^{k})$. Esta propriedade é preservada para os pesos associados com o total de configurações de γ.

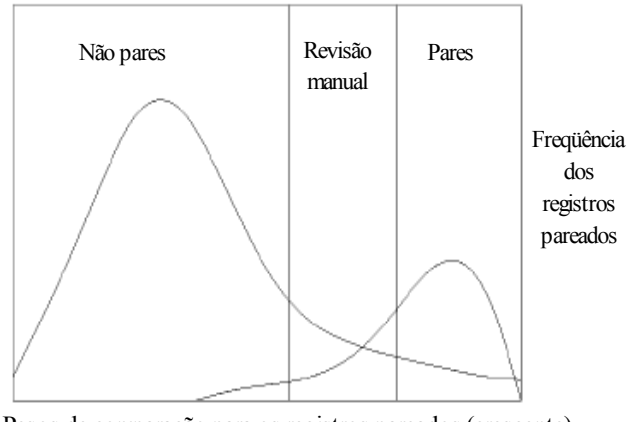
O total número de configurações para γ é $n_1, n_2, ..., n_k$, mas pela propriedade aditiva dos pesos definida para as componentes isto é suficiente para determinar apenas $n_1 + n_2 + ..., +$

n_k pesos. Então o peso associado para qualquer γ é encontrado utilizando a propriedade aditiva.

Na prática têm-se utilizado diferentes metodologias para encontrar os pesos. Fellegi e Sunter (1969) propõem duas metodologias para calcular o peso utilizado em seus modelos. O primeiro método pressupõe que a informação, *a priori,* está disponível na distribuição dos campos usada na comparação, como também as probabilidades de erros diferentes que podem acontecer nos registros. O segundo método utiliza informação dentro dos arquivos ou bases a serem relacionadas para estimar as probabilidades $m(\gamma)$ e $u(\gamma)$. Outra metodologia desenvolvida pelo White (1997), considera a aproximação Bayesiana (JENSEN,2004).

Intuitivamente, poderia pensar-se que existem muito mais registros pareados não pares, que os pares. Na FIG. 2.2, observa-se o típico histograma dos pesos dos registros pareados. O modelo não par é maior que o modelo dos pares. O grau de separação entre os modelos está indicando o nível de dificuldade da taxa de relacionamento e valor do erro de tipo I e II que podem resultar.

Figura 2.2 – Histograma dos pesos para comparar no modelo probabilístico, para os pareados e não pareados, e o grau de superposição (onde há uma indefinição)

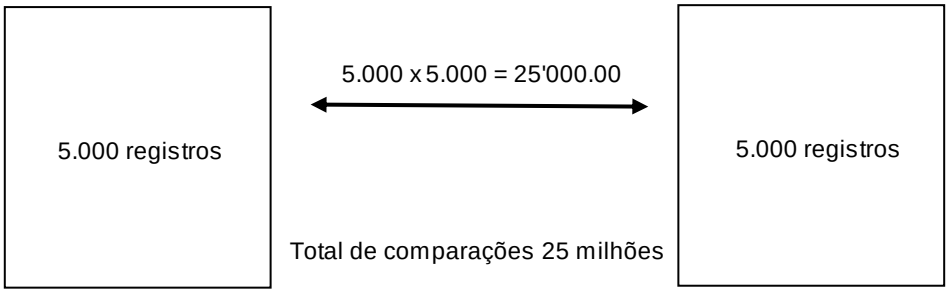


v). Blocagem

Um problema central no relacionamento de registro é que na maioria dos casos os arquivos ou bases de dados que se utilizam são de tamanhos grnades e por conseqüência, a base que contem a combinação dos registros será também de tamanho grande. Assim, quando as

bases de dados a serem relacionadas contêm 5.000 registros cada uma, então 25 milhões de registros de comparação podem ser realizados, parecendo ser impraticável analisar todas as comparações $(\alpha, \beta) \in A \times B$ (Ver FIG 2.3)

FIGURA 2.3 – Total de registros a serem comparados sem considerar a blocagem quando as bases de dados a serem comparadas contém 5.000 registros cada uma (Exemplo hipotético)



Como muitos proceso de relacionamentos de base ded ados envolve volumens grandes de registros, é importante criar subconjunto de registros de comparação, para limitar tempo, orçamento e aumentar a eficiência dos sistemas computacionais. A redução de registros de comparações é determinada pela combinação de registros semelhantes em grupos de comparação (ou blocos). A Blocagem é executada ordenando dois registros sobre um ou mais campos (variáveis) presentes em cada arquivo ou bases de dados. As comparações de registro são restringidas para pares de registros dentro de um determinado bloco, o que diminui o número de comparações de registros a ser feito.

O objetivo da blocagem é permitir que o processo de relacionamento se faça de forma mais otimizada. Por meio deste processo, as bases de dados são logicamente divididas em blocos mutuamente exclusivos, limitando-se as comparações aos registros pertencentes ao mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros (CAMARGO e COELI, 2002a).

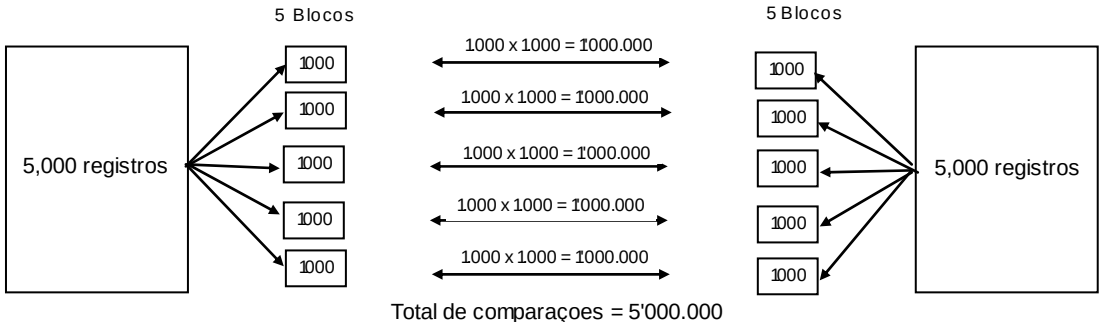
Para prover uma estrutura computacional sólida para comparar os registros dos arquivos ou bases de dados, o número de comparações a ser examinado pode ser restringido a um subespaço de Γ, digamos Γ^*. O subespaço Γ^* pode ser adquirido dividindo o arquivo de comparação em blocos mutuamente exclusivos, utilizando o campo ou variável da base de dados tal como sexo, sobrenome, entre outros. Isto proporciona a realização de comparações explícitas entre registros dentro de cada bloco, e um bloco pode ser criado utilizando qualquer campo ou variável da base de dados. Porém, é melhor utilizar um

campo que é comum em ambas ás bases de dados, que apresente menos erros na sua grafia e que sejam iguais ou, ao menos, bastante semelhantes.

O subespaço Γ^* é então o conjunto de γ para o qual o campo ou variável de blocagem tem o status de concordância, todos os outros γ são implicitamente não pares positivos.

O diagrama da FIG 2.4 ilustra a redução em comparações para o caso onde há cinco blocos de igual tamanho em cada arquivo de dados.

FIGURA 2.4 – Total de registros a serem comparados considerando 5 blocos, quando as bases de dados a serem comparadas contêm 5.000 registros cada uma e cada bloco 1000 registros. (Exemplo hipotético)



Considerando o exemplo apresentado na FIG. 2.1, para o par de registros Maria Souza/Maria Sousa, se a variável sexo fosse utilizado como uma variável de blocagem, os dois registros ainda seriam comparados, mas se o ano de nascimento fosse utilizado como uma variável de blocagem, então elas não seriam comparadas.

Com a aplicação prática do relacionamento de registros, as comparações não são selecionadas aleatoriamente de $A \times B$. Disto concluímos que as probabilidades de erro μ e λ, são interpretadas como as proporções de erro em vez de probabilidades de erro. Assim, é importante notar que um evento particular A_1 ou A_3 não é de preocupação ao determinar as probabilidades de erro, mas a proporção de ocorrências de concordância e discordância para a população total nos permitirá derivar um subconjunto de registros para as comparações.

vi) Valores limiares.

Depois que os pesos forem calculados, o limiar mínimo e o máximo são estabelecidos. O limiar máximo é o peso acima do quais todos os registros pareados são determinados como pares verdadeiros. Nesta região, usualmente, existe um único par de registros relacionado, outros possíveis pares podem ser ignorados ou considerados como registros duplicados. O limiar mínimo é o peso no qual todos os registros pareados são determinados como não pares verdadeiros (ver FIG 2.5).

Depois que a especificação de todas as configurações pertinentes de γ_j^k forem feitas, junto com os pesos associados γ_j^k, valores limiares T_u e T_λ precisam ser fixadas. Em conjunto com estes valores de limiar, a proporção de fracassos necessita ser estimada, permitindo determinar as disposições positivas de comparações a serem realizadas.

O número de configurações de γ_j^k em qualquer comparação será provavelmente muito grande, quando se criar uma inscrição completa e ordenando então, provando configurações dentro de um conjunto de treinamento onde os status de *M* e *U* são conhecidos para poder estimar T_u e T_λ. Isto, porque os vetores de componente γ_j^k são independentes de um ao outro, as configurações da componente $\gamma_1^k, \gamma_2^k,\ldots, \gamma_{jk}^k$ podem ser uma amostra independentemente com probabilidades $z_1^k, z_2^k,\ldots, z_{jk}^k$, então a configuração total $\gamma_j = (\gamma_1^k, \gamma_2^k,\ldots, \gamma_{jk}^k)$ é uma amostra com probabilidade $z_1^k, z_2^k,\ldots, z_{jk}^k$. Assim, não todas as configurações de γ são necessárias para a amostragem, apenas a configuração de γ^k, para cada *k* é suficiente. A amostra pode, então, ser ordenada pelos valores decrescentes de

$$w = w_1 + w_2 + \ldots + w_k$$

Seja γ_h *(h = 1, 2, ...,* S onde S é o número de configurações dentro da amostra*)* o h^{th} elemento da relação ordenada de uma amostra. Então $P[w(\gamma) < w(\gamma_h) \mid \gamma \in M]$ é estimado por

$$\lambda_h = \sum_{h'=h}^{S} \frac{m(\gamma_{h'})}{\pi(\gamma_{h'})}, \text{ onde } \pi(\gamma_h) = \frac{S}{2} \cdot z'(\gamma_h)$$

e

$$z'(\gamma_h) = z_{h_1}^1 \cdot z_{h_2}^2 \cdots z_{h_k}^k + z_{n_1-h_1+1}^1 \cdot z_{n_2-h_2+1}^2 \cdots z_{n_k-h_k+1}^k$$

enquanto

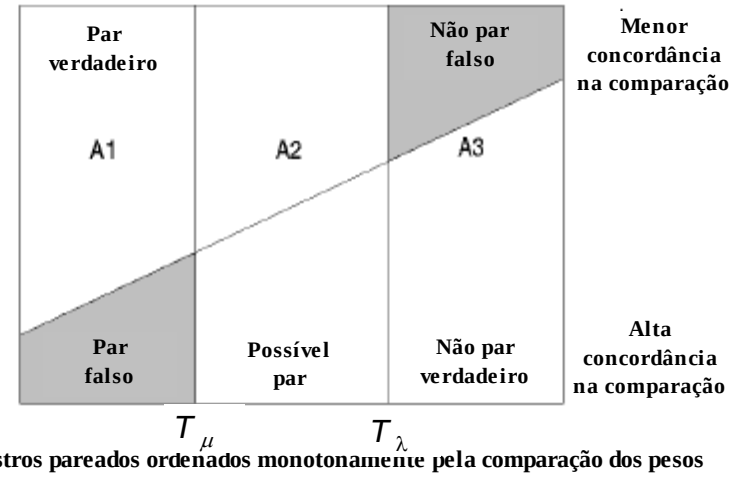
$$P'\left[w(\gamma) < w(\gamma_h) \mid \gamma \in U\right]$$

é estimado por

$$u_h = \sum_{h'=1}^{h} \frac{u(\gamma_{h'})}{\pi(\gamma_{h'})}, \text{ onde } \pi(\gamma_h) = \frac{S}{2} \cdot z'(\gamma_h)$$

Portanto, os valores dos limiares $T(\lambda_h)$ e $T(u_h)$ são simplesmente os pesos $w(\gamma_h)$ e $w(\gamma_h)$. Além disso, estes valores são utilizados como um critério na determinação da classificação de cada registro.

Figura 2.5 – As três regiões do modelo de probabilidade.



2.3.3 Vantagens dos programas computacionais para o relacionamento.

Nos últimos anos, o avanço da tecnologia computacional conduziu a melhorias na metodologia e eficiência do relacionamento probabilístico (JENSEN, 2004). Nesse contexto, temos o Algoritmo de Maximização – Expectativa, mais conhecido como o algoritmo EM, que tem como objetivo calcular as distribuições de probabilidade fundamentais para cada variável ou campo, e que foi apresentado por Winkler (1989, 1990, 1993?, 1994), ajudando na simplificação do processo de estimação.

A metodologia geral de Fellegi e Sunter (1969) especificamente não foi ajustada para registros pareados únicos. Jaro (1989) modificou esta metodologia para obter pareamentos de registros únicos, desenvolvendo um algoritmo para a comparação de campos "caracteres", levando em conta a inserção, extração, troca e transposição aleatória de caracteres. A Metodologia descrita por Jaro (1989) foi implementada no *software AUTOMATCH***,** que requer especificar: variáveis de blocagem que ajudam a reduzir o número de comparações a serem realizadas de fato; variáveis de comparação, cujos valores podem ser comparados por pares de registro; valores de inicial das probabilidades de *m* e *u* para cada um dos identificadores considerados; e os limiares sobre os pesos que determinam as três categorias de decisões A_1; A_2 e A_3. Nos últimos anos, nos países onde freqüentemente utiliza-se esse procedimento para combinar informações de diferentes fontes de dados, programas de *software* de computador executando as rotinas de relacionamento de registro foram desenvolvidos. Entre os outros programas desenvolvidos a partir do *AUTOMACH*, podem-se mencionar os seguintes: *Febrl* Free (relacionamento de registros biomédicos livremente extensível), *Link Plus Free* (Relacionamento de dados para prevenção e controle de doenças), *SimMetrics Free* (proporciona uma configuração do relacionamento de dados, com base aos campos e gráficos de similaridades), *The Link King* (programa desenvolvido a partir de um algoritmo determinístico, para os serviços administrativos de abuso de sustâncias proibidas e saúde mental)e D-Dupe (Integração e exploração de dados duplicados).

No caso do Brasil, o algoritmo desenvolvido por Jaro (1989) foi resultado de uma tentativa de inseri-lo como parte de um *software* para o relacionamento probabilístico de registros (*AutoStan-AutoMatch; MatchWare Technologies, Inc*.), contudo, seu custo foi considerado infactível e portanto a utilização deste programa não foi possível (Camargo e Coeli, 2000; Machado, 2002). Como solução a esses inconvenientes e dificuldades, Camargo e Coeli (2000) desenvolveram um *software*, denominado ***RecLink*****,** em uma linguagem C++ com o ambiente de programação *Borland* C++ *Builder* versão 3.0 (*Borland International Inc*., 1998a; Reisdorph, 1998). Este *software* corresponde ao sistema de relacionamento de bases de dados fundado na técnica de relacionamento probabilístico de registros segundo a teoria de Fellegi e Sunter (1969)

2.4 Evidências do relacionamento de bases de dados

O relacionamento de base de dados é uma ferramenta utilizada em muitos países do mundo, sobretudo nos mais desenvolvidos. Entre os tipos de relacionamento utilizados, há: o relacionamento das mesmas pessoas em uma única base de dados, para criar históricos de saúde; o relacionamento de dados de coortes; o relacionamento de dados de referências geográficas para adicionar novas informações na cartografia; o relacionamento como parte de um ambiente do sistema operacional (exemplo de registros de câncer); e o relacionamento para adicionar variáveis analíticas (FAIR, 1999). Além disso, há trabalhos com dados provenientes de diferentes fontes de dados e que podem corresponder a estatísticas vitais, censos, dados administrativos e *surveys*, com o objetivo de melhorar a qualidade e consistências dos dados, preparar registros específicos para estudar doenças, acompanhar coortes para determinar *status* vitais do indivíduo e atividades habituais, construir estruturas de amostragem, e estudar histórias genealógicas ou históricas. (SMITH, 1985; GOLDACRE, 1987; GILL e BALDWIN, 1987; JENSEN, 2004).

O relacionamento de dados tem sido extensivamente utilizado em vários países na área da saúde, especificamente nos estudos epidemiológicos (como a saúde infantil e Neoplasias) (Machado, 2002). Esta metodologia tem aplicação também em outras áreas, como no *marketing*, relacionando dados para administrar a fidelidade dos clientes de um produto no mercado, descoberta de fraude e data *warehousing*. As entidades do governo também utilizam o relacionamento de dados para executar leis, regulamentos e políticas. Todas estas aplicações podem ser classificadas como “administrativas”, porque o relacionamento é utilizado para fazer decisões e ações relacionadas com uma entidade individual (FAIR, 1999?).

Nas últimas décadas o Canadá e os Estados Unidos têm fomentado um sistema similar de acompanhamento das causas de mortalidade, utilizando o relacionamento de dados. Países escandinavos, como a Noruega, Suécia, Dinamarca, e Finlândia, também têm aproveitado a utilização de procedimentos de relacionamentos de bases de dados para o acompanhamento de indivíduos durante vários anos com vários objetivos específicos, tomando como base o número de identificação individual emitido aos residentes no momento do nascimento. No Reino Unido, durante vários anos, atividades semelhantes utilizaram um número de serviços para a saúde nacional como identificador, com o

objetivo de pesquisar a incidência de câncer e a mortalidade. Além disso, há mais de dez anos foram desenvolvidos sistemas de relacionamento de dados complexos, e/ou em alguns casos estão sendo desenvolvidas ferramentas computacionais de relacionamento de dados em países como a Austrália, França, Índia, Israel, Japão, e a antiga União Soviética (M. Carpenter, Estatísticas da Canadá, Ottawa, Ontario, pessoal de comunicação da Canadá, 1998)[11]

2.4.1 Evidências do relacionamento de bases de dados aplicadas no Brasil.

No caso brasileiro, os trabalhos de relacionamento de dados probabilísticos (e alguns determinísticos) foram realizados na área da saúde. O relacionamento probabilístico foi especialmente aproveitado nos estudos da mortalidade infantil (MACHADO, 2002).

Dentre os primeiros trabalhos que se discute o relacionamentos de dados "*record linkage*" com funções automáticas, aponta-se o de Noronha et *al* (1997), em que é feita uma comparação entre os sistemas de informações de mortalidade e de nascidos vivos para analisar o grau de concordância do preenchimento de dados comuns a eles e recuperação de informações. Para isto utilizou-se um relacionamento automático e determinísticos dos registros de nascimentos pertencentes à coorte de nascidos de 1998 e cujas mães residiam no município de Rio de Janeiro (MACHADO, 2002).

Almeida e Jorge (1996) relacionaram as informações do SIM e do SINASC, para estudo de mortalidade neonatal, com possibilidade de determinação de medidas de risco para os nascidos vivos. Este estudo foi realizado no município de Santo André, Região Metropolitana de São Paulo, Brasil.

Fernandes (1997) fez um relacionamento de informações sobre óbitos e nascimentos, partindo, inicialmente, da localização manual dos registros. Neste trabalho considerou os nascidos em 1989, 1990 e 1991 de Brasília-DF, comparandor o nome da mãe em ambas as bases de dados, de forma manual.

---

[11] No ANEXO I são apresentados alguns trabalhos que tratam do relacionamento de bases de dados nos paises desenvolvidos.

Carvalho e Mello et al (1998), com o objetivo de analisar a sobrevida em pacientes hospitalizados por Acidentes Vasculares Encefálicos (AVE), realizaram o relacionamento entre a base contendo os 6531 casos de AVE identificados na base de dados dos formulários AIH e os bancos das DO de 1998 (110.820 óbitos, por todas as causas) e de 1999 (105.644 óbitos, por todas as causas). O método probabilístico foi escolhido por não ter um campo identificador unívoco entre os bancos de referência (como por exemplo, o campo CPF, normalmente não preenchido), o que não possibilita a busca direta pelo caso. Os campos utilizados para o relacionamento foram nomes, data de nascimento e sexo.

Bohland (2003) utilizou em seu estudo as informações do SIM, SINASC, SIH e Sistema de Informação da Atenção Básica para melhorar a qualidade da informação sobre óbitos de mulheres em idade reprodutiva.

Os trabalhos anteriormente mencionados foram feitos utilizando um relacionamento exato. Outros trabalhos relacionando informações de registros entre os registros de mortes e os de nascimento também foram realizados, mas neste caso por meio do relacionamento probabilístico.

Machado (2002) utilizou o relacionamento probabilístico de registros das bases de dados de SIM e SINASC para estudo da morbi-mortalidade infantil. No estudo identificou todos os nascimentos da cidade de São Paulo durante 1998, extraindo 209.628 registros de nascimento. Depois de ter a informação combinada, Machado fez uso da regressão logística multivariada para ajustar o efeito de cada variável independente sobre o escore de Apgar indicando: menos de sete a um minuto e menos de sete a cinco minutos**.**

Coeli et *al*. (2003) utilizaram o relacionamento probabilístico para obter a concordância entre a informação de internação hospitalar obtida por inquérito domiciliar e o registro hospitalar da internação mencionada. Este estudo contou com um total de 2.288 entrevistas domiciliares que foram realizadas em Duque de Caxias, Rio de Janeiro. As informações sobre a ocorrência de ao menos uma hospitalização durante o ano que precedeu a entrevista foi obtida de um total de 10.733 moradores. Os 130 registros de moradores que relataram ao menos uma hospitalização na rede pública foram relacionados a uma base de dados hospitalares contendo 801.587 registros.

Por último, um trabalho em que se utilizou o processo de relacionamento de base de dados é de Miranda-Ribeiro (2007), trabalho este que utiliza o processo de relacionamento para a

reconstrução de história de nascimentos, com o objetivo de tornar completa a história de nascimentos das mulheres entre 15 e 64 anos de idade, para os quinze anos anteriores ao censo ou pesquisa. Especificamente o relacionamento, consiste em buscar, no universo de histórias de nascimentos completos, aquela que mais se aproxima da história de nascimentos parcial, com base na comparação de algumas variáveis (MIRANDA-RIBEIRO, 2007).

2.5. Dados de pesquisa de campo e registros administrativos

As informações aceitas como o resultado do processamento, manipulação e organização dos dados, podem ser coletadas por vários métodos, tais como entrevistas, questionários, observações ou revisão de registros administrativos, cada um dos quais apresenta vantagens e desvantagens. Não raro, essas formas de coleta de dados, complementam-se com o objetivo de ajudar a assegurar uma pesquisa completa (Floridi, 2005). É nessa perspectiva de relacionamento de dados que este trabalho se realiza e partirá de diferentes fontes de informação: uma pesquisa de campo de domicílios e outra dos registros administrativos. Nesta seção ressalta-se a importância da informação das pesquisas de campo e dos registros administrativos.

2.5.1. Informações das Pesquisas de Campo.

Uma pesquisa de campo é aquela utilizada com o objetivo de conseguir informações e/ou conhecimentos acerca de um problema para o qual se procura uma resposta, ou de uma hipótese que se queira comprovar, ou ainda, descobrir novos fenômenos ou as relações entre eles (MARCONI e LAKATOS, 2003). O que caracteriza esta pesquisa como uma pesquisa de campo é, principalmente, o levantamento no campo das percepções das pessoas, usuários ou operadoras sobre os temas que se está pesquisando.

Freqüentemente as ciências e áreas de estudo que utilizam informações de pesquisa de campo para o estudo de indivíduos, grupos, comunidades, instituições, têm como objetivo compreender os mais diferentes aspectos de uma determinada realidade ou, em alguns casos, visam diagnosticar e formular políticas públicas (RAMOS e SANTANA, 2002). Além disso, as pesquisas de campo exigem determinadas técnicas de coleta de dados mais

apropriadas à natureza do tema e, ainda, à definição das técnicas que serão empregadas para o registro e análise. Dependendo das técnicas de coleta, análise e interpretação dos dados, a pesquisa de campo poderá ser classificada como quantitativa (descritiva) ou qualitativa (RICHARDSON, 1999). As informações de pesquisas de campo quantitativas caracterizam-se pelo processo de quantificação, tanto no processo de coleta de informações, como no tratamento destas por meio de técnicas de estatísticas e procedimentos matemáticos; enquanto as qualitativas diferem do quantitativo na medida em que não emprega, necessariamente, um instrumental estatístico como base no processo de análise de um problema (CERVO e BERVIAN, 2002).

As informações das pesquisas de campo são coletadas seguindo um conjunto de regras, que dependem do método de coleta eleito. Entre os métodos mais importantes estão: entrevistas, questionários e observações, nas quais o analista obtém e desenvolve um sistema de informação para atingir suas metas e objetivos. Independente do método de coleta escolhido os analistas ou pesquisadores devem demonstrar e desenvolver conhecimento e manifestar a sua honestidade, imparcialidade, habilidade, objetividade, controle, comunicação compressão e amabilidade para conseguir informações adequadas (ALFONSO, 2001).

Uma das técnicas mais utilizadas nas pesquisas de campo são as entrevistas estruturadas e individuais. Estas adotam, como critério básico, que a coleta de dados se baseie na auto-declaração dos indivíduos da população alvo, desta forma a coleta de informações permite um padrão estruturado na pesquisa. Como estas informações podem ser influenciadas por aspetos subjetivos, como opiniões ou atitudes, as entrevistas devem ser realizadas em um ambiente que facilite a conversação. Além disso, como a declaração dos entrevistados é de suma importância nas informações que serão analisadas, deve evitar-se adiantar ou sugerir as respostas às questões formuladas durante a entrevista (Bartholomew, 1961). Neste método de coleta de dados, faz-se necessário tomar cuidados especiais: as pesquisas de campo devem considerar a estrutura geral, não negligenciar os erros de não respostas parciais ou globais, que deverão ser controlados por uma adequada qualificação e supervisão dos entrevistadores; garantir o anonimato; motivar o respondente a cooperar; e iniciar o questionário com questões interessantes e pouco controversas.

No caso brasileiro, as informações que provêm de pesquisas de campos, e que são de grandes repercussões, pertencem às pesquisas de campo realizadas pelo Instituto Brasileiro

de Geografia e Estatística (IBGE), cujos objetivos estão relacionados à avaliação das condições e situações na qual a população brasileira desenvolve-se. Entre as pesquisas de maior importância realizada pelo IBGE e que estão vigentes desde inicio dos anos 70 e 80 tem-se, a Pesquisa Mensal de Emprego – PME (produz indicadores mensais de trabalho sobre a condição de atividade da população); Pesquisa Nacional de Saneamento Básico – PNSB; (oferta e qualidade dos serviços de saneamento básico no país); Pesquisa Nacional por Amostra de Domicílios – PNAD (Informação anual sobre características demográficas e socioeconômicas da população); Pesquisa da Pecuária Municipal – PPM (informação sobre efetivo das espécies animais criadas e dos produtos da pecuária); Pesquisa de Assistência Médico-Sanitária – PAMS (oferta de serviços de saúde e as condições de assistência médico-sanitária); Pesquisa Industrial Anual - Empresa e Produto – PIA (informações econômico-financeiras sobre o setor industrial brasileiro). INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE, 2007?).

Existem também outras instituições, tais como Universidades, Centros de pesquisas privadas e ONGS, que obtêm importantes e valiosas informações provenientes de pesquisa de campo. A utilização deste tipo de fonte são os meios mais diretos, sendo em alguns casos única forma de obter informações. Dessa forma, as pesquisas de campo requerem, cada vez mais, um forte rigor metodológico para obter de forma adequada a informação, sobretudo nos trabalhos cujo objetivo é estudar as melhorias das condições de vida e a eficácia das políticas públicas implementadas (FOWLER, 1996).

Uma das informações que serão utilizadas neste trabalho, pertence à Avaliação de Impacto do Programa Bolsa Família (AIBF), que conduziu uma pesquisa de campo para coletar os dados dos domicílios necessários e úteis, e foi realizada por meio de entrevista semi-estruturadas aos integrantes das famílias alvo.

2.5.2. Informação dos Registros Administrativos.

As informações dos registros administrativos são resultados das necessidades sociais, fiscais, tributárias ou outras, criadas com o objetivo de viabilizar a administração ou operacionalização dos programas de governo, ou mesmo para fiscalizar e controlar a execução de obrigações legais por parte de determinados segmentos da sociedade (CEPAL, 2003a).

As características prioritariamente desejáveis dos registros administrativos para uma adequada utilização são: a) determinar a cobertura e alcance dos registros; b) utilizar unidades estatísticas uniformes que garantam a consistência temporal dos resultados, evitando duplicações e omissões nos registros dos dados; c) designar um número único de identificação do informante, que normalmente é um código legal designado pelo órgão que administra os registros, ampliando a capacidade de articulação entre as diversas fontes, inclusive as administrativas; d) determinar adequadamente as variáveis pesquisadas e seus respectivos conceitos e definições, como também a qualidade com que os dados são respondidos e processados; e e) definir a freqüência com a qual os dados estão disponíveis para a fins estatísticos (CEPAL 2003b; TINTÓ, 2004?).

Entre as principais vantagens da utilização de informação dos registros administrativos com propósitos estatísticos, tem-se: a) obtenção de dados a custo baixo; b) contribuição para reduzir o trabalho de preenchimento de formulários para os informantes; c) evitar a duplicação de esforços nas instituições informantes, porque os mesmos dados podem ter sido informados a outros órgãos do estado; d) garantia de uma cobertura completa da população alvo; e) não contêm erro de amostragem, manipulam menores volumes de erros de não-resposta e permitem separação específicas de sub-população, tais como nível geográfico, tamanho, atividades econômica entre outros; e f) a qualidade da informação pode aumentar consideravelmente ao acesso de informações atualizadas para a utilização dos diretores das instituições responsáveis dos registros administrativos (CEPAL, 2003b).

Embora existam várias motivos para se empregar as informações dos registros administrativos, estas nem sempre são fáceis de utilizar ou acessar, e apresentam um conjunto de desvantagens com fins estáticos, tais como: a) falta de correspondência nas definições das unidades entre os sistemas administrativos e as áreas ou agências estatísticas que obrigam a realizar um processo de conversão de unidades administrativas a unidades estatísticas; b) diferenças nas definições das variáveis; c) utilização de diferentes classificações que acarretam a construção de tabelas de conversões para transformar os códigos da classificação administrativa em informações utilizadas pelas instituições ou áreas estatísticas; d) disponibilidade temporal de dados e períodos de referências não coincidentes; e) os registros administrativos do setor público podem ser influenciados por mudanças de aspectos políticos; f) o órgão ou área estatístico responsável deve realizar uma conciliação dos dados, o que facilita se houver algum número identificador comum

nos registros; e g) inconsistência de dados de diferentes fontes que obriga a estabelecer regras de prioridade de sua utilização (CEPAL 2003a).

Embora muitos registros administrativos tenham como objetivo ser fontes de informação estatísticas contínuas, este caso nem sempre pode ser observado. Existem, registros administrativos que não são capazes de ser uma base de dados para análises econômicas ou sociais. No entanto, pela quantidade de informação que possuem, podem se converter em fontes de informação estatística para o qual precisam passar por um tratamento ou trabalho de aprimoramento, para descartar incoerências, contradições, contornar mudanças na cobertura, etc. (RAMOS e SANTANA, 2002).

Portanto, considerando vantagens e desvantagens no manuseio das informações dos registros administrativos, estas têm sido importantes fontes complementares para as pesquisas de campo na elaboração de sistemas de informações estatísticas, em especial na elaboração e manutenção dos responsáveis das áreas ou oficinas estatísticas. No entanto, precisam-se avaliar aspectos relevantes das informações dos registros administrativos e das pesquisas de campo relacionados à qualidade, cobertura, definição de conceitos, metodológicos, classificações e variáveis pesquisadas, entre outros, antes que estes tipos de informações sejam adotados como fontes principais.

No caso do Brasil, são exemplos mais usuais de Registros Administrativos: a RAIS (Relação Anual de Informações Sociais), o CAGED (Cadastro Geral de Empregados e Desempregados), o banco de dados do SUS (Sistema Único de Saúde), o banco de dados do Seguro-Desemprego, o Censo Escolar, entre outros (RAMOS e SANTANA, 2002).

Os registros administrativos que serão utilizados neste trabalho, correspondem ao Cadastro Único de Beneficiários dos Programas Sociais do Governo Federal (CadÚnico), que é uma ferramenta utilizada pelo Governo Federal para identificar os potenciais beneficiários dos programas sociais Bolsa Família, Agente Jovem, Programa de Erradicação do Trabalho Infantil (PETI), Tarifa Social de Energia Elétrica e outros. Esta informação também é utilizada por vários estados e municípios para identificação do público-alvo dos seus programas (BRASIL, 200-?c).

2.5.3. Integração de informações de duas fontes de dados diferentes.

Um dos objetivos deste trabalho é organizar uma base de dados que contenha informações anexadas de duas fontes de informação: da pesquisa de campo AIBF e dos registros administrativos CadÚnico. A base de dados organizada com os dois tipos de informações será utilizada na análise que será apresentada ao longo deste trabalho. Com base neste objetivo, impôs-se um crivo rigoroso na interpretação e correção de problemas de códigos, inconsistências próprias de ambas as bases de dados utilizadas, procurando gerar uma base de dados organizada e o mais consistente possível.

Como exposto anteriormente, volta-se a enfatizar que, quando se utiliza bases de dados de diferentes fontes, deve-se fazer frente aos problemas de conciliação de dados. Se existe algum número identificador único comum dos registros, o problema será facilitado, mas, caso contrário, ao buscar relacionar os dados há que se levar em consideração outras variáveis, tais como nome, sexo, data de nascimento, código de município, entre outros. Neste último caso, é provável contar com um conjunto de erros no relacionamento que deve ser previsto na ocasião da analise da base de dados organizada com ambas as informações.

Outro problema que freqüentemente é encontrado quando se utiliza múltiplas fontes de dados, é a consistência entre os dados. Isto porque os dados de uma fonte podem contradizer os da outra, devido a diferentes definições, classificações ou, inclusive, por erros em uma das fontes. Para resolver tais problemas, é necessário estabelecer regras de prioridade, definindo qual fonte é mais confiável para cada variável. Uma vez estabelecidas às hierarquias das fontes de dados de maior primazia para cada variável, é possível assegurar-se de que um dado de uma fonte de maior prioridade não será substituído por outro de menor prioridade.

Uma última interrogante que se apresenta quando utilizamos mais de uma fonte de dados é definir qual delas tem maior gradação de qualidade. Para essa questão não existe uma resposta simples, muitas medidas em conjunto podem ser aplicadas para respondê-la. Assim, entre as fontes de informação pode-se comparar a cobertura e precisão das variáveis, preferencialmente por meio de um tipo de processo de qualidade, para estabelecer os valores corretos de determinada variável. Embora existam muitas discussões sobre a decisão de qual é a melhor fonte de informação, o mais importante é destacar ou

aproveitar a melhor informação que cada fonte de dados possa dispor, tanto da pesquisa de campo quanto dos registros administrativos. Dessa forma, poder-se-á contar com a maior informação útil para responder aos objetivos dos estudos que precisam conter informações complementares de ambas as fontes de dados.

A importância que têm as informações combinadas de duas fontes de dados é diversa. No caso das informações do Cadúnico e da pesquisa de campo AIBF que serão utilizadas neste trabalho é possível aumentar consideravelmente a qualidade e quantidade de informação estatística. Assim, o trabalho de relacionar essas bases de dados proporciona um instrumento de coordenação e harmonização das diferentes fontes de dados utilizadas que permite contar com um marco ótimo para as pesquisas futuras dirigidas para domicílios, contando com informação demográfica de tipo longitudinal e oferecendo atualização e manutenção contínua de informações familiares, bem como para designar códigos fixos a cada domicílio, que sejam unívocos e de boa qualidade.

Neste trabalho o relacionamento de base de dados, em primeiro lugar, teve como objetivo recuperar o número de NIS para as pessoas que não contam com essa informação na pesquisa de campo; em segundo lugar, alocar as famílias nos grupos de comparação segundo os registros administrativos. Como conseqüência das informações que foram recuperadas, pode-se contar com uma grande base de dados com informação completa de ambas as fontes de informação.

O exercício que possibilita os resultados deste relacionamento refere-se também a obtenção de outros resultados importantes ou a aplicação de outras técnicas não-experimentais, ambas úteis para avaliar o impacto dos programas sociais. Entre os exercícios possíveis de realizar tem-se:

1. Contrastar a distribuição da renda dos registros administrativos dos programas sociais com a informação de renda obtida na pesquisa de campo para avaliar o impacto destes programas, aplicação que permite avaliar em certa forma o grau de focalização dos programas e seus efeitos sobre a desigualdade de renda.

2. Realizar exercícios iniciais sobre a obtenção de uma variável de controle sobre o tempo de exposição dos beneficiários do programa (efeito duração) ou sobre o valor do benefício recebido durante a totalidade do período (efeito dose), porque o relacionamento permitirá obter algumas variáveis utilizadas como "*proxys*" para analisar estes efeitos.

3. Utilizar uma forma particular de identificar os grupos potencialmente beneficiários e não-beneficiários dos programas sociais, para avaliar o impacto potencial entre os beneficiários da variação de algumas remunerações básicas determinadas (por exemplo, restringir a amostra a famílias beneficiárias com renda per-capita entre 40 e 60 reais e avaliar o impacto).

Considerando o item 3, o relacionamento de dados nos possibilita aplicar a técnica de Regressão Descontínua *Sharp* (RD) que utiliza as descontinuidades no processo de alocação ao programa para identificar o efeito causal, e supõe que uma variável contínua pré-tratamento (Z) influi nas variáveis resultados (Y), assim como na variável que define a participação no programa (D), que, por sua vez, afeta o resultado Y. Assim, Z tem um impacto direto em Y e um efeito indireto através de D (Thistlethwaite e Campbell, 1960). A luz do exposto, a aplicação desta técnica somente é possível quando se utiliza a variável continua (Z) que para este estudo é a "renda familiar" dos registros administrativos do CadÚnico, onde se pressupõe é pré-tratamento e não está influenciada pela renda que recebem os beneficiários, mas que influiria nos resultados de impacto do PBF e na participação das famílias beneficiárias desse programa.

3 AVALIAÇÃO DO IMPACTO E OS PROGRAMAS SOCIAIS

Neste capítulo, são discutidas diversas questões sobre os programas sociais que objetivam aliviar ou combater a pobreza, como também a realização de processos de avaliação deste programas e especificamente do Programa Bolsa Família, analisando as metodologias de avaliação e o conjunto de ações e etapas indispensáveis para avaliar adequadamente os impactos dos programas. Na seção inicial, explica-se sobre alguns enfoques que definem o teórico e o metodológico da avaliação de impacto dos programas sociais. Em seguida, descreve-se as etapas indispensáveis para uma adequada avaliação de impacto; definindo os métodos a serem utilizados, as quais dependem dos tipos de experimentos ou desenhos metodológicos para estimar o impacto e que variam na forma e critério utilizado na construção do contrafactual. Enfatizam-se as técnicas dos desenhos não-experimentais que serão utilizando na tese: método de pareamento[12] (matching) e regressão descontínua. Na seção seguinte, abordam-se os programas de transferência condicionada de renda no Brasil e, especialmente, o Programa Bolsa Família, descrevendo o desenvolvimento, cobertura e eficiência adquirida nos últimos anos no Brasil. Em seguida à apresentação do Programa Bolsa Família realiza-se uma revisão breve do desenvolvimento da pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF), ressaltando a implementação e o método utilizado da avaliação, como também, alguns resultados importantes conseguidos. São apresentados alguns estudos empíricos de avaliação de impacto dos programas de transferências condicionadas de renda (TCR) na América Latina, esboçando resumidamente o que está relacionado ao benefício oferecido pelo programa e à cobertura deste, implementação da avaliação, método de avaliação de impacto do programa e alguns resultados da avaliação de impacto. Finalmente, apresenta-se a estratégia alternativa para alocar às famílias nos grupos de comparação segundo o registro administrativo do Cadastro Único, com base na utilização do relacionamento de bases de dados da pesquisa de campo AIBF com o CadÚnico

[12] Neste caso e para a analise da avaliação de impacto o termo "pareamento" será utilizado para referir-se à técnica de Pareamento (ou Matching) por Escore de propensão (PSM), que compara resultados de famílias similares do grupo de tratamento com as do grupo de comparação ou controle.

3.1. Avaliação de impacto.

Nos últimos anos, diversos enfoques que definem a teoria e metodologia de uma avaliação têm sido apresentados. De forma geral o termo de avaliação pode ser definido como una atividade gerencial interna ou externa que tem como propósito assegurar a pertinência do desenho de um programa, por meio dos métodos de implementação que atingem tanto objetivos específicos como gerais (COHEN et. al, 2001)

Segundo Cohen e Franco (1988), o termo de avaliação de impacto define-se como um processo orientado a determinar, sistemática e objetivamente, a eficiência e eficácia dos impactos das atividades realizadas tratando à avaliação como um processo organizativo para melhorar as atividades ainda em andamento e ajudar a administrar o planejamento, programação e decisões futuras.

Segundo o Banco Mundial (2003a), a avaliação de impacto é a mensuração das mudanças no bem-estar dos indivíduos que podem ser atribuídas a um programa ou a uma política específica. Seu propósito geral é determinar a efetividade das políticas, programas ou projetos executados (PATTON, 2002). Tsl como outras técnicas de avaliação acumulativas, a avaliação de impacto pode ser utilizada para determinar até que ponto os resultados planejados foram produzidos ou atingidos, assim como para melhorar outros projetos ou programas em andamento ou futuros (BROUSSEAU e MONTALVÁN, 2007?).

Na atualidade, a avaliação de impacto é uma das técnicas de resultados mais utilizadas na valoração dos efeitos das intervenções sociais, especialmente os de médio e longo prazo. Neste contexto, os países da América Latina têm gerado um grande interesse por incorporar a avaliação de impacto como uma ferramenta complementar aos métodos de avaliação financeira, econômica e social (com base na analise custo-benefício) que têm sido utilizados nas últimas décadas pelos sistemas de investimento público na região (NAVARRO, 2005).

A avaliação de impacto mede a magnitude das mudanças geradas e sua causalidade com os componentes e benefícios outorgados pelas intervenções (estudo de causalidade). Diante desta situação, a avaliação, as políticas, e os programas correspondem às causas, e seus efeitos são todas as mudanças nas condições dos beneficiários (no curto, médio e longo

prazo), medidos como as mudanças em determinadas variáveis de impacto (ou variáveis de resultado) que são atribuíveis à intervenção (HECKAM e VYTLACIL, 2005).

A avaliação de impacto para medir os efeitos dos programas sociais é uma tarefa complexa, em grande parte devido à presença de fatores externos às intervenções que influem nas condições de vida dos beneficiários, o qual torna difícil a valoração das transformações que são exclusivamente gerados pela intervenção. Estes fatores externos podem-se classificar em observáveis e não observáveis. Os primeiros estão relacionados com as características individuais dos beneficiários (idade, sexo, educação, estado civil, renda, entte outras), com as características de suas famílias (número de membros, renda per capita domiciliar, taxa de participação do trabalho, entre outras) ou com a comunidade (infra-estrutura social, crescimento econômico, capital social, entre outras). Em relação aos fatores não observáveis, estes associam-se especialmente com os valores morais, motivações, interesses pessoais, entre outros (RAVALLION, 1999).

Neste sentido, ressalta-se a importância de tratar a avaliação de impacto como um processo e como algo que faz parte da gestão de um programa ou política social, e não como algo isolado. Desta forma o desenho, o modelo e as variáveis utilizadas para uma avaliação tornam-se pilares importantes para obter resultados de avaliação robustos (NAVARRO, 2005).

3.2. Metodologia de avaliação do programas sociais.

A complexidade das intervenções dos programas sociais e a variedade de ferramentas de pesquisa disponíveis fazem com que não exista uma estratégia única e predefinida da avaliação do impacto. Desta forma, o trabalho da avaliação pode ser considerado flexível para combinar diferentes instrumentos que estão relacionados com as características e o contexto da intervenção (NAVARRO, 2005).

Quando as estratégias para avaliar o impacto do programas sociais apontam para a mensuração dos efeitos da intervenção do programa, os métodos de avaliação utilizados geralmente são quantitativos, considerando duas características principais: verificação da hipótese e comparação. A verificação da hipótese especifica o método dedutivo da pesquisa, na qual se formula a hipótese a partir da teoria do programa, e se coleta a informação para determinar se aceitamos ou rejeitamos a hipótese. A comparação

determina a causalidade entre a intervenção e as mudanças experimentadas pelos beneficiários. Este método de avaliação é aquele que determina a causalidade através da construção de um cenário contrafactual (MOHR, 1999).

3.2.1. Etapas da avaliação de impacto

A) Método de avaliação

Uma das primeiras etapas a ser considerada na avaliação de impacto é definir o método a ser utilizado no trabalho de avaliação. No método de avaliação quantitativa a variável independente considerada é o programa de intervenção e a variável dependente é o efeito ou variável de impacto. A relação entre as duas variáveis pode ser positiva ou negativa. Na avaliação de impacto existem diferentes metodologias para se utilizar, as quais dependem dos tipos de experimentos ou desenhos metodológicos para estimar o impacto e que variam na forma e os critérios que se utilizam para construir o contrafactual.

B) Identificação dos efeitos do programa

Depois de ter definido o desenho de avaliação, a tarefa seguinte é identificar os diferentes efeitos do programa que serão avaliados. Esta identificação é importante porque, por meio desta, minimiza-se o risco de não considerar na avaliação algum efeito relevante do programa. Entre os principais recursos disponíveis para determinar os efeitos do programa destacam-se, o conhecimento dos objetivos e teoria do programa e o diagnóstico do problema.

Na avaliação dos programas o diagnóstico é a descrição da situação atual do problema que se deseja avaliar, fornecendo informação referente a quatro aspectos básicos: (1) a magnitude e severidade do problema, (2) efeitos sobre os diferentes grupos da população interveniente, (3) possíveis causas e (4) as conseqüências deste (NAVARRO, 2005). O insumo principal na elaboração do diagnóstico dos problemas sociais é a informação sobre as condições de vida dos beneficiários; que pode ser quantitativa ou qualitativa. Tanto os métodos quantitativos como qualitativos têm vantagens e desvantagens e, sua aplicação depende, entre outros, do tipo de problema que se analisa, do custo e da disponibilidade da informação. Assim, quando os programas sociais estão orientados à diminuição da

pobreza, o consenso mais amplo acerca das vantagens é combinar informação quantitativa e qualitativa nos diagnósticos da pobreza (RAVALLION, 2002; WHITE; 2002).

C) Seleção das variáveis de impacto.

Quando se precisa avaliar os efeitos dos programas sociais, estes devem ter um significado claro e serem operacionalizados por meio de variáveis ou indicadores que permitam sua valorização. Estas características são de especial importância nos programas de redução da pobreza, nos quais se definem os possíveis efeitos das intervenções através de conceitos abstratos que podem ter mais de um significado, dependendo da perspectiva teórica e o contexto em que são utilizados.

Vaus (1986) propõe os seguintes passos para converter conceitos abstratos e de difícil compreensão a um conjunto de variáveis e indicadores que permitam avaliar os efeitos da intervenção do programa: primeiro, formulam-se os possíveis efeitos do programa; segundo, são identificados os diferentes aspectos ou dimensões que conformam o efeito a avaliar; e terceiro, selecionam-se as variáveis de impacto para as dimensões que serão avaliadas.

Após identificar as dimensões dos possíveis efeitos do programa, o passo final é a seleção de variáveis que permitem a mensuração de impacto da intervenção. Como a pobreza apresenta várias dimensões, o impacto tem que ser analisado através de distintas variáveis: níveis de renda e consumo dos domicílios, indicadores sociais e indicadores de vulnerabilidade (BANCO MUNDIAL, 2003b). Igualmente, a partir destas variáveis e indicadores, constroem-se um conjunto de índices de pobreza com dois propósitos: a) definir quando um domicílio é pobre e obter uma medida agregada que indique a magnitude da pobreza em um determinado grupo de indivíduos ou domicílios e b) comparar diferentes dimensões da pobreza.

D) Coleta de informação.

Esta é uma etapa relacionada estreitamente com a qualidade dos resultados que se espera encontrar. Nas avaliações dos programas sociais que estão relacionados com a pobreza, esta atividade é particularmente mais complexa, pelas diversas características que os beneficiários apresentam. Além disso, muitas das avaliações deparam com restrições de tempo e orçamento, fatores que influenciam no planejamento da coleta da informação.

Outra característica importante nas avaliações de impacto é que estas requerem que se coletem o mesmo tipo de informação para os beneficiários e não beneficiários (grupos de tratamento e comparação). Assim, os instrumentos de coleta devem ser instrumentos padronizados e que permitam pré-estabelecer categorias da classificação (WEISS, 1998). Estes instrumentos de coleta de informação podem ser entrevistas estruturadas, questionários, registros administrativos, entre outros. A informação coletada através destes meios permite, e facilita, a transformação numérica da informação. No entanto, inconvenientes encontrados na avaliação de impato, referem-se principalmente as limitações de uma baixa qualidade da informação, a qual pode ser compensado com um adequado planejamente na coleta dos dados

Uma das fontes de informação mais utilizadas na avaliação de impacto são as pesquisas domiciliares. O objetivo destas pesquisas é coletar informação sobre as características demográficas e socioeconômicas dos domicílios e pessoas. No entanto, é importante destacar que, utilizar registros administrativos ou fontes de informação secundária podem ser úteis para conferir e avaliar as classificações dos domicílios alvos, segundo critérios de elegibilidades entre os beneficiários e não beneficiários de determinado programa social (COADY et al., 2004).

A importância na seleção de variáveis e na coleta de informação, para medir adequadamente o impacto do programas sociais e a focalização destes, induziu alguns países de América Latina a desenvolverem seus próprios índices de pobreza, principalmente com o objetivo de focalizar, apropriadamente, as intervenções sociais. Por exemplo, na Colômbia índice SISBEN[13] e no México um índice de elegibilidade multidimensional. Estes índices são muito importantes na avaliação de impacto do programa porque, através destes, podem-se ordenar as famílias ou domicílios alvos, permitindo a pré-classificação de beneficiários e não beneficiários. O resultado destes gera um cadastro para a seleção de beneficiários que possa ser atualizado, considerando que algumas variáveis podem perder o poder de prever a pobreza ou de discriminar beneficiário e não beneficiário (COADY et al., 2004).

13 SISBEN é utilizado para a seleção de beneficiários de subsídios de gasto social na saúde, educação, moradia, bem-estar familiar, entre outros (ver seção 3.4).

Neste sentido, para o presente trabalho o relacionamento da base de dados da pesquisa de campo AIBF e dos registros administrativo do CadÚnico, torna-se importante, devido ao fato de que a pesquisa de campo não utilizou o cadastro de famílias do programa[14] para conferir e avaliar as famílias dentre dos grupos de beneficiários e não beneficiários. O relacionamento de dados permitirá recuperar a variável que classifica ou aloca as famílias entrevistadas na pesquisa de campo nos grupos de beneficiários e não beneficiários segundo os registros administrativos do órgão responsável pelo monitoramento das famílias beneficiárias do Programa Bolsa Família. Desta forma, poderá avaliar a robustez dos resultados obtidos com os grupos alocados segundo a pesquisa de campo e segundo o registro administrativo.

E) Análise da informação

A informação coletada no método de avaliação quantitativo se expressa numericamente, e devem ser utilizadas ferramentas estatísticas para a sua análise. O objetivo de utilizar estas ferramentas é estimar o impacto médio do programa e o seu nível de significância. Na seção seguinte discute-se sobre os métodos de analises de informação a ser utilizada na avaliação de impacto dos programas sociais.

3.2.2. Os Métodos de avaliação de impacto

Diversos tipos de metodologias, experimentos ou desenhos de avaliação têm sido delineados para se estimar o impacto de programas sociais. Estas metodologias variam, prinicpalmente, na forma e nos critérios que se utilizam para construir o contrafactual (DIAZ e HANDA, 2004). Dois tipos de cenário contrafactual complementares são comumente utilizados: o primeiro compara as condições de vida dos indivíduos que participam do programa (grupo de tratamento ou beneficiários) com pessoas que não formam partes do grupo de beneficiários, mas apresentam características similares às dos beneficiários; e, o segundo cenário, que consiste em comparar a situação dos beneficiários em diferentes momentos do tempo (antes e depois da intervenção) com os não beneficiários. Dependendo destas características, os experimentos ou desenhos classificam-se em experimentais e não experimentais (BAKER, 2000).

[14] Na seção 3.5 discute-se sobre a não utilização do cadastro de famílias do programa

Segundo Schutt (2001), os métodos de avaliação assumem uma definição baseada na lei de causalidade, que permite que a execução do programa leva à uma variação nas variáveis de impacto (efeitos), quando todos os outros fatores permanecem constantes (ceteris paribus). Desta forma, o contrafactual procura isolar o efeito de fatores externos ao programa que puderam também ter causado as variações nas variáveis de impacto, para determinar o efeito liquido ou impacto do programa.

Determinar o cenário contrafactual é essencial para o desenho da avaliação, porque este pode ser realizado utilizando diversas metodologias classificadas em duas categorias gerais: desenhos experimentais (aleatórios) e desenhos não-experimentais (não aleatórios). No entanto, é complicado separar o efeito do programa das condições hipotéticas que podem ser afetadas pela história e o viés de seleção (BAKER, 2000). A seguir analisa-se com mais detalhe cada um destes desenhos.

a). Desenhos experimentais

Os desenhos experimentais sociais constituem a metodologia considerada como a mais robusta para a construção do cenário contrafactual na avaliação de impacto e são as referências para as avaliações das políticas públicas (HECKMAM, 1992). Para aplicar esta metodologia requer-se que a seleção de beneficiários e não beneficiários, do programa em estudo, seja realizada aleatoriamente, garantindo que os grupos de comparação sejam realmente comparáveis (EZEMINARI, RUDQVIST e SUBBARAO, 2002). Os grupos que constituem os experimentos sociais são denominados de grupo de tratamento, quando se trata dos beneficiários do programa, e grupo de controle, constituído pelos não beneficiários do programa. O grupo de tratamento diferencia-se pelos componentes ou combinações de componentes que recebem do programa. Embora, a maioria das avaliações considere dois grupos de comparação (tratamento e controle), em uma avaliação de impacto é possível formar múltiplos grupos de tratamento e controle.

A aleatorização realizada na seleção dos grupos de tratamento e controle garante que, em média, as diferenças entre estes grupos devam-se apenas ao fato de participar ou não no programa que se está avaliando, controlando assim, a incidência de outras variáveis independentes associadas com a variável de impacto e a participação no programa. Esta

característica permite que o grupo de comparação forneça informação do que aconteceu aos beneficiários, se estes não tiveram participado na intervenção (LALONDE, 1986)

Estes tipos de experimentos possuem uma notável tradição nos Estados Unidos, enquanto na Europa os estudos com dados obtidos com uma finalidade diferente à avaliação têm uma maior difusão. Embora estes sejam os melhores desenhos para avaliar um programa, estes, por sua vez, também apresentam algumas desvantagens na implementação ideal de uma avaliação. Na Europa, por exemplo, existem algumas reticências de ordem moral ou ética ao momento de excluir um grupo selecionado aleatoriamente para os escolhidos ao grupo de tratamento (HECKMAN e SMITH, 1995). Além disso, existem problemas do chamado viés de substituição causado pela possibilidade que dispõem a um membro do grupo do controle de participar em um tratamento externo similar ao programa que se pretende avaliar. Junto a este viés, também se observa o viés de abandono provocado pela negativa dos indivíduos selecionados de participar no programa (BURTLESS, 1995).

Durante as três décadas passadas muitos programas sob auspício federal e estadual nos Estados Unidos têm avaliado estes programas utilizando aproximações experimentais. Estas avaliações aleatorizadas têm sido utilizadas em muitos estudos de avaliação para execução de métodos não-experimentais, porque provém um método apropriado de referência. Muitas das intervenções têm sido em programas de emprego e treinamento de trabalho (voluntários e obrigatórios). Entre os voluntários, há o National Supported Work Demonstration (NSW), o AFDC Homemaker-House Health Aide Demonstration, e o The National Job Training Partnership Act Study (JTPA); entre os programas obrigatórios, há o State Welfare-to-Work Demonstrations e o Outside labor programs Tennessee's Student Teacher Achievement Ratio (Project STAR), este último foi um estudo experimental que avaliou o impacto de redução do tamanho da turma sobre os escores de um teste (DIAZ e HANDAL, 2004).

Na América Latina, há alguns exemplos conhecidos de avaliação de impacto com desenhos experimentais. Uma avaliação de impacto muito conhecido no México é o realizado pelo programa PROGRESA (atualmente OPORTUNIDADES), programa que tem como objetivo integrar simultaneamente as intervenções em matéria de saúde, educação e nutrição, entendendo que, com as melhoras destas dimensões, permita-se reduzir a pobreza. Em 1990, a administração do programa adotou como marco empírico para avaliar o seu efeito um método flexível para resolver o problema da avaliação. As vantagens

derivam de duas características principais: primeiro, trata-se com um desenho experimental na que se alocam em forma aleatória as localidades, e não domicílios ou pessoas, os grupos de tratamento e de controle. Em segundo lugar, reúnem-se os dados de todos os domicílios das localidades de tratamento e de controle antes e depois do início do tratamento. A combinação destas duas características permitiu aos pesquisadores avaliar o efeito direto médio do tratamento sobre os sujeitos ou, melhor dito, o efeito do programa sobre os participantes utilizando qualquer dos estimadores disponíveis na bibliografia sobre avaliação (SKOUFIAS, 2006). Na Nicarágua outra avaliação importante foi realizada ao programa "*Red de Protección Social*" (RPS). Este programa apresenta uma nova proposta na formação de redes de seguridade para as famílias mais pobres do país. O programa RPS foi desenhado em duas fases que abarcariam um período de cinco anos, iniciando no ano 2000, com uma fase piloto de três anos também chamado de Fase I. Para esta fase piloto selecionaram-se aleatoriamente 21 comarcas dos departamentos de Madriz e Matagalpa. Para manter um caráter experimental, selecionaram-se outras 21 comarcas, também de forma aleatoria, para serem observadas como um grupo controle de não intervenção. Assim, o primeiro componente da avaliação se centrou no programa piloto e utilizou um desenho experimental que incluíram trabalhos de campo entrevistas para estudar o impacto do programa em gastos e seguridade alimentares; escolaridade infantil e trabalho infantil; atenção na saúde de crianças menores de cinco anos (MALUCCIO, 2005). Outro exemplo de avaliação de impacto utilizando um desenho experimental é o realizado com o programa "*Proempleo*", executado na Argentina durante o período 1998-2000. O objetivo da avaliação foi determinar a eficácia de prover um subsídio salarial e capacitação aos atuais beneficiários de programas públicos de emprego transitório como meio para facilitar sua transição a trabalhos regulares no setor privado. O público Alvo do programa foram os beneficiários que estavam participando nos programas de emprego temporário administrados pelo Ministério do Trabalho da Argentina. Selecionaram-se três amostras aleatórias, em que um grupo dos selecionados recebia o subsídio salarial, outro grupo o subsídio salarial e a capacitação, e o último grupo não recebia qualquer dos benefícios, representando, assim, o grupo de controle (GALASSO, RAVALLION e SALVIA, 2001).

b) Desenhos não experimentais.

Os experimentos sociais constituem o método referencial para estimar o impacto dos programas sociais, mas usualmente estes experimentos nem sempre estão disponíveis, por diversas razões. Por um lado, os experimentos sociais são custosos e transcorre muito tempo desde o inicio do experimento até a obtenção dos resultados para sua avaliação. Por outro lado, existem algumas reticências de ordem moral ou ética no momento de excluir um grupo selecionado aleatoriamente para fazer parte do grupo de tratamento. Conseqüentemente, testar a confiabilidade dos métodos não experimentais é um assunto central na literatura de avaliação de programas (DIAZ e HANDA, 2004).

Comparando com o desenho experimental, este desenho não seleciona aleatoriamente os grupos de beneficiários e não beneficiários. No desenho não-experimental existem alternativas para selecionar o grupo de comparação de uma forma adequada. Estes métodos encontram ou identificam grupos de indivíduos que não participaram no programa, mas que cumpram com os critérios de seleção do programa e sejam similares às pessoas que formam parte do grupo de tratamento nas características observáveis que poderiam incidir na variável de impacto e na decisão dos indivíduos de participar ou não no programa (MOFFITT, 2003).

A vantagem principal dos desenhos não-experimentais é que é possível ter como base fontes de dados existentes e, portanto, freqüentemente são mais rápidos e menos custosos de implementar. Além disso, a avaliação pode ser realizada, quando o programa está em andamento, com a condição de que existam dados suficientes. As principais desvantagens das técnicas não-experimentais são, primeiro, que com freqüência reduz-se a confiabilidade dos resultados; e segundo, estes métodos podem ser estatisticamente complexos[15] (BAKER, 2000).

As técnicas não-experimentais podem ser de dois tipos: (1) metodologias não-experimentais com base em dados longitudinais, ou com dados transversais repetidos; e (2) os métodos baseados com dados transversais. Independente do tipo de dados que se

[15] Uma terceira desvantagem refere-se à possibilidade de que os estimadores apresentem um problema de viés de seleção.

disponha, as técnicas mais conhecidas dentro da avaliação de impacto com desenhos não-experimentais, são:

- Método diferença em diferença: baseados em dados longitudinais ou transversais repetidos.

- Comparações reflexivas: com base em dados longitudinais ou transversais repetidos.

- Método das variáveis instrumentais: baseados em dados transversais.

- Método de pareamento: com base em dados transversais.

- Método da regressão descontínua: baseados em dados transversais.

Quando um grupo de comparação é gerado e não alocado aleatoriamente, muitos fatores podem afetar a validade dos resultados. LaLonde (1986) apresentou alguns questionamentos sobre a confiabilidade dos estimadores de impacto do programa obtidos pela metodologia não-experimental. Analisando o programa NSW, demonstrou que, com base nos supostos comuns feitos por econometristas para justificar os estimadores não-experimentais, os métodos transversais, antes-depois e diferenças em diferenças não conduzem a estimadores confiáveis, se estes são comparados com estimadores experimentais. Por outro lado, Friedlander e Robins (1995) mostraram evidências no desempenho de métodos de ajuste de regressão pareamento como estimadores com métodos não-experimentais para programas com intervenções do emprego. Eles comparam as estimativas de impacto produzidas por este procedimento não-experimental com os de um experimental, no mesmo tempo e mesmo local dado, concluindo que um viés significativo surge somente ao comparar participantes do programa que residem em diferentes áreas geográficas, e não nas mesmas áreas.

Nos últimos anos, nos países em desenvolvimento, o desenho mais utilizado nas avaliações de impacto foi o não-experimental. Navarro (2005) na revisão dos desenhos de avaliação utilizados em algumas das avaliações de impacto realizadas durante a última década na América Latina, mostrou que 15 dos 19 programas que foram avaliados utilizaram só o desenho não-experimental; três avaliações aplicaram o desenho experimental; um ambos dos desenhos, e, um programa, o não-experimental.

3.3 Métodos de estimação de impacto para desenhos não experimentais

A seguir descrevem-se resumidamente as duas técnicas dos métodos não-experimentais da avaliação de impacto, que serão utilizados dando ênfases nas técnicas do pareamento e regressão descontínua, uma vez que, para fins desta tese, serão utilizadas estas metodologias.

3.3.1 Método diferença em diferença ou diferença dupla.

Este método consiste em comparar um grupo de tratamento e um de controle antes (primeira diferença) e depois de um programa (segunda diferença) (HECKMAN et al, 1998).

3.3.2 Comparações reflexivas.

Nesta técnica realiza-se uma pesquisa de referência junto aos participantes antes da intervenção do programa, com a qual é construído o contrafactual. Logo se realiza uma pesquisa de acompanhamento quando o programa está em andamento. Assim, são comparados os participantes de programa antes e depois da intervenção.

3.3.3 Método das variáveis instrumentais.

Este método utiliza uma ou mais variáveis que influem na participação do programa, mas não nos resultados dada a participação. Identifica a variação exógena nos resultados atribuíveis ao programa, reconhecendo que o estabelecimento não é aleatório, mas intencional[16] .

Observação importante:

Com a implementação da técnica de Pareamento por Escore de propensão (PSM) ou Matching de Escore de Propensão, que compara resultados de famílias similares do grupo de tratamento com as do grupo de comparação ou controle, deve-se ter em consideração a diferença que existem entre o termo "Pareamento para o relacionamento de base de dados" e "Pareamento (ou Matching) para a técnica utilizada na avaliação de impacto". Ambos os

[16] Para mais detalhes deste método, do método diferença em diferença ou diferença dupla e comparações reflexivas, ver ANEXO II.

termos mencionados, na sua definição estrita têm significados semelhantes, mas para nosso caso, com o objetivo de diferenciar e clarear as diferenças que existem entre as técnicas aplicadas no trabalho para cada procedimento que tem diferentes propósitos, realiza-se as seguintes observações:

O termo "relacionamento" será utilizado quando nos referimos a relacionamento das bases de dados realizados entre a base da pesquisa AIBF como os registros administrativos do CadÚnico, para não utilizar o termo de pareamento, e tem como objetivo realizar a re-alocação alternativa que se propõe neste trabalho para a distribuição dos grupos de comparação com os registros administrativos.

No entanto, o termo de "Pareamento" será referido para a técnica utilizada na avaliação de impacto dos programas sociais com o escore de propensão, cujo objetivo é construir pares sobre as observações de controle e o tratamento que são similares em termos das características observáveis. Logo, se mensura as diferenças das variáveis de impacto na educação do PBF entre o grupo de tratamento e o grupo de comparação ou controle, isto é, para ambos os procedimentos da alocação das famílias.

3.3.4 Métodos de Paramento (*matching*)

O método de "Paramento – *matching*" é uma aproximação não paramétrica para o problema de identificação do tratamento de impacto sobre os resultados. Isto é, no senso geral, nenhuma especificação em particular precisa ser assumida. Além disso, pode ser combinado com outros métodos, produzindo estimativas mais precisas e permitindo suposições menos restritivas. Contudo, o método também se baseia em suposições fortes e exigências sobre o tipo de informação que se precisa. O propósito principal do pareamento é restabelecer as condições de um experimento, quando os dados não estão disponíveis (BLUNDELL e COSTA, 2002).

O pareamento pode ser realizado por indivíduo ou por grupo de comparação. Quando se utiliza o pareamento por individuo procura-se que os pertencentes ao grupo de tratamento sejam comparáveis aos indivíduos do grupo de comparação (controle). O pareamento por grupo é menos exigente, mas requer que os grupos de tratamento e comparação sejam, em média, iguais. Assim o pareamento por indivíduo parece ser mais preciso e proporciona resultados mais confiáveis que o grupo de pareamento por grupo (FREEMAN, ROSSI, e

WRIGHT, 1980). Embora, as aplicações do pareamento por individuo sejam estatisticamente mais desejáveis que o de pareamento por grupo, em geral as avaliações de impacto utilizam o método agregado.

O pareamento consiste em construir pares sobre as observações de controle e tratamento que sejam similares em termos de suas características observáveis. Quando as diferenças relevantes entre duas observações são capturadas nas variáveis observáveis (pré-tratamento), o qual acontece quando o resultado é independente da alocação do tratamento, dada as variáveis pré-tratamento (suposto de independência condicional), então o método pareamento produz uma estimativa não enviesada do impacto do tratamento.

O pareamento é um procedimento simples de aplicar quando poucas características dos indivíduos afetam a variável de impacto e a decisão de participar no programa. Em geral os problemas que procuram resolver os programas sociais estão determinados por mais de duas variáveis, o que dificulta a aplicação do método de pareamento. Além disso, quando o pareamento não inclui todas as variáveis que determinam a variável de impacto e a participação no programa, poderia existir viés na estimação de impacto. Isto devido a que os grupos de tratamento e comparação não seriam estatisticamente comparáveis (RAVALLION, 1999).

Umas das vantagens na estimação do impacto do método de pareamento, é que os grupos de tratamento e comparação não têm necessariamente que se formar antes de iniciar a operação do programa. A outra vantagem é que o método de pareamento não exige que se proíba o ingresso ao programa de indivíduos que são parte da população objetivo da intervenção (RAVALLION, 1999). Em relação às desvantagens, observa que, quando se quantifica o impacto de um programa social com este método, encontram-se diferenças não observáveis entre os grupos de tratamento e comparação, que geram um "viés de seleção". Este viés gera-se pelo fato de que o ingresso ao programa é uma decisão do beneficiário e não de um processo aleatório como no caso do desenho experimental. Isto implica que as pessoas que decidem participar do programa poderiam ter características não observáveis pelo avaliador que influem na sua decisão de participar e, por sua vez, determinar a variável de impacto do programa.

Para ter maior facilidade na aplicação do pareamento têm sido desenvolvidos modelos econométricos que permitem controlar os efeitos de variáveis observáveis e identificar

aqueles indivíduos que são similares às pessoas que integram o grupo de tratamento. Os modelos de pareamento desenvolvidos estimam a probabilidade dos indivíduos de participar no programa através de modelos *probit* ou *logit*, utilizando como variáveis independentes uma série de características socioeconômicas dos indivíduos relevantes ao programa que se avalia. Um tipo particular deste método é a técnica de Pareamento de Escore de propensão (PSM) como um estimador de impacto (DIAZ e HANDA, 2004).

O PSM leva em consideração as diferenças entre os indivíduos que participaram do programa e os que não participaram, e pode ser resumida nos seguintes passos: primeiro, estima-se a probabilidade de que um indivíduo receba o tratamento; segundo, separa-se a amostra em duas sub-amostras, os tratados (os que receberam o tratamento) e os de comparação (os que não receberam o tratamento), e ordenam-se ambas as sub-amostras de forma descendente, e no último passo, para cada indivíduo do grupo de tratamento procura-se um indivíduo do grupo de comparação com similar escore, formando os pares.

O PSM, no transcurso dos estudos de avaliação de impacto realizados, apresentou defensores, mas também detratores. Rosenbaum e Rubin (1983) forneceram um rol central no estudo das relações de causalidade. Dehejia e Wahba (1998) destacaram que o PSM permite estimar com êxito o impacto de programas de trabalho e que se simplifica a tarefa de controlar por diferenças em variáveis prévias ao programa. Estes mesmos autores, em 2002, ressaltaram as boas propriedades do PSM ainda quando tem poucos casos de comparação (controles) com que comparar as unidades (forma mais geral que indivíduos) que receberam o tratamento. Entre os detratores, temos que Heckman, Ichimura e Todd (2003) desenvolveram um método de emparelhamento com base em distribuições de kernel e demonstraram que o PSM não implica necessariamente uma diminuição na variância dos estimadores. Por sua parte, Shadish et al. (2002) indicaram que se requer amostras grandes, com suficiente diferença entre grupos, e que existe algum viés devido ao fato de que o PSM só controla as variáveis observáveis.

Segundo Smith e Todd (2001), o PSM pode ser considerado como uma metodologia adequada se as seguintes condições são cumpridas:

1. A população que vai ser parte do grupo de tratamento e os do grupo de comparação deve pertencer à mesma amostra (ou pelo menos ao mesmo tipo de pesquisa), de tal forma que as variáveis sejam medidas da mesma forma.

2. Ambos os grupos participem do mesmo problema em estudo.

3. As bases de dados contenham um número suficiente de variáveis para modelar a decisão de participar no programa.

Nos últimos anos têm sido produzidos significativos avanços nas técnicas de correspondência do Escore de Propensão. Este método é muito atrativo para os avaliadores que tem restrições de tempo e não dispõem de dados de referência, uma vez que se pode utilizar, contando com apenas dados de corte transversal. Assim, parece que as estimações para dados com PSM, como um estimador de impacto, são levemente melhores que outros estimadores não-experimentais (ROSENBAUM e RUBIN, 1985; JALAN e RAVALLION, 1998).

I. Fundamentos matemáticos do método pareamento e estimadores de escore de propensão.

Para determinar a eficiência de uma medida dirigida aos problemas sociais é necessário descrever corretamente o conceito causal do problema. Isto é, o fundamental no estudo de avaliação é distinguir entre o efeito causal da participação em um programa social, D, e a correlação estatística entre a participação e a variável resultado, Y (DURAN, 2004). Uma extensa discussão do conceito de causalidade utilizado na econometria e na estatística pode ser encontrada em Cox (1992), Dawid (2000) e Holland (1986).

Com a finalidade de estudar a avaliação econométrica dos programas sociais será utilizado o modelo de resultados potenciais proposto inicialmente por Neyman (1923) e desenvolvido posteriormente por Rubin (1974) e Heckmam e Vytlacil (2000). Assim, a exposição formal do modelo mais simples, assume uma perspectiva estática e supõe que o estado de participação apenas toma dois valores 0 e 1.

Segundo Rubin (1974), a idéia básica do modelo é comparar os resultados potenciais de um indivíduo no caso de participar em um programa Y_1 com o resultado de não participar Y_0. A diferença entre os resultados potenciais $Y_1 - Y_0$ é o efeito causal, mas com base ao suposto de independência dos resultados individuais da participação de outros indivíduos. Para completar a especificação do modelo, define-se o estado de participação do indivíduo mediante a variável estocástica binária D cujas realizações são observáveis.

Conseqüentemente, a variável–resultado observada Y é uma função de D e dos resultados potenciais de interesses:

$$Y = Y_0\ (1 - D) + Y_1\ D = Y_0 + D(Y_1 - Y_0) \qquad [3.1]$$

a) Considerações iniciais do pareamento.

Neste trabalho utiliza-se a técnica de pareamento com base no escore de propensão. Esta técnica constrói pares dos beneficiários e os não beneficiários de um programa com base na sua probabilidade estimada de participação do programa $p(X)$. Está técnica é utilizada porque, em muitas aplicações de interesses, a dimensionalidade das características observáveis é alta, pelo que é difícil determinar sobre que dimensões fazer os pares ou que esquema de pesos a utilizar. Além disso, a técnica é muito útil, porque apresenta um esquema de pesos naturais que produz estimadores não enviesados de impacto do tratamento (ROSENBAUM e RUBIN 1983).

Uma característica importante é que o pareamento não requer uma restrição acerca de uma forma funcional a qual está implícita nas regressões comuns. Assim, se o pressuposto de independência condicional cumpre-se, mas a linearidade não, então o pareamento é consistente, enquanto a regressão não é. Além disso, o pareamento permite considerar o problema de suporte (*support problem*), que se refere ao suporte comum da distribuição do conjunto de valores para as quais se tem uma densidade positiva, isto é, o conjunto de valores com probabilidades diferentes de zero. Isto é importante quando se realiza o pareamento, porque em alguns casos os valores de X ou de $p(X)$ que estão presentes no grupo de beneficiários, não estão presentes no grupo de não-beneficiários. Assim mesmo, o suporte comum pode não incluir todas as observações dos participantes de um programa, mas, para calcular o impacto médio do tratamento sobre os tratados, apenas requer-se que existam observações parecidas com o grupo de tratamento no grupo controle.

b) Aleatoriedade.

Nos desenhos experimentais, os grupos de controle e de tratamento são eleitos aleatoriamente da mesma população e a diferença que há entre os dois grupos é por efeitos do recebimento dos benefícios do programa. Mas, quando não é possível construir um desenho experimental, o efeito do programa não pode ser observado diretamente.

Formalmente, seja *i* o índice para a população em consideração, Y_{i1} o valor do resultado quando a unidade i pertence ao tratamento (1), e Y_{i0} o valor da mesma variável quando a unidade pertence ao grupo de controle (0). O impacto do tratamento com base a um desenho experimental para uma observação, digamos $\mathbf{t}_i$, define-se como $\mathbf{t}_i = Y_{i1} - Y_{i0}$. Por outro lado, quando se está trabalhando como desenhos não-experimentais, o interesse é conhecer o efeito esperado do tratamento para a população tratada, por tanto:

$$\begin{aligned} t/_{D=1} &= E(t_i / D_i = 1) \\ &= E(Y_{i1} / D_i = 1) - E(Y_{i0} / D_i = 1) \end{aligned} \qquad [3.2]$$

Em que, D_i=1 (=0) se a *i-th* unidade se aloca ao tratamento (controle). O problema da não observação está explicado porque somente pode estimar $E(Y_{i1} / D_i = 1)$, mas não $E(Y_{i0} / D_i = 1)$. Uma forma de estimar o efeito do programa será estimando a diferença: $E(Y_{i1} / D_i = 1) - E(Y_{i0} / D_i = 0)$. Este é um estimador com viés da diferença ***t*** porque se esta aproximando $E(Y_{i0} / D_i = 1)$ com os não participantes auto-eleitos $E(Y_{i0} / D_i = 0)$. Este viés conhece-se como o viés de seleção[17], objeto de estudo nos desenhos não-experimentais.

c) Pareamento – *Matching* exato.

Quando não se conta com grupo de tratamento e controle eleitos aleatoriamente da mesma população, não é possível estimar o efeito do programa com a diferença dos resultados entre os dois grupos. Neste caso é possível substituir a ausência de unidades experimentais de controle se assumimos que os dados podem ser obtidos de um conjunto de potenciais unidades de comparação, as quais não necessariamente procedem da mesma população que as unidades de tratamento, mas as quais se podem observar o mesmo conjunto de variáveis pré-tratamentos, X_i.

d) Pressuposto de Independência Condicional:

Este suposto estabelece que, uma vez condicionados o vetor de características X, a participação no programa é independente do resultado no grupo controle. Assume-se que, tomando a alocação ao tratamento como aleatória, dadas algumas variáveis X; e em particular, as variáveis não observáveis não têm papel na alocação do tratamento (Rubin,

[17] O viés é igual a $E(Y_{i0} / D_i = 1) - E(Y_{i0} / D_i = 0)$

1977). Com base neste suposto, o efeito condicional do tratamento, $\mathbf{t}|_{D=1}$, se obtém primeiro estimando $\mathbf{t}|_{D=1,X}$ e logo uma média sobra a distribuição de *X* dado *D=1*. Esta proposição satisfaz-se se *X* inclui todas as variáveis que afetam tanto a participação, como o resultado. Assim, as diferenças destacáveis entre duas observações, são captadas nas variáveis observáveis pré-tratamento – que aconteceu quando o resultado é independente da alocação ao tratamento dada as variáveis pré-tratamento – podendo assegurar-se que os métodos de pareamento produzem um estimador não enviesado do impacto do tratamento (DEHEJIA, WAHBA, 1998).

e) O pressuposto do pareamento.

Este suposto é necessário para identificar alguma medida de impacto da população. Este é dado por:

$$0 < \Pr\{D = 1 \mid X = x\} < 1 \qquad [3.3]$$

Esta suposição assegura que para cada valor de *x* existam casos no grupo de tratamento e controle. Existe uma sobreposição entre uma sub-amostra dos tratados e não tratados, assim, para cada unidade do grupo de tratamento existe outra unidade dos não tratados com similar característica *X*.

f) O pressuposto da média condicional.

Chamada também como pressuposto da independência da média condicional:

$$\mathrm{E}\{y_0 \mid D = 1, x\} = \mathrm{E}\{y_0 \mid D = 0, x\} = \mathrm{E}\{y_0 \mid x\} \qquad [3.4]$$

O qual implica que y_0 não determina a participação.

g) O pareamento usando o Escore de Propensão.

Rosenbaum e Rubin (1983, 1985a, b) definem o escore de propensão como a probabilidade condicional de receber o tratamento dado um vetor de variáveis pré-tratamento:

$$p(x) \equiv \Pr\{\mathrm{D} = 1/X\} = E\{D/X\} \quad e\ p(x) < 1 \qquad [3.5]$$

No qual D = {0,1} é o indicador de exposição ao tratamento e X é o vetor multidimensional das características pré–tratamento.

A equação 3.5 é importante porque permite reduzir o problema da dimensionalidade no pareamento. Quando temos muitas variáveis, é difícil determinar sobre qual dimensão realizar o pareamento ou que esquema de pesos seguir.

O escore de propensão mensurado pode ser calculado dado o conjunto de dados (D_i, X_i) utilizando métodos paramétricos ou semi-paramétricos.

Um pressuposto que tem um papel importante na avaliação do tratamento é a condição de balanceamento, dado por:

$$D \perp X = x \mid p(x) \qquad [3.6]$$

Alternativamente, pode-se expressar que, para indivíduos como o mesmo escore de propensão a alocação ao tratamento é aleatório e pode ser visto identicamente em termos de qualquer vetor de *X*.

Rosenbaum e Rubin (1983) utilizando a independência condicional dado $p(x)$, definem:

$$Y_{i1},\ Y_{io} \perp D_i \mid X{=}x \Rightarrow Y_{i1},\ Y_{io} \perp D_i \mid p(x) \qquad [3.7]$$

Rosenbaum e Rubin (1983) mostram que se a exposição ao tratamento é aleatório dentro dos grupos definidos por *X*, isto é também aleatório dentro dos grupos definidos pelos valores de uma só variável *p(x)*.

h) Efeitos do tratamento e viés de seleção.

Para o pareamento utilizando o escore de propensão traz consigo um esquema de pesos, que determina os pesos que coincidem com as unidades de comparação quando calculamos o efeito estimado do tratamento. O valor desta técnica é que podemos aproximar o resultado de uma avaliação experimental, na que se tenta estimar o impacto médio de algum programa. Neste sentido, duas medidas de efeitos do tratamento são apresentadas: o efeito médio sobre o total de indivíduos e os efeitos médios sobre os tratados.

Parâmetros importantes:

Seja Δ a diferença entre os resultados dos tratados e não tratados, assim:

$$\Delta = Y_1 - Y_0 \qquad [3.8]$$

Considerando que Δ não é diretamente observável, já que o mesmo indivíduo não pode ser observado em ambos os grupos. Logo o valor populacional do efeito médio do tratamento (ATE) e efeito médio do tratamento sobre os tratados (ATT)[18], define-se:

$$ATE = E[\Delta],$$

$$ATT = E[\Delta \mid D = 1] \qquad [3.9]$$

As estimativas destes valores são:

$$\overline{ATE} = \frac{1}{N}\sum_{i=1}^{N} E[\Delta_i],$$

$$\overline{ATT} = \frac{1}{N_T}\sum_{i=1}^{N_T} E[\Delta_i \mid D_i = 1], \qquad [3.10]$$

Na qual, $N_T = \sum_{i=1}^{N} D_i$. Considerando que estes termos contêm uma componente não observável que precisa ser estimada, utilizando algum pressuposto.

A medida *ATE* é relevante no caso que o tratamento tenha aplicação universal, sendo razoável considerar que os ganhos hipotéticos do tratamento para uma seleção aleatória dos membros da população. No caso do *ATT*, é útil quando se considera o ganho médio do tratamento sobre os tratados (Heckman e Vytlacil, 2002).

Um dos parâmetros em que os estudos de avaliação centram-se é o efeito médio do tratamento sobre o tratado (*ATT*):

$$ATT = E(Y_1 - Y_0 \mid D = 1) = E(Y_1 \mid D = 1) - \underbrace{E(Y_0 \mid D = 1)}_{\text{Náo observado}} \qquad [3.11]$$

Dado que este é uma medida que reflete os efeitos do tratamento sobre aquelas pessoas que realmente têm participado no programa, seria um indicador mais eficaz da política implementada.

---

[18] Na literatura internacional: *Average treatment effect* (ATE) e *Average treatment effect on treated* (ATT).

O último termo na expressão [3.11] é o contrafactual de interesse, mas este não pode ser observado nos dados.

Uma alternativa para estimar esse contrafactual é utilizar $E(Y_0 \mid D=0)$, que é a média do resultado potencial no estado dos não tratados e que pode ser observado. No entanto, em geral, espera-se que $E(Y_0 \mid D=1) \neq E(Y_0 \mid D=0)$, o qual na estimação dos efeitos médios resultará em um viés, que surge devido a diferenças nas características observáveis e a diferenças nos atributos não observáveis entre os grupos de tratamento e controle.

O viés de seleção:

Algumas vezes o viés $B = E(Y_0 \mid D=1) - E(Y_0 \mid D=0)$ é causado pelas características que estão correlacionadas com a seleção dos participantes D como com o resultado *Y*. Quando as variáveis X são conhecidas e estão disponíveis, é possível resolver o problema de seleção controlando a estimação por estas variáveis. Rubin (1979) mostra que, para um valor dado dessas variáveis, não se produz, por definição, viés algum: $E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D=0) = E(Y \mid X, D=0)$.

No caso em que *D* e Y_0 sejam independentes para cada valor de X, esta condição recebe o nome de *pressuposto de independência condicional*. Assim, aplicando a lei de esperanças interativas pode-se escrever,

$$E(Y_0 \mid D=1) = E\{E(Y_0 \mid X, D=1) \mid D=1\} = E\{E(Y_0 \mid X, D=0) \mid D=1\} \qquad [3.12]$$

Logo, a expressão resultante do efeito médio do tratamento sobre os tratados (*ATT*) pode ser estimada de forma consistente a partir dos análogos amostrais, dado que depende apenas das variáveis observáveis.

$$E(Y_1 \mid D=1) - E(Y_0 \mid D=1) = E(Y_1 \mid D=1) - E\{E(Y_0 \mid X, D=0) \mid D=1\} \qquad [3.13]$$

A este parâmetro comumente conhece-se como “impacto médio do tratamento dos tratados”

O *ATT* utilizando o Escore de Propensão:

Dado uma população de unidades denotada por *i*, se o escore de propensão $p(x_i)$ é conhecido, o efeito médio do tratamento sobre os tratados (ATT) pode ser estimado como segue:

$$\begin{aligned} \overline{ATT} &\equiv E\{Y_{1i} - Y_{0i} / D_i = 1\} \\ &\equiv E\{E\{Y_{1i} - Y_{0i} / D_i = 1, p(X_i)\}\} \\ &\equiv E\{E\{Y_{1i} / D_i = 1, p(X_i)\} - E\{Y_{0i} / D_i = 1, p(X_i)\} / D_i = 1\} \end{aligned} \qquad [3.14]$$

no qual, a esperança externa é sobre a distribuição de $(p(X_i)|D_i = 1)$ e Y_{1i} e Y_{0i} são os resultados potenciais nos duas situações contrafactuais dos tratados e os não tratados. Utilizando a expressão (3.14) a estimação do escore de propensão não é suficiente para estimar o *ATT*. Isto porque a probabilidade de observar duas unidades com exatamente o mesmo valor do escore de propensão é, em princípio zero, dado que $p(X)$ é uma variável contínua. Vários métodos têm sido propostos na literatura para solucionar este problema e quatros dos mais utilizados são: o pareamento do vizinho mais próximo (*Nearest Neighbour Matching – NNM*), o pareamento do raio (*Radius Matching - RM*), pareamento de Kernel (*kernel Matching - KM*) e pareamento estratificado (*Stratification Matching - SM*) (BECKER e ICHINO, 2002).

II. Tipos de pareamento baseados no Escore de Propensão.

O objetivo nesta parte da avaliação é decidir que tipo de pareamento utilizar, para isso, a seguir os tipos de pareamento com base no Escore de propensão que são comumente referidos na literatura e que descrevem a metodologia utilizada neste trabalho e que é mencionado de forma sucinta a seguir:[19].

- O pareamento de vizinho mais próximo (NNM) consiste em selecionar as unidades não tratadas para o grupo controle de forma que minimize a diferença absoluta da probabilidade de participação da unidade tratada e não tratada.

[19] Para mais detalhes dos tipos de matching baseados no Escore de Propensão ver o ANEXO III.

- O pareamento Raio (RM), a unidade tratada só será pareada com uma unidade do grupo de controle, quando este possuir um valor de escore de propensão que se encontra em uma distância pré-definida (o raio) do escore de propensão.

- O pareamento de Kernel (KM) realiza-se uma média ponderada dos resultados das observações mais próximas a cada participante. Os pesos são alocados de forma inversamente proporcional à distancia entre os escores de propensão dos grupos tratamento e controle.

- O pareamento Estratificado (SM), método que se baseia no mesmo procedimento de estratificação utilizado para estimar o escore de propensão.

3.3.5 Método da regressão descontínua

A regressão descontínua é um método utilizado quando os dados provêm de um desenho não-experimental, caracterizando-se por considerar que a probabilidade de receber os benefícios do programa (ser parte do grupo de tratamento) é uma função descontínua de uma ou mais variáveis fundamentais para a elegibilidade do programa **(**Buddelmeyer e Skoufias, 2004**).**

Nos últimos anos, a regressão descontínua (RD) tem-se convertido na base da avaliação padrão para solucionar temas causais com dados não-experimentais. Uma característica intrínseca deste método é que o grupo de tratamento é dado para indivíduos se e somente se uma covariável observada intercepta um limiar conhecido. Assim, sob as condições dadas, a probabilidade de receber os benefícios do programa próximo ao limiar da variável se comporta aleatoriamente. Este é o único desenho que permite identificar o efeito causal do programa sem impor restrições exclusivas arbitrárias, suposições sobre o processo de seleção, forma funcional ou o pressuposto da distribuição do erro (BLACK, GALDO e SMITH, 2005).

A idéia do método de RD foi utilizada pela primeira vez por Thistlethwaite e Campbell (1960) com o objetivo de estimar o efeito de receber uma subvenção ao estudo sobre as subseqüentes aspirações de curso profissionais. Dado que a subvenção apenas é outorgada se os aspirantes superam um determinado escore obtido em uma determinada prova, o

status de tratamento de subvenção outorgada depende da forma descontínua do escore obtido.

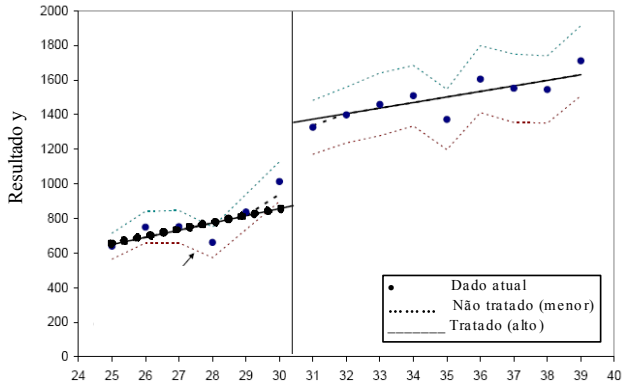
Por outro lado, Hahn, Todd, e van der Klauuw (2001) foram os primeiros a relacionar o desenho RD para a literatura de avaliação de programas e, juntamente com Porter (2003), estabeleceram formalmente menos condições para a identificação. As propriedades estatísticas da aleatorização no status de tratamento próximos ao ponto limiar é formalmente mostrado por Lee (2003), e algumas aplicações empíricas, incluindo Angrist e Lavy (1999), Black (1999), Van der Klaauw (2001), Lee (2003), Lemieux and Milligan (2004), Chen e Van der Klauuw (2004). Eles explicam também que a variação aleatória próxima do ponto de descontinuidade resolve o problema do viés de seleção. O ponto concordante em todos estes estudos empíricos é a confiança outorgada aos dados observacionais, que evitam a avaliação do desempenho dos estimadores econométricos RD, resolvendo o problema de avaliação. Embora haja várias discussões e aplicações do método RD na literatura de avaliação de programas sociais, importantes questões ainda permanecem no concernente à fonte de identificação e às formas de estimações dos efeitos do tratamento baseados nas restrições mínimas paramétricas (BUDDELMEYER e SKOUFIAS, 2004; BLACK, GALDO e SMITH, 2005).

I. Fundamentos matemáticos da regressão descontínua.

Usualmente no método de RD a literatura distingue dos cenários gerais do desenho, o desenho de regressão descontinua Sharp e Fuzzy (SRD e FRD respectivamente) (Trochim, 1984, 2001;HTV). Com o desenho Sharp (SRD) o tratamento, digamos “x”, é conhecido e depende em uma forma determinística de algumas variáveis observávei, enquanto o desenho *Fuzzy* (FRD) a variável “x” é uma variável aleatória, dadas as variáveis observáveis, mas a probabilidade condicional conhecida no ponto descontínuo que a variável observável toma o valor do limiar. Um exemplo é mostrado em Van der Klaauw (1996), no qual, a probabilidade que o estudante recebe ajuda financeira é uma função descontínua de um índice de estudante conhecido dos escores CPA e SAT. No entanto, existem outros fatores, alguns dos quais são não observáveis, que afeta à decisão de receber a ajuda financeira, e assim o ajuste dos dados deve ser realizado com um desenho *Fuzzy*, e não o *Shap* (DURÁN, 2004).

Para operacionalizar o desenho RD, deve existir a informação adicional para a regra de seleção, isto é, conhecer os mecanismos de designação ao tratamento, os quais dependem do valor de uma variável contínua observável, relativa ao umbral dado, ou ao *score* de corte, de tal forma, que a correspondente probabilidade de obtenção dos tratados (propensity score) é uma função descontínua desta variável no score de corte (ver FIG 3.1)

Figura 3.1 – Exemplo do um desenho de regressão descontínua.



Existem dois tipos de desenho de RD, o desenho *Sharp* e o chamado desenho *Fuzzy*. No primeiro, o tratamento x_i é conhecido e depende de uma forma determinística de alguma variável observável r_i. O desenho *Fuzzy* difere do primeiro, em que a atribuição ao tratamento não é uma função determinística de r_i (HAHN, TODD e VAN DER KLAAUW, 1999).

Neste estudo, revisaremos o desenho denominado de "Sharp"**,** no qual os indivíduos são alocados para o grupo tratamento ou controle somente com base em uma medida observável contínua *S*, chamada variável de seleção. Aqueles que estão acima do corte $\overline{S}$ não recebem tratamento e constituem o grupo controle, enquanto, aqueles que estão abaixo do corte *S*, recebem tratamento (*D=1*). Isto é, a alocação ao tratamento acontece por meio de uma decisão determinística mensurável e conhecida: $D_i = I\left[S_i > \overline{S}\right]$. Na figura 3.2, o desenho *Sharp* é mostrado com a linha sólida.

No desenho *Sharp* RD, temos:

$$E\left[u/TRAT, r\right] = E\left[u/r\right], \qquad [3.15]$$

no qual *u* denota o erro na equação do resultado potencial. Dado que *r* é apenas sistematicamente determinante de *TRAT*, *r* poderia capturar alguma correlação entre *TRAT* e *u*.

Com $TRAT_i = I\left[r_i > \bar{r}\right]$, a dependência entre $TRAT_i$ e u_i uma regressão de MQO deveria apresentar um estimador inconsistente de β_1. Previamente, mencionamos que uma aproximação da estimação do efeito do tratamento deve especificar e incluir a função média condicional $E\left[u/TRAT, r\right]$ como uma "função controle" na equação de resultados potenciais. Assim,

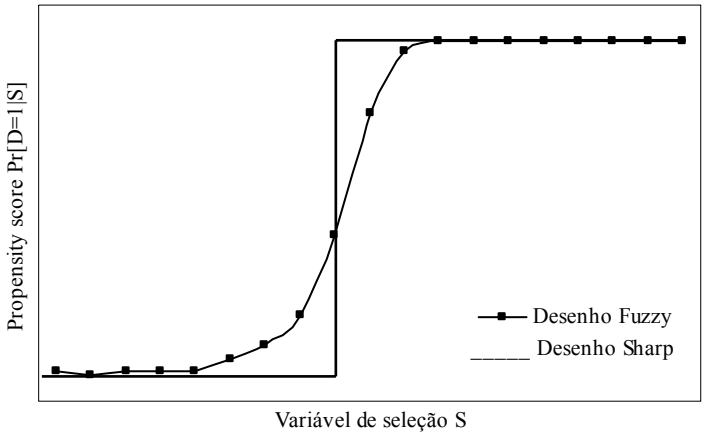
$$Y_i = \beta_0 + \beta_1 TRAT_i + \delta(r_i) + \sum_{j=1}^{J} \theta_j X_{ij} + \varepsilon_i \qquad [3.16]$$

$$y_i = \beta + \alpha D_i + k(r_i) + \varepsilon_i \qquad [3.17]$$

em que, $\varepsilon_i = Y_i - E\left[Y_i/TRAT_i, r_i\right]$. Se $k(r)$ é corretamente especificada, a regressão poderia consistentemente estimar β_1.

Em um caso especial em que $k(r)$ seja linear, β_1 poderia ser estimado pela distância entre duas linhas de regressão paralelas lineares no ponto de corte, que é igual à diferença entre os dois interceptos. Assim, se a função controle é linear, o estimador do efeito comum do tratamento é não viesado.

Figura 3.2 – Desenho Regressão Descontínua: Desenho *Sharp* e *Fuzzy*



Estimação e identificação no desenho RD

Intuitivamente, neste modelo, uma amostra de indivíduos em uma pequena vizinhança do corte poderia ser similar a um experimento aleatorizado no mesmo ponto de corte, dado que eles apresentam essencialmente a mesmo valor *S*. Espera-se que aqueles que apenas estão abaixo do corte sejam muito similares, aos que estão pouco acima deste. A comparação da média *y*, valor daqueles acima e aqueles abaixo do corte poderiam produzir uma boa estimação do efeito médio do tratamento.

Quando se incrementa o intervalo em torno do corte, este pode influenciar na estimação do efeito do tratamento, especialmente se a variável de alocação foi por si só relacionada à variável de resultados potenciais, condicionado ao estado de tratamento. Se uma suposição sobre a forma funcional desta relação pode ser feita, então é possível utilizar mais observações e extrapolar acima e abaixo do ponto de corte (Como um experimento aleatorizado teria mostrado no ponto de corte). Esta dupla extrapolação, combinada com exploração do experimento aleatorizado ao redor do ponto de corte, foi a idéia principal, base para a análise da regressão descontínua (VAN DER KLAAUW, 2002)

No desenho de RD, se deve garantir que,

$$\lim_{r\downarrow \bar{r}} E[Y/r] - \lim_{r\uparrow \bar{r}} E[Y/r] = \beta_1 + \lim_{r\downarrow \bar{r}} E[u/r] - \lim_{r\uparrow \bar{r}} E[u/r], \qquad [3.18]$$

Para formalizar esta expressão, na ausência do tratamento, indivíduos no pequeno intervalo em torno de $\bar{r}$ deveriam ter resultados médios similares se é observado o seguinte:

- A função média condicional $E[u/r]$ é continua em $\bar{r}$

- A função média do efeito do tratamento $E[\beta_1/r]$ é contínua à direita em $\bar{r}$:

$$Y_i = \beta_0 + \beta_1 TRAT_i + \delta(r_i) + \sum_{j=1}^{J} \theta_j X_{ij} + \varepsilon_i, \qquad [3.19]$$

em que, $\varepsilon_i = Y_i - E[Y_i/TRAT_i, r_i]$.

II. Implementação da Regressão Descontínua (RD).

Considerando o desenho de regressão descontinua (RD) definido anteriormente, temos:

- "r" é uma variável de alocação do tratamento.

- "x_i" é o nível de tratamento relativo a uma variável de alocação *r*, o qual apresenta descontinuidade digamos "r_0", ponto de descontinuidade.

- Sobre um vizinhança de r_0 e com base em alguns pressupostos, o ponto descontinuo sobre a variável de resultados de impacto "*Y*", pode ser atribuído à mudança no nível de tratamento.

Da mesma forma, sobre os dois desenhos RD, "*Sharp*" e "*Fuzzy*" definidos, alguns esclarecimento podem ser feitas para implementar tal desenho.

No desenho *Sharp*, o tratamento x_i, aumenta de zero a um em r_0, enquanto, no desenho *fuzzy*, o tratamento incrementa descontinuamente, ou a probabilidade do incremento do tratamento descontinuamente, mas não de zero a um, assim, prefere-se considerar as mudanças pelo incremento esperado de x_i, em r_0, construindo uma estimativa do impacto causal de uma mudança de uma unidade em x_i. Assim, no RD *Sharp*, a descontinuidade (ou "saltos") nos resultados Y_i, em r_0, é a estimação do impacto causal de x_i, enquanto que o RD *Fuzzy*, o deslocamento nos resultados Y_i pelo deslocamento em x_i, em r_0 é a estimação

local de Wald (equivalente ao loca de estimação de variáveis instrumentais) do impacto causal.

Os pressupostos e testes chaves para a implementação.

Por outro lado, para realizar esta implementação no programa de computador, precisa-se de alguns supostos que permitam inferir o efeito causal sobre Y_i, devido à mudança abrupta de x_i em r_0:

PS1. A mudança de x_i em r_0 é verdadeiramente descontínua

PS2. r é observado sem erro

PS3. Y_i é uma função contínua de r em r_0 na ausência do tratamento

PS4. Os indivíduos não são ordenados por meio de r_0 na sua sensibilidade ao tratamento.

Embora haja a necessidade de se utilizar estes pressupostos, nenhum deles pode ser testado diretamente, mas existem testes que permitiram a utilização, tal como se mostra a seguir:

T1. Testar na hipótese nula que nenhuma descontinuidade do tratamento acontece em r_0. ($\Delta x_i\,(r_0) \neq 0$).

T2. Testar que não há qualquer outra descontinuidade diferente de x_i ou Y_i longe de r_0. ($\Delta x_i\,(r \neq r_0) = 0$ e $\Delta Y_i\,(r \neq r_0) = 0$).

T3 e **T4**. Estes dois testes predeterminarão que as características e a densidade de exibição de r não pulam em "salto" a r_0, assim, a própria estimativas normalmente provê um teste que o efeito de tratamento é não zero (Y_i "salta" em r_0 porque x_i "salta" em r_0). ($\Delta x^c(r_0) = 0$) e ($\Delta f(r_0) = 0$).

Para estimar o tamanho de um "salto" descontínuo é possível realizar uma comparação de médias em pequenas caixas à esquerda e direita de r_0, ou via uma regressão com vários controles de r, um indicador D para $r>r_0$, e interações de todas as condições de r em D, mas desde que o objetivo é estimar o efeito precisamente no ponto (r_0) utilizando só

observações adjacentes a este r_0, a aproximação padrão é utilizar a regressão local que minimiza o viés (FAN e GIJBELS, 1996)[20].

Tendo escolhido usar regressão linear local, a escolha de largura da banda e kernel serão fundamentais. Assim, várias técnicas estão disponíveis para escolher larguras da banda, destacando o triângulo de Kernel, porque apresenta propriedades boas no contexto de RD (CHENG et al. 1997). A seguir apresenta-se a implementação para cada um doa cinco testes mencionados:

T1. $\Delta x_i\ (r_0) \neq 0$. Neste caso, primeiro estimam-se os erros padrões utilizando a regressão linear local *bootstrap*[21]. Neste programa, a variável de alocação r_0 assume pela definição que o ponto de corte é $r_0=0$. Utiliza-se o triangulo de kernel e o largo da banda padrão. Além disso, a regressão linear local (com lpoly) é calculada duas vezes, a primeira utilizando as observações ao lado do corte, para o qual $r<0$, e um para $r>0$. Logo a estimação do salto utiliza apenas as predições no corte $r_0=0$.

T2. ($\Delta x_i\ (r \neq r_0) =0$ e $\Delta Y_i\ (r \neq r_0) = 0$). Para este teste, precisa-se assumir só a continuidade x_i e Y_i em r_0, desta forma assegura-se que se rejeita a nulidade só em 5% de casos, e tendo definido um programa da descontinuidade, é possível escolher aleatoriamente 100 pontos de corte placebos $r_p=r_0$, sem substituição e testar a continuidade de x_i e Y_i em cada um.

T3. ($\Delta x^c{}_i\ (r_0) =0$) Considerando que o incremento no tratamento x_i é produto da alocação aleatória na vizinhança do ponto de corte r_0, características predeterminadas x^c dos indivíduos não deveriam apresentar descontinuidade no ponto r_0. No caso da RD simplesmente precisa-se testar que a estimação do salto em cada x^c predeterminada é zero no ponto r_0, ou $\Delta x^c(r_0) =0$ para todo x^c.

T4. $\Delta f(r_0) = 0$. Segundo McCrary (2007), a violação de permutação de observações em torno do ponto de corte r0, pode ser observado quando os indivíduos manipulam sua alocação, alterando seus dados ou ocultando, assim, os indivíduos próximos a r0 podem mudar cruzando o limite. Isto produz a descontinuidade na densidade de r em r_0. No entanto, McCrary (2007) aponta que a ausência de uma descontinuidade na densidade de r

[20] No Programa de STATA este procedimento é realizado com o comando "*lpoly*".

[21] Isto é implementado em programa de estimação "***discont***", que forma parte da regressão descontinua (incluído no comando "***rd***") no pacote estatístico STATA versão 9.

em r_0 não é necessária nem suficiente para a permutação, mas uma falha para rejeitar a hipótese nula que a densidade no salto de r em r_0 é zero está apresentada[22].

Estimador do efeito do tratamento. Este está relacionado com a estimação do efeito causal. Assim, temos que, no caso da RD *Sharp*, no qual x_i "salta" de um a zero de forma direta, enquanto que, no RD *fuzzy* para estimar o "salto" na escala de Y_i pelo "salto" de x_i e r_0, é dado pelo estimador de Wald local, para o qual precisa modificar o programa para estimas ambas das descontinuidades, e o qual esta já implementado no programa "***rd***"

Finalmente o programa que implementa a RD precisa de três argumentos, a variável de resultado Y_i, x_i, e r_0, assumindo que r0=0, e utilizando um *hardwired* padrão de *bandwidth* de 0.06[23]

3.3.6 Resumo dos métodos de avaliação

Como resumo dos desenhos e métodos de avaliação de impacto, na FIG 3.3 apresentam-se as principais características dos três tipos de desenho utilizados na análise quantitativa do impacto gerado por programas sociais. Pode-se concluir que existe uma relação inversa entre a aplicabilidade destes desenhos e a confiabilidade dos resultados que se podem obter ao aplicar cada um dos desenhos. Além disso, deve-se considerar que, em muitas avaliações, nestes desenhos, substitutos são utilizados como alternativas complementares da avaliação.

---

[22] No caso da implementação do RD, um programa utilizando o comando ***kdensity*** é proposto com o objetivo de estimar a densidade à esquerda e direita de r_0.

[23] O programa "***rd***" do STATA é similar ao espírito descrito na implementação acima mostrado para a estimação do efeito, mas considerando mais opções

FIGURA 3.3 – Métodos de formação de grupos contrafactuais segundo desenhos dos experimentos sociais

Desenho do experimento social	Método segundo a conformação do grupo contrafactual (Variável X)
Experimental	
	Aleatorização X = 1 (beneficiários) X = 0 (não beneficiários)
Quase-experimental ou não experimental	
Dados longitudinais	Comparações reflexivas ou Método *difference in differences* X = 1 (beneficiários na situação com projeto) X = 0 (beneficiários na situação sem projeto)
Dados de corte transversal.	pareamento, IV e Regressão descontinua X = 1 (beneficiários) X = 0 (não beneficiários)

Entre estas técnicas de desenho não-experimental, em geral considera-se que as técnicas de comparação que utilizam o pareamento são as alternativas sub-ótimas ao desenho experimental. Além disso, nos últimos anos a regressão descontínua tem conseguido colocar-se entre umas das técnicas de avaliação preferidas quando o desenho é não-experimental. Grande parte da bibliografia sobre metodologias de avaliação que centram a utilidade deste tipo de avaliações indica com freqüência as comparações pareamento e ultimamente a regressão descontínua (ROSENBAUM e RUBIN, 1985; JALAN e RAVALLION, 1998).

3.4 Os programas sociais no Brasil e o programa Bolsa Família

3.4.1 Os programas sociais no Brasil

As políticas públicas que vigoram na atualidade no Brasil estão alinhadas nas reformas realizadas pelo Governo Federal desde inícios dos anos 1990. Assim, diversas políticas públicas têm sido criadas para promover o bem-estar social da população, sendo planejadas e executadas na sua maioria pelo Governo Federal, objetivam ajudar as famílias de baixa renda (KASSOUF, 2004). Estas políticas introduziram novos conceitos de programas sociais, tais como focalização, descentralização e transferências de renda. Estas

características aplicam-se com diferentes ênfases nas políticas e programas sociais na atualidade.

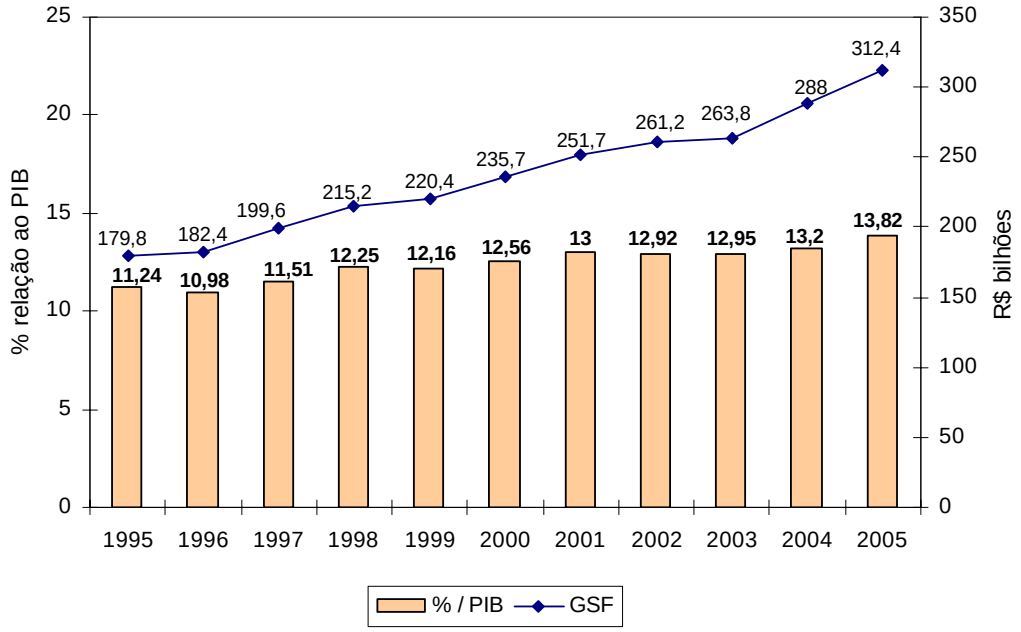
O objetivo dos programas focalizados é aumentar a efetividade do gasto social, alocando os recursos escassos nos grupos com maiores carências econômicas e sociais. Além disso, minimizam comportamentos dependentes dos usuários das políticas sociais, distinguindo o tipo de benefícios que recebem e as condições requeridas para o acesso aos programas (COADY, GROSH e HODDINOTT, 2004).

Embora nos últimos anos as políticas de combate à pobreza no Brasil tenham aumentando a sua cobertura e a sua eficiência, nas décadas passadas, estas estiveram assentadas mais no que se entende por políticas compensatórias e assistencialistas do que em políticas sustentáveis. Entende-se por políticas assistenciais e compensatórias aquelas que procuram ações imediatas e temporárias, no marco da compensação das desigualdades e da exclusão social, para aliviar os problemas sociais ou, especificamente, a pobreza (VACCARISI, 2005). Este grupo de políticas públicas é considerado como boa iniciativa por parte do Estado para controlar os problemas, mas necessitam condições concretas através do desenvolvimento integrado e sustentável das ações para erradicar os problemas sociais. Entre os casos representativos deste tipo de políticas sociais estão os programas dirigidos aos efeitos da seca do Nordeste sobre a fome e pobreza, que, nas décadas passadas, foram mantidos inalteráveis para solucionar o problema (ARBACHE, 2003).

Nas políticas e programas voltados a solucionar os problemas da população, o volume de gastos sociais é um fator importante para os resultados das intervenções, porque este representa as despesas públicas alocadas aos programas sociais nos níveis governamentais. No Brasil, os gastos sociais federais (GSF) têm crescido na última década, embora estes gastos ainda representem uma quantia pequena, quando comparamos a participação com o Produto Interno Bruto (PIB). Em 1995, o gasto federal destinado à área social esteve em torno dos R$179,8 bilhões, atingindo em 2000 os R$235,7 bilhões, e chegando a um montante de R$312,4 bilhões em 2005. Estes valores significam que, entre 1995 e 2005, os GSF cresceram em termos reais 74% e, que de forma agregada, representam em torno de R$11 bilhões ao ano para as políticas sociais.

Comparando os valores totais dos gastos sociais federal com o produto interno bruto, a posição relativa mostra um progresso durante os 11 anos analisados, crescendo de 11,24% em 1995 para 12,3% em 2005.

GRAFICO 3.1 – Evolução do Gasto Social Federal (GSF)[1] e porcentagem de participação em relação ao PIB. Brasil: 1980-2003.



Fonte: Disoc/Ipea.
Nota: 1 Valores deflacionados mês a mês, para dezembro de 2005 pelo IPCA.

Com as novas tendências das políticas públicas e o incremento do gasto federal na área social, os programas adquiriram um formato mais integral tanto nas instituições públicas, como no seu funcionamento. Assim, no Brasil, nos anos 90, foram integrados ministérios e programas sociais mais específicos, com o objetivo de diminuir a desigualdade social e econômica do país. Além disso, como resultados das políticas públicas, os programas que impulsionam a melhoria das condições econômicas e sociais da população, podem ser agrupados nas seguintes categorias (KASSOUF, 2004):

1. Os voltados para educação e erradicação do trabalho infantil;

2. Aqueles que atendem à criança e adolescente;

3. Dirigidos a aliviar ou combater a pobreza;

4. Os que estimulam a geração de emprego e renda;

5. Programas cujo objetivo é melhorar as condições de saúde da população;

6. Programas que promovem o desenvolvimento rural; e

7. O último grupo, destinado a investir na moradia popular e infra-estrutura urbana.

Destes grupos de categorias, iremos destacar aqueles de transferência de renda como Bolsa Família, Bolsa Escola, PETI, Bolsa Alimentação e Auxílio-Gás, que tem como objetivo principal aliviar ou combater a pobreza e, conseqüentemente, atendem à criança e ao adolescente, melhorando as condições de educação e erradicação do trabalho infantil, como também as condições de saúde da população. Alguns destes programas, inclusive, estão passando por um processo de integração ou sofrem modificações para melhorar sua eficácia; outros estão sendo executados pelos governos federal, estaduais e municipais em conjunto.

Os programas de transferência condicionada de renda.

Os programas de transferências condicionadas de renda (TCR) consistem na transferência direta de dinheiro a famílias ou indivíduos pobres sempre e quando se comprometam a certas condições, geralmente aquelas que implicam um investimento no capital humano como a freqüência regular de seus filhos à escola ou a centros de saúde. Como os programas TCR têm atingido êxitos no seu desenvolvimento, sobretudo na América Latina e no Caribe, estes têm sido objeto de rigorosas avaliações quanto a sua eficácia (RAWLINGS, RUBIO, 2003). No Brasil, a idéia central dos Programas TCR é proceder a uma articulação entre transferência monetária e políticas educacionais, de saúde e de trabalho direcionadas a crianças, jovens e adultos de famílias pobres (SILVA, 2006?).

A seguir menciona-se, de forma sucinta, os principais programas coordenados e fiscalizados pelo governo federal[24], e que objetivam aliviar ou combater a pobreza, melhorar as condições de educação e saúde das crianças e adolescentes e erradicar o trabalho infantil.

[24] Embora muitos destes programas atualmente já tenham sido fusionados ao programa bolsa família, no ANEXO IV, descrevem-se de forma detalhadas os programas coordenados e fiscalizados pelo governo federal para ter uma idéia da focalização destes programas.

- Bolsa Escola, programa pioneiro no que diz respeito aos programas de transferência condicionada de renda, sendo um programa de garantia de renda mínima vinculada à educação.

- Auxílio Gás, programa criado em 2001 com o objetivo subsidiar o preço do gás liquefeito de petróleo para famílias de baixa renda.

- Bolsa Alimentação, programa de Renda Mínima vinculado à saúde, que consiste em melhorar as condições de saúde e nutrição de gestantes, mães que estão amamentando filhos menores de seis meses, e crianças de 6 meses a 6 anos e 11 meses.

- Cartão Alimentação, criado, em 2003, com o objetivo de conceder um benefício às famílias em situação de insegurança alimentar.

- Benefício de Prestação Continuada (BPC), que garante um salário mínimo mensal a idosos com 67 anos ou mais e a pessoas portadoras de deficiência incapacitadas para o trabalho e para a vida independente, seja por deficiência física, seja por deficiência mental.

- Programa de Erradicação do Trabalho Infantil (PETI) tem como objetivo eliminar, em parceria com os diversos setores dos governos estaduais e municipais e da sociedade civil, o trabalho infantil em atividades perigosas, insalubres e degradantes.

3.4.2 O programa Bolsa Família (PBF).

Programa criado pelo Governo Federal por meio da medida provisória n°163, de 20 de outubro de 2003, e que tem "por finalidade a unificação dos procedimentos de gestão e execução das ações de transferência de renda do Governo Federal". Os programas unificados[25] foram o Bolsa Escola, o Bolsa Alimentação, o Cartão Alimentação (Fome Zero) e o Auxílio Gás, utilizando os dados do Cadastramento Único. Logo que as famílias são cadastradas no Cadastro Único do Governo Federal, para as famílias selecionadas com renda mensal de até R$60,00 por pessoa, o PBF deposita, mensalmente, um benefício fixo de R$60,00, tenham filhos ou não. Além desse valor fixo, as famílias com filhos abaixo de

[25] Até que todas as famílias que atualmente recebem algum benefício dos programas existentes sejam incorporadas ao Bolsa Família, nenhum desses programas será interrompido.

15 anos têm um benefício variável de R$18,00 por filho, até o limite de três benefícios. Para as famílias com renda mensal acima de R$60,00 e até R$120,00 por pessoa, o Bolsa Família deposita, mensalmente, o benefício variável de R$15,00 por filho com menos de 15 anos, até o limite de três benefícios. Dado que este trabalho utiliza os dados da avaliação do PBF, na seção seguinte explana-se melhor este programa.

O Programa Bolsa Família (PBF), do Governo Federal, integra o Programa "Fome Zero",[26] os seus objetivos principais são a promoção do alívio imediato da pobreza, o reforço ao exercício de direitos sociais básicos nas áreas de saúde e educação e a coordenação dos programas complementares, que têm por objetivo o desenvolvimento das famílias, de modo que os beneficiários do Bolsa Família consigam superar a situação de vulnerabilidade e pobreza (BRASIL, 200-?c).

O Programa Bolsa Família, para cumprir os seus objetivos, realiza pagamentos que variam de R$18,00 (dezoito reais) a R$112,00 (cento e doze reais), de acordo com a renda mensal por pessoa da família e o número de crianças, gestantes e nutrizes. No caso de famílias que migraram de programas remanescentes, o valor do benefício pode ser maior, tendo como base o valor recebido anteriormente.

Os benefícios financeiros estão classificados em dois tipos: para as famílias em situação de extrema pobreza (com renda mensal per capita de até R$ 60,00), o benefício parte de um valor básico de R$60,00 para aquelas sem ocorrência de crianças, gestantes e nutrizes, e as famílias em situação de pobreza (com renda mensal per capita de R$ 60,01 a R$ 120,00) adiciona-se um valor variável de R$18,00 para cada ocorrência de crianças, até o teto de três (BRASIL, 200-?c).

A taxa de cobertura no PBF é dinâmica, devido ao grau de expansão dos dois últimos anos. Assim, pode-se dizer que ultrapassou o seu objetivo de 3,6 milhões de famílias em seus primeiros três meses de operação em 2003 (alcançando 3,615 milhões). Ao final de 2004, um número total de 6,5 milhões de famílias foi atingido; até janeiro de 2005 atingiram-se aproximadamente 6,6 milhões de famílias e, em outubro de 2005, aproximadamente 8,0

[26] O FOME ZERO é uma estratégia impulsionada pelo governo federal para assegurar o direito humano à alimentação adequada às pessoas com dificuldades de acesso aos alimentos.

milhões de famílias residentes em todos os municípios brasileiros eram atendidas pelo programa (Brasil, 200-?c).

3.5 A Pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF)

A Avaliação de Impacto do Programa Bolsa Família (AIBF) é uma pesquisa que foi realizada em 2005 com o objetivo de avaliar o impacto do programa social Bolsa Família nas dimensões decorrentes das restrições orçamentárias e da operação de aspectos comportamentais ligados às condicionalidades do programa: Estrutura Relativa de Gastos, Antropometria, Saúde, Educação, Trabalho Infantil e da Mãe. Esta pesquisa ganhou muita importância pela abrangência que o Programa Bolsa Família tem atingido na população brasileira (OLIVEIRA et al, 2007).

3.5.1 Implementação da avaliação

Na análise de impacto, a primeira tarefa a realizar é estimar o cenário contrafactual ou simulado alternativo, isto é, o que teria acontecido se o programa não tivesse sido implementado ou o que teria ocorrido "normalmente". Para determinar o cenário contrafactual, precisa-se separar o efeito das intervenções de outros fatores; uma tarefa um tanto complexa. Isto é conseguido com a ajuda de grupos de comparação ou de controle (aqueles que não participam de um programa nem recebem benefícios), que se comparam com o grupo de tratamento (pessoas que recebem a intervenção).

Para o PBF, foi implementada uma avaliação não-experimental, dado que o programa foi criado a partir da migração e integração de vários programas prévios, sem possibilidade de definir um momento "antes" no qual a realização do experimento (aleatorização) pudesse ser efetuada. Além disso, uma vez que o programa tem como meta a universalização entre a população abaixo da linha da miséria e da linha de pobreza, o estabelecimento de um grupo de controle aleatório criaria um problema ético de negação do benefício a um determinado número de famílias necessitadas (OLIVEIRA et al, 2007).

Considerando a avaliação não-experimental, o AIBF optou pela elaboração de uma pesquisa de linha de base domiciliar, de cunho observacional. A pesquisa foi desenhada para servir como âncora a várias outras pesquisas, dentro do mesmo plano amostral, que,

no conjunto, constituem um painel longitudinal para a delimitação do impacto do programa ao longo do tempo (OLIVEIRA et al, 2007).

Para realizar esta pesquisa de linha de base domiciliar, o AIBF utilizou o procedimento de amostragem em 3 estágios: seleção de municípios (estratificados por cobertura do Programa Bolsa Família), seleção de setores (estratificado por renda) e seleção de domicílios (identificados no campo através do instrumento *screening*[27]), tendo como meta abranger 15.000 (quinze mil) domicílios. O tamanho da amostra foi definido para obter representatividade para três grandes áreas do país, a região Nordeste (NE), as regiões Sudeste e Sul (SE-S), em conjunto, e as regiões Norte e Centro-Oeste (N-CO), também em conjunto.

A amostra foi distribuída em domicílios identificados com os seguintes perfis: domicílios beneficiários do programa (casos); domicílios com famílias cadastradas no Cadastro Único, mas ainda não-beneficiarias do programa (controle 1); e domicílios sem famílias beneficiárias ou cadastradas (controle 2); dando probabilidade diferente para cada grupo, com as seguintes proporções: 30% (casos), 60% (controle 1), 10% (controle 2).

A seleção da amostra de domicílios foi feita por amostragem estratificada simples, sendo os estratos formados por setor e por classe de domicílios. A alocação de domicílios para esta fase da amostra foi, conforme mencionado anteriormente, feita na razão de 3 – 6 – 1, para casos, controles tipo 1 e controles tipo 2, respectivamente, a cada 10 domicílios selecionados. A coleta de dados foi executada durante o mês de novembro de 2005 e resultou em um total de 15.426 questionários completos. Para o estrato SE-S, este total foi de 5.887. Os estratos NE e N-CO apresentaram totais de 5.106 e 4.433, respectivamente (OLIVEIRA et al, 2007).

3.5.2 Método de avaliação de impacto do programa

A técnica não-experimental utilizada na pesquisa foi a estimativa dos diferenciais, a partir do pareamento de grupos de tratamento e controle por intermédio do escore de propensão

---

[27] *Screening* é uma listagem completa de cada setor, com quesitos simples que captem informação de características dos domicílios e que não incluam renda, alocando os domicílios nos três grupos segundos os perfis definidos (cadastro estratificado por estas categorias).

(Propensity Score Matching - PSM). A técnica do pareamento por escore de propensão garante a similaridade entre os grupos de tratamento e controle no que tange aos atributos observáveis, mas não assegura os atributos não-observáveis. Esta técnica possui o pressuposto de independência condicional aos atributos observáveis dos grupos de tratamento e controle. Isto significa que, se somente os atributos observáveis causam viés nas medidas de impacto, então a estimativa não-experimental dará uma boa medida de impacto.

Grupo de tratamento / comparação e elegibilidades

Para a classificação dos domicílios, o critério de elegibilidade considerou os seguintes cortes de renda domiciliar per capita: o primeiro corte constituído pelos domicílios que, na data da pesquisa, declaram ter uma renda domiciliar per capita até R$50,00 (valor que coincide com as famílias em extrema pobreza). O segundo corte incluiu os domicílios que, na data da pesquisa, declararam uma renda domiciliar per capita mensal até R$100,00 (valor que coincide com o limite de renda oficial definido para elegibilidade ao programa). Um terceiro corte de renda considerou os domicílios com renda domiciliar per capita até R$200,00 (corte que foi utilizado para garantir a representatividade amostral em todos os grupos).

Considerando o critério de elegibilidade, os domicílios foram re-classificados em três grupos. O primeiro grupo chamado “Tratamento” (T), constituído pelos domicílios que declararam estar recebendo, na data da pesquisa, o benefício do Bolsa Família. O segundo grupo, denominado de “Comparação 1” (C1), composto pelos domicílios que, na data da pesquisa, estavam recebendo outros benefícios. O último grupo, denominado “Comparação 2” (C2), foi composto pelos domicílios que declararam nunca terem recebido qualquer tipo de benefício, independentemente de serem cadastrados em algum programa público. O restante da amostra não incluída nos grupos de comparação, é constituído pelos domicílios que já haviam recebido algum tipo de benefício, mas que não recebiam mais na data da pesquisa, e de domicílios cuja renda domiciliar per capita era maior que R$200,00. A amostra total com informação válida contém 15.240 domicílios, incluindo 4.435 no grupo de Tratamento, 3.496 no grupo de C1 e 4.941 no grupo de C2, além de 2.368 domicílios não classificados em nenhum dos grupos (OLIVEIRA et al, 2007).

A justificativa para a formação de dois grupos de comparação decorre da possibilidade de se investigar dois tipos distintos de resultados do programa. O primeiro tipo, envolvendo a comparação do grupo de tratamento com o grupo C2, caracteriza-se como um resultado preliminar "puro" do Programa Bolsa Família, na medida em que compara os domicílios beneficiários com domicílios semelhantes em termos de probabilidade de participação no programa, mas que não recebem qualquer tipo de transferência de renda. Na segunda comparação, analisamos os resultados obtidos na amostra de beneficiários do Programa Bolsa Família em relação aos beneficiários de outros programas federais de transferência de renda. Essa análise merece muita cautela uma vez que esse segundo grupo é bastante heterogêneo em termos de transferência de renda e presença de condicionalidades. Por último, vale enfatizar que a análise é baseada na autodeclaração dos domicílios acerca do recebimento dos benefícios de programas sociais.

3.5.3 Resultados da avaliação de impacto

Entre os resultados mais importantes destaca-se o impacto positivo sobre os índices de freqüência e de evasão escolar. A redução dos índices de evasão escolar observada, entretanto, foi acompanhada do aumento do número de reprovações, o que confirma que o programa, ao intervir apenas na demanda, não é capaz, por si só, de impactar positivamente em todos os aspectos educacionais.

Outro resultado foi o impacto positivo do programa Bolsa Família na participação da força de trabalho, sobretudo entre as mulheres. Adultos assistidos pelo programa tiveram participação no mercado de trabalho 2,6% maior do que aqueles não assistidos, sendo a participação das mulheres beneficiadas pelo programa 4,3% maior que a de homens assistidos. Tal fato contraria as críticas feitas aos programas de transferência de renda como o Bolsa Família, segundo as quais tais iniciativas estimulariam as pessoas assistidas a pararem de trabalhar ou a não procurarem trabalho.

3.5.4 Limitações da AIBF:

1. Os resultados que suscitam da aplicação da metodologia, devem ser tomados com cautela na interpretação, pois a metodologia não assegura que atributos não-observáveis

evitem a presença de algum viés na medida de impacto. No entanto, esta é a única medida possível, por não ser possível implementar um desenho experimental.

2. Outra limitação relacionada à interpretação dos resultados é que os diferenciais são captados apenas em um ponto temporal, que não se refere a um momento anterior ao início do programa.

3. Ressalva-se que na aplicação metodológica, o diferencial obtido na linha de base não é uma medida de impacto, isto é, uma medida que possa ser considerada como tal, sem sombra de dúvidas. Para tal conclusão, torna-se necessário conduzir uma segunda rodada de pesquisa, de forma a se construir uma base longitudinal. Ainda assim, deve-se ter cautela, porque não existe um controle sobre o tempo de exposição dos beneficiários ao programa (efeito duração) e nem sobre o valor do benefício recebido durante a totalidade do período (efeito dose). Uma avaliação definitiva do impacto deverá resolver metodologicamente a incorporação destes dois efeitos. O método de pareamento de grupos de tratamento e controle não resolve este problema.

Finalmente, é importante destacar que, apesar das limitações, estas não invalidam os resultados da AIBF nem o rigor técnico da sua execução. Apenas delimita o grau de cautela necessária para a interpretação dos resultados.

3.6 Algumas aplicações empíricas de avaliação de impacto dos programas de transferências condicionadas de renda (TCR) na América Latina.

Os programas TCR criados na América Latina são geralmente identificados como uma nova geração de políticas contra a pobreza. As avaliações confirmam que estas transferências de renda atingem, de fato, os pobres (Zepeda, 2008). As experiências das avaliações de impacto dos programas aplicados no México, Brasil, Colômbia e Nicarágua não só indicam resultados alentadores e eficazes para promover a acumulação de capital humano nas famílias pobres, mas também os avanços conseguidos em matéria de aplicação de métodos de avaliação experimentais e não-experimentais (RAWLINGS e RUBIO, 2003). A seguir, apresentam-se as características e estratégias de avaliação para mensurar o impacto dos principais programas de TCR em México, Colômbia e Nicarágua.

QUADRO 3. 1. Ano de inicio, objetivos e componentes dos benefícios dos programas de transferências condicionadas de renda (TCR) na América Latina e Caribe.

N.	Nombre	País	Ano de inicio	Objetivos	Benefícios		
					Educação	Saúde e nutrição	População objetivo
1	Progresa/Oportunidades (PROP)	México	1997/2002	Melhorar o nível educativo, estado de saúde e nutricional das familias pobres, particularmente de crianças e as mães.	1- Transferência em dinheiro. 2- Apoio para matéria escolar. 3- Fortalecimento de entrega de qualidade de serviços educativo.	1- Transferência em dinheiro. 2- kit básico de serviços de saúde. 3- Educação nutricional. 4- Suplementos nutricionais.	-Educação: Famílias pobres com crianças de 8-18 anos. -Saúde: Mulheres grávidas com filhos em período lactente, Crianças de 4-24 meses e desnutridas entre 2–5 anos.
2	Familias en Acción (FA)	Colômbia	2001	1- Aumentar a inversão em capital humano entre famílias de pobres extremos. 2- Atuar como red de proteção social.	- Transferências em dinheiro, por bimestre.	1- Transferências em dinheiro. 2- Educação em saúde.	-Educação: Famílias pobres com crianças de 7 a 17 anos. -Saúde: Familias pobres com crianças de 0-6 anos que não participam de outros programas.
3	Red de Protección Social (RPS)	Nicarágua	2000	Promover a acumulação de capital humano entre os domicílios em extrema pobreza	1- Transferência em dinheiro. 2- Apoio para material escolar. 3- Incentivos de oferta.	1- Transferência em dinheiro para alimentação. 2- Educação nutricional/saúde. 3- Medidas básicas de saúde para crianças < 5 anos.	- Educação: Famílias com crianças pobres de 6-13 anos. -Saúde: Serviços de atenção destinados a familias pobres com crianças de 0-5 anos.

QUADRO 3. 2. Implementação do programa, método de avaliação de impacto e resultados obtido pelos programas de transferências condicionadas de renda (TCR) na América Latina e Caribe.

N.	Nombre	Implementação do programa	Avaliação de Impacto	
			Método	Principais resultados
1	Progresa/Oportunidades (PROP)	- Comunidade rural com índice de marginalidade maior que 50, com menos de 2,500 habitantes e ter acesso a uma escola primaria, secundária e um centro de saúde. - Dentro das localidades elegíveis, os domicílios beneficiários identificam-se por meio de uma análise discriminatório da renda da família e outras características.	- Desenho experimental com dado de painel: distribuição aleatória de localidade em grupos de tratamento. - Estimador antes - depois, diferença em diferença, e primeira diferença. - Pareamento por escore de propensão (PSM) e regressão descontínua.	- Aumenta das taxas de matrícula escolar, maior freqüência aos consultórios de saúde e uma menor morbidade entre as crianças beneficiárias de 0 a 2 anos. - Melhor nutrição e cuidado preventivo
2	Familias en Acción (FA)	- Municípios que não sejam capitais de departamentos com menos de 100, habitantes. - Municípios de não participem de outros programas e que tenham oferta de serviços educativos e de saúde e bancos. - Municípios com base de dados SISBEN (sistema de informação que identifica aos domicílios pobres e vulneráveis) atualizados. - Familias do nível 1 de SISBEN.	- Desenho não-experimental. - Estimador por seleção observáveis e diferenças em diferenças (DD).	- Nas áreas rurais aumento da freqüência escolar de crianças entre os 7 e 12 anos e 13 e 17 anos; e melhor nutrição em crianças acima de 36 meses. - Nas áreas urbanas o único impacto significativo é o aumento da freqüência à escola secundária.
3	Red de Protección Social (RPS)	- Departamentos e municípios com incidência de pobreza extrema, com acesso a escolas e centro de saúde. - Municípios elegíveis divididos em áreas censais, classificados em 2 grupos segundo um índice de marginalidade. O primeiro grupo participa da Fase piloto 1 (áreas de censo com menos de 14.1 de hectare-as e não tenham veiculo); enquanto o segundo grupo participará na Fase piloto 2 (Elegibilidade do domicílio segundo uma formula de alocação de escores).	- Desenho experimental com dados de painel: distribuição aleatória das áreas censais em grupos de tratamento e controle. - Estimador de diferença em diferença ou dupla diferença.	- Impactos positivos nas crianças entre 7 e 13 anos matriculadas na escola primária. - Maior proporção de crianças menores de 3 anos com controles de crescimento e maior proporção de crianças entre 12-23 meses com todas as vacinas ao dia.

3.7 O relacionamento como alternativa para alocar às famílias segundo o registro administrativo do Cadastro Único.

Na pesquisa da Avaliação de Impacto do Bolsa Família (AIBF), a estratégia da amostragem do projeto se baseou na realização de pesquisa de campo de linha base domiciliar, sem depender do cadastro de famílias do programa, porque este foi avaliado e considerado precário na ocasião do planejamento da pesquisa. Com esta estratégia, tornou-se possível cobrir na pesquisa toda a população de famílias do país, inclusive uma pequena amostra de famílias não elegíveis para o programa (OLIVEIRA et al, 2007).

Neste sentido, foi realizado um *screening*, o qual foi necessário para atualizar o cadastro de domicílios dos setores censitários sorteados no procedimento de amostragem para a Pesquisa de Avaliação de Impacto do Programa Bolsa Família (AIBF). Especificamente, o *screening* teve uma função fundamental de classificar os domicílios segundo três categorias: (1) beneficiários do Programa BF; (2) cadastrados no Cadastro Único (CadÚnico) e/ou beneficiários de outros programas de transferência de renda do Governo Federal; e (3) não-beneficiários não-cadastrados. Este procedimento é crucial para a seleção aleatória dos domicílios nos quais foi aplicado o questionário da Pesquisa AIBF (OLIVEIRA et al, 2007).

Para realizar este *screening* os principais cadastros utilizados foram: Arquivo Agregado de Setores do Censo Demográfico de 2000 (base para seleção das amostras de municípios e setores); Base Operacional Geográfica do Censo Demográfico de 2000 (mapas e descritores das áreas selecionadas, para apoiar coleta) e o Cadastro Único de Beneficiários dos programas do Governo Federal (CadÚnico), mas apenas as informações agregadas, utilizadas para apoio à estratificação da amostra de municípios.

Embora a informação coletada no *screening* sobre o recebimento do benefício pelas famílias entrevistadas seja considerada adequada para análise na AIBF, nas pesquisas de campo, apesar da coleta de dados seguir um conjunto de regra, é possível que as respostas estejam influenciadas por aspetos subjetivos, como opiniões ou atitudes das pessoas. Assim, algumas variações ou diferenças de informação podem alterar a significância estatística dos impactos ou diferenciais dos resultados da avaliação.

Considerando esta ponderação, compete indagar sobre a possibilidade de utilizar o registro administrativo CadÚnico para alocar as famílias ao grupo de tratamento ou controle, segundo este registro, mas considerando algumas ferramentas estatísticas. O registro administrativo CadÚnico caracteriza-se por ser desenhado para registrar informações socioeconômicas das famílias com renda per capita por mês até meio salário mínimo, permitir a identificação das necessidades e características da família e seus membros, utilizar para selecionar beneficiários dos diversos programas sociais e possibilitar a geração de um número único nacional de identificação para os programas sociais (NIS) , evitando duplicidades.

Considerando a possibilidade de utilizar os registros administrativos do CadÚnico, seria preciso uma técnica que identificasse às famílias que foram entrevistadas no AIBF, nos grupos de comparação segundo o registro administrativo CadÚnico, mas simultaneamente nos grupos de alocação segundo os resultados do *screening* (pesquisa de campo AIBF). Uma das técnicas possível é o relacionamento de bases de dados, que foi descrito no capítulo 2, que define a comparação de dois ou mais registros das bases que contêm informações de identificação para determinar se estes registros referem-se à mesma entidade (HOWE, 1988). Com o resultado do relacionamento da base de registro administrativo CadÚnico com a pesquisa de campo AIBF, surge à possibilidade de se estudar e analisar as presumíveis variações ou diferenças dos resultados de impactos ou diferenciais da avaliação, entre ambas as fontes de informação utilizadas para alocar às famílias no grupo de comparação. Além disso, os trabalhos de avaliação de impacto sugerem utilizar várias configurações de informações disponíveis, para realizar a avaliação de um programa, porque os procedimentos de seleção dos beneficiários podem enfrentar uma série de dificuldades e limitações (financeiras e políticas) no momento da implementação do programa e da avaliação (SKOUFIAS, 2006).

4 REALIZANDO O RELACIONAMENTO DE DADOS

Neste capítulo, o objetivo final é encontrar a nova alocação dos domicílios familiares[28] nos grupos de comparação para avaliação do PBF segundo os registros administrativos do CadÚnico. Para conseguir o mencionado objetivo precisa-se, primeiro, recuperar informação do Número de Identificação Social (NIS) de ao menos um integrante do domicílio que foi entrevistado na pesquisa de campo do AIBF, a partir da qual será possível recuperar os benefícios que foram outorgados a esta família no mês da pesquisa de campo, mas segundo os registros administrativos.

Na primeira parte deste capítulo, descreve-se o que está relacionado às duas fontes de dados usadas: a) Dados da pesquisa de campos de domicílios AIBF e b) Dados dos registros administrativos do Cadastro Único (CadÚnico). O capítulo esclarece o desenho amostral da pesquisa de campo; além disso, avalia a consistência da informação e cobertura para ambas as bases e as características especiais do CadÚnico. Seguidamente, descrevem-se as tarefas que devem ser realizadas antes de iniciar o processo do relacionamento. Segundo Gill (2001), nos esforços que se realizam para a implementação do relacionamento de dados, 75% deles centra-se em preparar a base de dados, 5% em conduzir o relacionamento e apenas 20% agrupa-se na avaliação dos resultados do relacionamento. Na parte seguinte deste capítulo, apresentam-se os passos e as tarefas realizadas e os resultados do procedimento do relacionamento dos dados determinístico e probabilísticos entre a base de dados da pesquisa de campo AIBF e os registros administrativos do CadÚnico. Na última seção deste capítulo, apresentam-se os resultados das famílias que foram encontradas com ambos os métodos de relacionamento utilizados e a nova alocação destas famílias nos grupos de comparação segundo os registros administrativos do CadÚnico.

[28] Um domicílio é a moradia onde o relacionamento entre seus ocupantes é ditado por laços de parentesco familiares, de dependência doméstica ou por normas de convivência. Neste caso, também se deve considerar que um domicílio foi considerado como uma unidade familiar, para efeitos de comparação das bases de dados.

4.1 Bases de dados utilizadas

Nesta seção, descrevem-se as duas bases de dados utilizadas na tese. A primeira é proveniente da pesquisa de campo AIBF e, a outra, dos registros administrativos do CadÚnico. Apresenta-se a sua estrutura, definição e descrição dos campos utilizados e uma apresentação em tabelas para descrição estatística e sucinta das variáveis mais relevantes, para familiarizar-se com ambas as bases. A seguir, passa-se a descrever as duas bases de dados utilizadas nesta tese.

4.1.1 Base de dados provenientes da pesquisa de campo AIBF

A estratégia utilizada na pesquisa AIBF foi realizar pesquisa de campo de base domiciliar, sem depender exclusivamente do cadastro de famílias do programa (CadÚnico), que foi avaliado e considerado precário na ocasião do planejamento da pesquisa AIBF. Com esta estratégia, foi possível cobrir toda a população de famílias do Brasil, inclusive uma pequena amostra de famílias não elegíveis para o programa.

Para definir a amostra da pesquisa, os principais cadastros utilizados foram: dados agregados de setores do Censo Demográfico de 2000 (para seleção das amostras de municípios e setores); Cadastro Único de Beneficiários dos programas do Governo Federal - CadÚnico (informações agregadas utilizadas para apoio à estratificação da amostra de municípios); e Base Operacional Geográfica do Censo Demográfico de 2000 (mapas e descritores das áreas selecionadas, para apoiar coleta).

O plano amostral empregado na pesquisa base foi a amostragem em duas fases. Na primeira fase, foi adotada amostragem conglomerada em uma ou duas etapas para seleção de setores censitários, com estratificação. Na segunda fase, foi feita seleção de domicílios por amostragem estratificada simples.

Na primeira fase, o plano amostral realizou uma estratificação geográfica e por tamanho do município. A estratificação dos municípios por tamanho foi feita em dois grupos: os 41 maiores municípios do país, conforme os dados de população do Censo 2000 foram alocados num estrato de “municípios grandes”. O segundo grupo, composto por todos os demais municípios, foi denominado de “municípios pequenos”. A estratificação geográfica

dividiu a população em três grandes áreas: residentes das regiões Norte e Centro-Oeste (N+CO), residentes do Nordeste (NE), e residentes do Sudeste e Sul (SE+S).

Concluída a seleção da amostra de setores, a segunda fase foi implementada após uma operação de *screening* ou varredura para cadastramento de domicílios em cada um dos 1.416 setores selecionados para a amostra. Esta operação buscou localizar, identificar e classificar todos os domicílios encontrados em cada um dos setores selecionados na fase 1. A classificação dos domicílios foi feita usando perguntas contidas na folha de coleta da pesquisa AIBF. Assim, os domicílios foram classificados em três categorias: domicílios com famílias cadastradas no Cadastro Único e beneficiárias do Programa Bolsa Família; domicílios com famílias cadastradas no Cadastro Único, mas ainda não beneficiárias do Programa Bolsa Família (podendo ser beneficiárias de outros programas de transferência de renda); e domicílios sem famílias cadastradas ou beneficiárias.

Considerando que teria uma perda de parte da amostra de domicílios por motivos diversos, a amostra inicialmente selecionada alcançou o total de 16.993 domicílios, mas a amostra final disponível, depois de descontadas as perdas por diversas razões ocorridas durante a operação de campo, ficou com um total de 15.426 domicílios com entrevistas completas. Desta forma, a meta inicial de ter uma amostra total de aproximadamente 15.000 domicílios foi cumprida. A TAB 4.1 apresenta as contagens finais de domicílios e pessoas na amostra coletada considerada disponível para as análises, por grandes áreas.

TABELA 4.1 – Contagens de domicílios e pessoas na amostra de domicílios com entrevista completa, por grande área.

Área	Domicílios com entrevista completa	Pessoas na amostra	Pessoas por domicílio
N+CO	4.433	21.314	4,8
NE	5.106	23.008	4,5
SE+S	5.887	25.360	4,3
Total	15.426	69.682	4,5

Fonte: Coleta de dados da pesquisa AIBF, 2005.

Na TAB 4.2, mostram-se às contagens finais de domicílios e pessoas na amostra coletada considerada disponível para as análises, segundo tipo de domicilio.

TABELA 4.2 – Contagens de domicílios e pessoas na amostra de domicílios com entrevista completa, por estrato de seleção dos domicílios.

Estrato	Domicílios com entrevista completa	Pessoas na amostra	Pessoas por domicílio
Casos	4.588	22.686	4,9
Controles tipo 1	9.036	41.068	4,5
Controles tipo 2	1.802	5.928	3,3
Total	**15.426**	**69.682**	**4,5**

Fonte: Coleta de dados da pesquisa AIBF, 2005.

Tomando em conta a classificação dos domicílios segundo o critério de elegibilidade, em que ponderou os cortes de renda domiciliar per capita até R$50,00, R$100,00 e R$200,00 tal como foi visto na seção 3.5, a amostra por grupos é de 4.588 no grupo de Tratamento (casos) , 9.036 no grupo de Comparação 1 – C1 (controles tipo 1) e 1.802 no grupo de Comparação - C2 (controles tipo 2).

Uma vez definida a amostra total, a seguir apresenta-se a composição da base de dados final, que está distribuída em três sub-bases, tal como se mostra no quadro a seguir:

QUADRO 4.1 – Composição final da Base de dados segundo sub-bases, seções incluídas do questionário e número de campos[29].

Sub-Bases	Descrição da base	Seções incluídas do questionário		
Domicílios	Estão contidos todos os dados levantados ao nível do domicílio, inclui, além da identificação do questionário.	01 04/c 04/d 10/a	10/b 10/c 11/a 11/b	11/c 12/a
Pessoas	Refere-se a todos os dados levantados pessoa a pessoa	02/a 02/b 03/a 03/b 04/a	04/b 04/e 05/a 05/b 06/a	07/a 07/b 12/b
Benefícios	Dados das pessoas que recebiam benefício de algum programa social.	12/c.		

Fonte: Coleta de dados da pesquisa AIBF, 2005.

[29] No ANEXO V apresentam-se algumas secções do questionário aplicada na pesquisa de campo AIBF.

4.1.2 Base de dados provenientes do registro administrativo CadÚnico.

O registro administrativo do Cadastro Único para Programas Sociais (CadÚnico) é um instrumento fundamental para identificar as famílias mais pobres do país, para conhecer suas vulnerabilidades e potencialidades, e para subsidiar a elaboração e implementação de políticas públicas destinadas a essas famílias. O CadÚnico foi criado em 2001, com o propósito de unificar os cadastros e a concessão de benefícios dos programas federais focalizados com caráter permanente. O CadÚnico pode contribuir, por meio das informações por ele disponibilizadas, para a construção e acompanhamento de políticas públicas que transformem a situação socioeconômica, reduzindo pobreza e desigualdade e promovendo uma maior eqüidade na sociedade brasileira (BRASIL, 200-?c).

O CadÚnico permite a concessão de benefícios do Bolsa Família, orienta o desenho e a implantação de políticas públicas, de responsabilidade de diferentes esferas de governo, voltadas para as famílias de baixa renda, quando possível, como foi mencionado no capítulo 3. Quando se identificam as características sócio-econômicas das famílias, é possível caracterizar melhor várias dimensões de pobreza e vulnerabilidade para além do rendimento monetário. O CadÚnico permite, ainda, identificar, por meio de variáveis multidimensionais, as famílias mais vulneráveis, prioritárias para acompanhamento familiar, e aquelas que podem, segundo suas características, ser incluídas em programas complementares ao Programa Bolsa Família (BARROS et al, 2002; RAMOS e SANTANA, 2002). Assim, o Cadastro Único compõe-se por três núcleos básicos de informações[30]:

• **Identificação da pessoa** (gera um número único, atribuindo a cada membro das famílias cadastradas um Número de Identificação Social (NIS) para os programas sociais, evitando duplicidade): nome completo, nome da mãe, data de nascimento, município de nascimento, algum documento de emissão nacional (CPF ou TE)

• **Identificação do endereço**.

• **Caracterização sócioeconômica:** composição familiar (número de pessoas, gestantes, idosos, portadores de deficiência), características do domicílio (número de cômodos, tipo

[30] No ANEXO VI apresenta-se o questionário do Cadastro Único – Domicílios e Pessoas.

de construção, água, esgoto e lixo), qualificação escolar dos membros da família, qualificação profissional e situação no mercado de trabalho, rendimentos e despesas familiares (aluguel, transporte, alimentação e outros).

Outra característica importante do CadÚnico é que este registro administrativo pode ser associado com uma Pesquisa de Campo Domiciliar, isto devido ao fato de o levantamento de dados abranger um conjunto de informações individuais e familiares, além de levantar dados sobre as condições de vida. Ou seja, não são levantadas unicamente informações úteis para um tipo de programa ou programas, mas também, contempla informações mais amplas, as quais são úteis para avaliar problemas sociais (BARROS et al, 2002). Este ponto é de relativa importância: a unidade pesquisada não é cada indivíduo isoladamente, senão o conjunto do ambiente familiar (RAMOS e SANTANA, 2002).

4.2. Descrição de algumas variáveis utilizadas para o relacionamento da base AIBF e CadÚnico.

Nesta primeira parte, descrevem-se algumas características gerais dos indivíduos, com o objetivo de conhecer, de forma geral ambas as bases utilizadas. Esta descrição não pretende ser analítica, mas sim informativa como forma de conhecimento das bases que são utilizadas nesta tese, como também, preparando a informação com que se conta para o relacionamento de bases de dados. Neste sentido, descrevem-se a seguir os indivíduos segundo distribuição por região, sexo, parentesco com o responsável pelo domicilio, idade, estado civil e raça.[31]

Distribuição de pessoas e domicílios segundo Região.

Sobre as pessoas, temos a distribuição por número de pessoas que foram entrevistadas na pesquisa AIBF e distribuição de pessoas cadastrados no CadÚnico. Analisando a distribuição de pessoas segundo Região, temos que o maior porcentagem de pessoas estão na região nordeste, tanto segundo a pesquisa AIBF como no CadÚnico (33% e 41% respectivamente), a segunda região com maior porcentagem de pessoas é a região Sudeste, seguida por Norte, Centro-Oeste e por último o Sul; esta distribuição é similar na amostra

[31] As variáveis utilizadas na avaliação de impacto serão descritas junto aos resultados da avaliação, isto é, capitulo 5.

da pesquisa AIBF e no CadÚnico. Quando se analisam os domicílios segundo Região, observa-se também a mesma distribuição que a das pessoas, isto é, maior proporção de domicílios no Nordeste e menor proporção na região Sul.

TABELA 4.3 – Distribuição de pessoas e domicílios por regiões segundo pesquisa AIBF e CadÚnico. Brasil. 2005.

Região	AIBF				CadÚnico			
	Pessoas		Domicílios[5]		Pessoas		Domicílios*	
	Casos	%	Casos	%	Casos	%	Casos	%
Norte	12.203	18,25	2.443	15,84	1.483.065	13,38	385.358	12,49
Nordeste	22.085	33,02	5.106	33,10	4.614.054	41,62	1.306.247	42,33
Centro-Oeste	8.028	12,00	1.990	12,90	1.055.964	9,53	288.838	9,36
Sudeste	20.663	30,90	4.913	31,85	3.300.502	29,77	934.265	30,27
Sul	3.902	5,83	974	6,31	631.820	5,70	171.312	5,55
Total	66.881	100,00	15.426	100,00	11.085.405	100,00	3.086.020	100,00

* Domicílio é o local estruturalmente separado e independente que se destina a servir de habitação a uma ou mais pessoas, ou que esteja sendo utilizado como tal.
Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

Quando se compara o resultado de ambas as fontes de dados, observa-se que a porcentagem de pessoas (e domicílios) na Região Nordeste, segundo o CadÚnico é maior em aproximadamente 8 pontos porcentuais, comparado com a porcentagem da amostra da pesquisa AIBF (Ver TAB 4.3). Este resultado pode ser explicado, pelo fato do CadÚnico registrar as famílias em situação de extrema pobreza e, segundo estudos de IBGE (Pesquisa Nacional por Amostra de Domicílios – PNAD, 2005), a maior proporção de pessoas pobres se encontram na Região Nordeste, então existe uma alta probabilidade que as pessoas que mais são cadastradas no CadÚnico sejam desta região. Embora a amostra da pesquisa AIBF considerasse uma ponderação da base operacional geográfica do Censo 2000, parece que o fato de trabalhar com todas as informações no CadÚnico, gera maior probabilidade de trabalhar com famílias em extrema pobreza.

Distribuição de pessoas segundo sexo.

Em relação à variável sexo, observa-se uma maior participação feminina na amostra da pesquisa AIBF e do registro administrativo do CadÚnico, assim como um toda a porcentagem de mulheres na amostra AIBF está em torno de 52% e no CadÚnico 57% (ver TAB 4.4).

Comparando a proporção de mulheres entre ambas as fontes de informação, observa-se que esta é maior no CadÚnico, isto porque as pessoas que são titulares ou responsáveis do PBF segundo o CadÚnico são mulheres, uma vez que a titularidade do cartão é concedida preferencialmente às mulheres (MDS, 2007). Portanto, nos registros do CadÚnico encontra-se uma porcentagem maior de mulheres, em comparação à pesquisa de campo AIBF, que registrou todos os membros da família, sem ter cotas por sexo (OLIVEIRA et al, 2007).

TABELA 4.4 – Distribuição por sexo das pessoas integrantes dos domicílios segundo pesquisa AIBF e CadÚnico. Brasil. 2005.

sexo	AIBF		CadÚnico	
	Freqüência	%	Freqüência	%
Feminino	34.505	51,59	6.271.096	56,57
Masculino	32.376	48,41	4.814.279	43,43
Total	66.881	100,00	11.085.375	100,00

Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

Distribuição das pessoas segundo parentesco com o responsável pelo domicilio.

Analisando a distribuição segundo relação de parentesco, a amostra da pesquisa AIBF e o registro CadÚnico indicam uma maior participação relativa de filhos(as) ou enteados(as), seguidos pelos chefes de famílias, cônjuges ou companheiros(as) e neto(a) ou bisneto(a) entre as principais categorias da relação de parentesco. Considerando as três primeiras categorias segundo porcentagem de participação, é possível dizer que as famílias em ambas as fontes de dados caracterizam-se como famílias nucleares.

TABELA 4.5 – Distribuição por relação de parentesco da família das pessoas integrantes dos domicílios segundo pesquisa AIBF e Cadastro CadÚnico. Brasil. 2006.

Parentesco	AIBF		CadÚnico	
	Freqüência	%	Freqüência	%
Chefes de famílias	15.098	22,57	3.075.285	27,74
Cônjuge, companheiro(a)	10.668	15,95	1.398.361	12,61
Filho(a), enteado(a)	33.241	49,70	5.213.626	47,03
Pai, mãe, sogro(a)	537	0,80	1.792	0,02
Neto(a), bisneto(a)	4.753	7,11	721.482	6,51
Irmão, irmã	585	0,87	170.767	1,54
Nora, genro	602	0,90	41.080	0,37
Outro parente	1.116	1,67	461.380	4,16
Agregado	229	0,34	693	0,01
Pensionista	12	0,02	0	0,00
Empregada doméstica	31	0,05	0	0,00
Parente de empregada doméstica	9	0,01	0	0,00
Sem dado	0	0	939	0,01
Total	66.881	100,00	11.085.405	100,00

Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

Distribuição das pessoas segundo idade.

Em relação à idade, observa-se que, no Brasil, a idade média da amostra segundo a pesquisa de campo AIBF é de 26,18 anos, enquanto segundo os registros administrativos CadÚnico verifica-se que esta média é menor em 3 anos. No caso da mediana, 50% das pessoas segundo a amostra da pesquisa AIBF são menores de 20 anos e 18 anos no CadÚnico. Resultados similares são encontrados quando se analisa os quartis, em que os valores são menores segundo o CadÚnico. Este comportamento era esperado porque no CadÚnico pressupõe-se que estão registradas famílias em extrema pobreza, as quais possuem maior número de filhos pequenos (RAMOS e SANTANA, 2002). Além disso, considerando a amostra AIBF, na qual uma parcela das famílias não necessariamente está

em condições de pobreza (com renda acima de R$200) e que podem ser famílias com menor número de crianças. Estes dois argumentos poderiam estar influenciando a idade mediana e quartis de ambas das fontes de dados.

TABELA 4.6 – Descrição da idade das pessoas integrantes dos domicílios segundo pesquisa AIBF e CadÚnico. Brasil. 2006.

Idade	AIBF	CadÚnico
Média	26,18	23,63
Mediana	20,00	18,00
Q1	10,00	11,00
Q2	20,00	18,00
Q3	37,00	35,00
Desvio padrão	39,73	16,16

Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

Distribuição das pessoas segundo estado civil.

Considerando o estado civil, observa-se uma maior concentração de solteiros e casados tanto na pesquisa de campo AIBF, como nos registros administrativos CadÚnico. Analisando as outras categorias do estado civil, as proporções são similares em ambas das fontes de dados. Comparando a proporção de casados entre a amostra da pesquisa AIBF e o CadÚnico, observa-se que esta é maior em aproximadamente 22 pontos porcentuais na AIBF. Para explicar estes resultados deve-se estar atento que é razoável, em primeiro lugar, que a alta porcentagem de casos sem dados possa estar influenciando nos resultados, já que as outras categorias apresentam proporções similares em ambas as fontes de informação. Em segundo lugar, relacionado à primeira proposição, está a forma em que se define a categoria de "casado". No caso de AIBF esta categoria inclui casado no civil e religioso, casado só no civil, casado só no religioso e união consensual, enquanto no CadÚnico esta categoria não é explicitada. Ambos os aspectos podem estar na origen da subestimação, de uma forma ou outra, da proporção de pessoas casadas segundo CadÚnico.

TABELA 4.7 – Distribuição por estado civil das pessoas integrantes dos domicílios segundo pesquisa AIBF e Cadastro CadÚnico. Brasil. 2006.

Estado Civil	AIBF		CadÚnico	
	Freqüência	%	Freqüência	%
Solteiro	40.385	60,38	7.398.299	66,74
Casado*	13.594	34,10	1.438.912	12,98
Divorciado	9.549	0,50	52.256	0,47
Desquitado/separado judicialmente e de fato	1.562	2,34	165.206	1,49
Viúvo	1.651	2,47	126.608	1,14
Sem dado	140	0,21	1904124	17,18
Total	66.881	100,00	11.085.405	100

* Para o caso da AIBF, nesta categoria estão incluídos casado no civil e religioso, casado só no civil, casado só no religioso e união consensual.
Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

Distribuição das pessoas segundo raça.

Considerando a variável raça para todo Brasil, na amostra da pesquisa AIBF e nos registros do CadÚnico os pardos são maioria (55% e 60% respectivamente). As outras categorias de raças, que concentram significativas proporções de pessoas, são os brancos e pretos. No entanto, no cadastro CadÚnico é ligeiramente maior nas categorias de preto e pardos. De uma maneira geral, tanto a amostra da pesquisa AIBF quanto os registros administrativos focalizam a população não-brancos, principalmente pardos.

TABELA 4.8 – Distribuição por raça das pessoas integrantes dos domicílios segundo pesquisa AIBF e Cadastro CadÚnico. Brasil. 2006.

Parentesco	AIBF			CadÚnico		
	Freqüência	%	% valido	Freqüência	%	% valido
Branca	22.095	33,04	33,28	2.419.655	21,83	27,93
Preta	6.624	9,90	9,98	957.612	8,64	11,06
Parda	36.981	55,29	55,71	5.212.025	47,02	60,17
Amarela	493	0,74	0,74	40.739	0,37	0,47
Indígena	194	0,29	0,29	32.107	0,29	0,37
Sem dado	494	0,74		2.423.267	21,86	
Total	66.881	100		11.085.405	100	

Fonte: Coleta de dados da pesquisa AIBF e CadÚnico 2005.

4.3 Preparando o relacionamento.

Antes de começar a realizar o processo de relacionamento, diversas edições dos campos (variáveis) das bases de dados precisaram ser feitas, porque, às vezes, os dados registrados ou captados apresentaram-se com diferentes formatos e classificações, contendo informações faltantes ou com erros. Neste sentido, a etapa de pré-processo de relacionamento tem como objetivo editar e padronizar os dados (etapa também chama de limpieza). A seguir, apresentam-se os possíveis problemas que podem ser encontrados nas bases de dados e quais são os procedimentos que podem ser utilizados para solucionar tais problemas.

4.3.1 Erros típicos nas variáveis de comparação.

Muitos erros nas variáveis escolhidas para o relacionamento acontecem durante o registro ou cadastramento e processamento das variáveis por parte dos administradores das bases de dados. Entre os principais erros encontrados nestas variáveis incluem: variação ortográfica, codificação e preparação dos dados, freqüência de “apelidos” nos nomes, nomes estrangeiros, uso de iniciais na variável nome, abreviação nas variáveis literárias,

utilização de nomes compostos, palavras faltantes ou extras (GILL, 2001). Entre os principais erros encontrados nas variáveis de comparação destacam-se:

Identificador único numérico: Quando este identificador é disponível pode ser uma variável adequada de comparação. Não obstante, possíveis erros podem ser encontrados, tais como: identificadores faltantes para alguns registros; inversão de dígitos; mesmo número identificador para mais de um registro (um exemplo de duplicados); ou as unidades podem recorrer a identificadores diferentes em bases diferentes.

Sobrenome: Os sobrenomes podem ter mudado, devido a casamentos ou divórcios, o que se torna o problema principal nesta variável. Em algumas sociedades estes erros são causados pelo aumento (ou retirada) de um sobrenome, mudança da ordem dos sobrenomes e utilização de sobrenomes compostos. Outro problema que comumente se encontra é a variação de ortografia dos sobrenomes, originada pelo efeito da transcrição destes através de vários sistemas de administração de bases de dados.

Primeiro nome: Um dos erros freqüentemente encontrado são as amplas variações na ortografia do primeiro nome, originadas pelo registro e transcrição com erros. Além deste problema, inclui-se também a utilização de apelidos e contrações que, muitas vezes, são identificáveis e em outras ocasiões não são. Um caso com o qual também se depara são registros que pertencem a recém-nascidos ou crianças pequenas, os quais têm anteposto ao nome o termo “Bebê” ou “Gêmeo”.

Endereço: Variável utilizada para confirmar pares de registros com incertezas, embora as divergências e rigidezes encontradas dificultem a sua utilização. Os erros neste caso são ocasionados pela mudança de endereços dos indivíduos; variações dos nomes dos endereços residenciais; e diferenças entre o endereço registrado nos registros administrativos e físicos ou encontrado na residência (WINKLER, 1993?).

Sexo: esta variável é geralmente bem reportada, exceto quando existem erros na transcrição e armazenamento dos computadores, mas continua sendo uma variável altamente confiável. Segundo Gill (2001), as possíveis dificuldades desta variável são originadas pelo não registro desta variável em alguns registros administrativos ou a existência de programas de ingressos de dados que geram uma variável de “sexo” através do primeiro nome, a qual não é completamente exata.

Data de nascimento: esta variável geralmente também é bem reportada; no entanto, alguns erros podem ser encontrados, quando a data de nascimento é provida por outras pessoas, por exemplo, no caso das crianças e dos idosos. Outros erros freqüentemente são encontrados na transcrição, quando se inverte o dia por mês, ou quando se invertem os dígitos nos anos[32] .

Adicionar títulos nos nomes: A muitos sobrenomes e primeiros nomes, antepõe-se títulos como Sr., Srª, Dr, Jr. Estes títulos deveriam ser eliminados ou separados antes de utilizar os nomes para o relacionamento.

4.3.2 Padronização: edição, análise gramática, formatação, concordância.

A padronização das variáveis é um processo importante para o relacionamento de dados. Os problemas da qualidade potencial dos dados determinam que algumas variáveis possam não ser satisfatórias para a utilização do relacionamento. O objetivo do exercício de padronizar é principalmente minimizar os erros. Entre os principais exercícios de padronização temos:

• Edição: é o processo de detectar e lidar com dados errôneos ou suspeitosos.

• Análise gramatical de um campo (ou variável): separa as entidades dentro do campo, para tornar a comparação mais fácil. Por exemplo, a variável que apresenta o nome do indivíduo contém primeiro nome e sobrenome; segundo essa análise estas deve ser separada em dois campos diferentes.

• Formatação: é o exercício necessário quando os campos são registrados em formatos diferentes, por exemplo, a data de nascimento "01Jan2002" em um arquivo e "010102" no outro arquivo de dados.

• Codificação consistente por arquivos (ou concordância): é um processo importante para as variáveis que requerem classificação, como, por exemplo, o sexo codificado como 1 e 2 em um arquivo e com as letras M e F codificados em outro.

---

[32] Por exemplo, a data de nascimento correta 10/12/1986 pode ser registrada como 12/10/1986 ou 10/12/1968

4.3.2.1 Edição:

A limpeza básica é necessária antes do relacionamento, porque através desta removem-se erros definidos na primeira exploração das variáveis comuns nas bases de dados. A edição ou revisão deveria ser realizada para identificar respostas inválidas, tais como caracteres string em variáveis numéricas, ou caracteres não alfanuméricos como "#", "$" ou "^" em respostas com caracteres de textos. Outra revisão pode ser feita para "valores fora de intervalos" ou respostas impossíveis, tais como data de nascimento com data futura. Quando se apresentam os casos descritos anteriormente e não existe forma de recuperar a informação correta, tratam-se estes casos como respostas faltantes ou *missing*.

4.3.2.2 Análise ortográfica e padronização das variáveis de relacionamento.

Este procedimento envolve a identificação da estrutura das variáveis de relacionamento e a representa em padrões comuns, de tal forma que possam ser utilizadas em tabelas, sistemas léxicos e codificações fonéticas (GILL, 2001). Desta forma, pode-se dizer que os elementos individuais padronizados são rearranjados em uma ordem comum e adequados. A padronização e análise ortográfica mais comum são as referentes ao nome da pessoa e endereço, as quais são explicadas a seguir.

Padronização de sobrenomes e primeiro nomes.

A padronização básica para este tipo de variável consiste em, primeiro, substituir muitas variações de ortografia e abreviaturas dos nomes e endereços por uma ortografia padrão e abreviações fixas; e, segundo, utilizar palavras-chave geradas durante o processo de padronização como sugestão para o desenvolvimento da análise gramatical das seqüências dos dados. Para o trabalho em questão, o objetivo de padronizar os nomes é permitir que o relacionamento das bases de dados utilizadas seja feito de uma forma mais eficiente e consistente.

Codificação fonética dos nomes.

Os nomes (e sobrenomes) são os identificadores mais difíceis no processo de relacionamento. Estas variáveis além das possibilidades de erros na entrada de dados, apresentam variações na ortografia e inversão de nomes, as quais são contidas nas bases de dados. Quando estes problemas não podem ser corrigidos, existe uma perda potencial de

uma fração significativa de pares de registros que poderiam ter sido pareados caso os erros tivessem sido corrigidos adequadamente.

A criação de sistema de codificação fonética é uma tentativa de direcionar o problema de uma variação da ortografia do mesmo nome, por exemplo, Antono e Antonio. Os códigos criados podem ser utilizados como alternativa dos nomes no processo de relacionamento de dados, assim, estes ajudam a reduzir a fração dos não-pares devido a erros nos nomes.

Dois sistemas de codificação fonéticos são geralmente usados: o sistema de codificação *Soundex* criado por Russell e Odell (KNUTH, 1973) e o Sistema de Informação de Inteligência Estatal de Nova Iorque (NYSIIS), publicado em 1970. Na estratégia do relacionamento probabilístico, estes códigos são utilizados com maior freqüência na criação dos blocos de subconjunto de registros nos arquivos a serem comparados. Desta forma, só os pares de registros formados a partir dos blocos de relacionamento nos arquivos são comparados e outros pares são ignorados. Com a escolha de uma boa variável de blocagem esta estratégia pode reduzir drasticamente o número de pares de registro a serem comparados no relacionamento, possibilitando significante poupança de tempo. No trabalho, será utilizada uma adaptação do sistema de codificação de *Soundex* para a língua portuguesa realizadas pelo Camargo e Coeli (2002).

O sistema de codificação *Soundex*.

A utilização dos códigos fonéticos do nome (primeiro e/ou último nome) é uma alternativa comumente utilizada, já que as chaves apresentam múltiplos valores com uma ocorrência de erros menor do que a seria esperada com a utilização direta do primeiro e/ou do último nome (CAMARGO e COELI, 2007). O *Soundex* é um dos códigos fonéticos freqüentemente utilizados. Este código é constituído por quatro dígitos: o primeiro representa a primeira letra da palavra a ser codificada, enquanto os outros três dígitos são representados por códigos numéricos segundo regras que buscam minimizar erros (por exemplo, eliminação de vogais e substituição de consoantes com sons similares por um código numérico comum) (NEWCOMBE et al., 1988). Por exemplo, o *Soundex* de Afonso é A152, enquanto o de José é J200. No Brasil o software que permite definir campos chaves para blocagem baseados na utilização da função *Soundex* (SOUNDEX (nome do campo)) é o *RecLink*.

Segundo Camargo e Coeli, 2002, quando se trabalha com bases de dados do Brasil encontra-se um problema de inadequação do código *Soundex* para alguns nomes brasileiros que apresentam variações de grafia da primeira sílaba para um mesmo som (por exemplo, Helena x Elena; Jorge x George), nomes que são mais sujeitos aos erros de registro. Como o código *Soundex* guarda a primeira letra do nome, as diferentes grafias recebem códigos diferentes, sendo conseqüentemente alocadas em forma diferente, o que aumenta a probabilidade da perda de pares verdadeiros. Devido a isto, Camargo e Coeli, em 2002, na implementação do software de *RecLink* acrescentaram uma rotina de padronização na "Subdivisão do nome", criando dois campos adicionais relativos ao primeiro e último nomes nos quais a primeira sílaba é modificada segundo as seguintes transformações:

- Primeira letra W e segunda A -> Primeira letra passa a V

- Primeira letra H -> Apagar primeira letra

- Primeira letra K e segunda A, O ou U -> Primeira letra passa a C.

- Primeira letra Y -> Primeira letra passa a I

- Primeira letra C e segunda E ou I -> Primeira letra passa a S

- Primeira letra G e segunda E ou I -> Primeira letra passa a J

Padronização de endereço.

Esta padronização opera de forma similar ao padrão dos nomes, por exemplo, abreviações como "R." ou "Av." deveriam ser substituídas pela apropriada expansão destas palavras "Rua" ou "Avenida" ou considerar uma abreviação padrão comumente utilizada pelas organizações estatais e privadas.

4.3.2 *Software* utilizado para o relacionamento de bases de dados.

A nível mundial, existe uma variedade de instituições comerciais, governamentais, educacionais e privadas que oferecem *softwares* de relacionamento probabilístico de bases de dados. Em nosso caso, foi utilizado o software denominado *RecLink* II, desenvolvido por Camargo e Coeli (1998-2002). O *software RecLink* foi desenvolvido na linguagem C++ com o ambiente de programação Borland C++ Builder versão 3.0 (*Borland*

International Inc., 1998a; *Reisdorph*, 1998). O programa é uma interface com bases de dados flexíveis que permite designar, de modo interativo, as regras de associação entre duas bases.

O processo do *RecLink* opera em dois níveis: no primeiro, criam-se blocos de registros (*Blocking*), como, por exemplo, o código *Soundex* dos campos selecionados (em princípio, contendo nomes) e, dentre os registros blocados segundo mesmo código, outras variáveis (denominadas pareamento, variando de uma a três) podem ser utilizadas para atribuir peso numérico à associação dos registros. No segundo nível, na atribuição de pesos, três algoritmos diferentes podem ser utilizados na comparação das respectivas variáveis: a comparação pura e simples, que somente retorna o valor verdadeiro caso o conteúdo seja rigorosamente idêntico; a comparação de seqüências de caractere a caractere e a comparação aproximada (CAMARGO e COELI, 2002).

O programa foi avaliado a partir dos dados coletados por um dos autores (COELI, 1998) para a realização de estudo que tem como objetivo avaliar a factibilidade para a implantação de sistema de vigilância do diabetes mellitus na população idosa residente na Área Programática 2.2 da cidade do Rio de Janeiro (CAMARGO e COELI, 2002).

4.4 O processo de pré-relacionamento de dados.

4.4.1 Identificação de duplicados.

Neste processo, removem-se os registros que pertencem à mesma entidade, dentro do mesmo arquivo de dados. Às vezes, se aceita certo nível de registros duplicados para planejamento e propósitos de pesquisas, mas recomenda-se remover os duplicados dos arquivos antes que o relacionamento se inicie, pois conservá-los pode complicar o relacionamento das bases de dados.

No caso do relacionamento dos dados utilizado neste trabalho, a base de dados proveniente da pesquisa de campo AIBF não apresentou nenhum registro duplicado, garantindo que a base de dados passou por uma adequada consistência. Na exploração dos dados do registro administrativo do CadÚnico, encontraram-se registros duplicados, como conseqüência da coleta, classificação e consistência dos dados desta base ser realizada pela instituição do

governo que administra e manipula os dados e que pode ter problemas nos diferentes processos que as bases suportam.

A seguir, apresentam-se os casos duplicados segundo as regiões trabalhadas:

TABELA 4.9 – Casos duplicados na base de dados do Registro Administrativo do CadÚnico. Brasil. 2006.

Regiões	Total de casos do CadÚnico	Total de casos sem duplicados do CadÚnico	% de casos duplicados
Norte	6.414.866	5.581.690	12,99
Nordestes	35.991.884	28.075.258	22,00
Centro-Oeste	3.279.262	2.875.596	12,31
Sudeste	19.839.466	16.418.989	17,24
Sul	8.047.575	7.172.828	10,87
Total	73.573.053	60.124.361	18,28

Fonte: Tabela elaborada com os dados do registro administrativo do CadÚnico. MDS. 2006

Ao observar a TAB 4.9., encontrou-se que a porcentagem de casos duplicados dos registros do CadÚnico em todo Brasil está em torno de 18%. Esta porcentagem indica que foi importante considerar esta etapa antes de iniciar o processo de relacionamento das bases de dados, porque dada a significativa porcentagem de casos duplicados, problemas no relacionamento teriam ocorrido e complicado o trabalho. Além disso, como se está trabalhando com grandes volumes de dados, as bases de dados combinadas como resultado de relacionamento apresentaria maior volume do que apresentou sem considerar os duplicados. Estes resultados confirmam a importância de identificar duplicados no presente estudo.

4.4.2 Variáveis comuns em ambas as bases.

Depois de realizar a primeira exploração dos dados de ambas as bases de dados para o relacionamento (seção 4.2), foram conferidas as informações e variáveis comuns a elas. As variáveis detectadas como comuns, foram aquelas que apresentavam na sua informação o mesmo conteúdo, independente do formato ou tamanho ser diferente. Na TAB 4.10,

mostram-se as variáveis comuns nas bases de dados. Esta etapa é útil para familiarizar-se com as bases e as variáveis ou campos que serão úteis para o relacionamento determinístico ou exato e o probabilístico.

TABELA 4.10 – Variáveis comum na base da pesquisa AIBF e CadÚnico. Brasil. 2006.

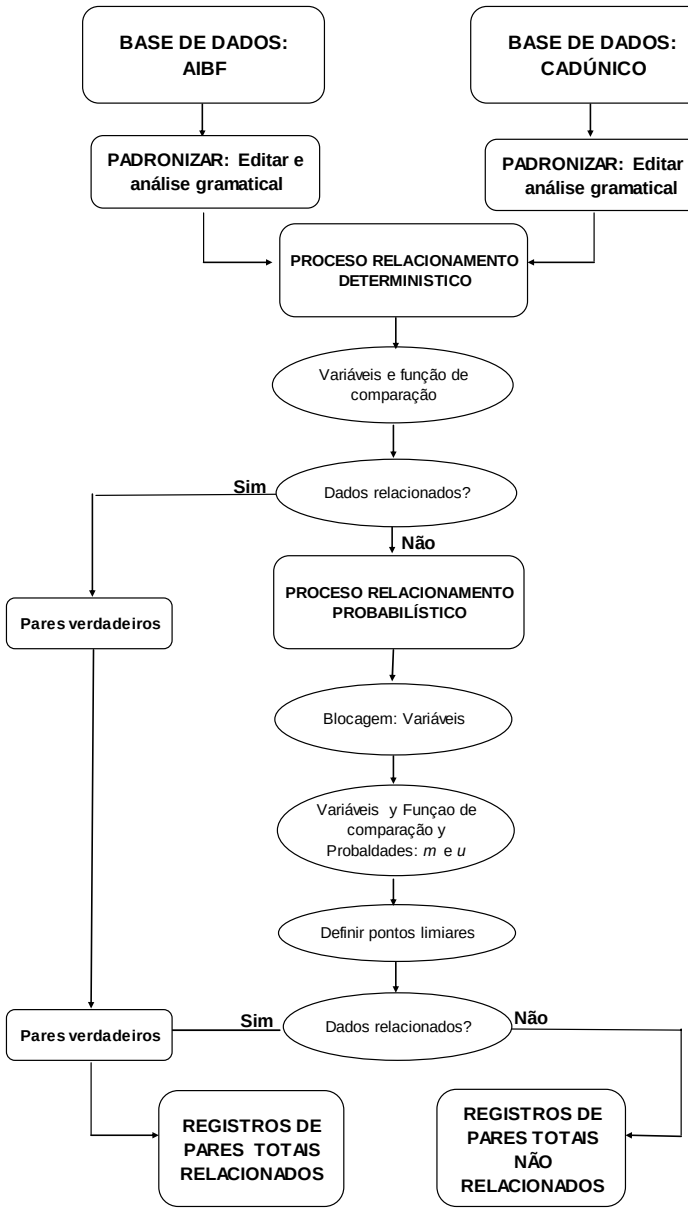
	Variável	AIBF		CadÚnico	
1	Número de identificação social (NIS)				
		Numérico 11 dígitos		Caractere[+] 11	
2	Nome e sobrenome da pessoa				
		Caractere 30		Caractere 70	
3	Sexo				
		0 Ignorado		- Ignorado	
		1 Feminino		F Feminino	
		2 Masculino		M Masculino	
4	Data de nascimento				
		Dia/Mês/Ano (dd/mm/aaaa)		MêsDiaAno (mmddaaaa)	
5	Município do domicílio				
		Numérico 7 dígitos (Código IBGE).		Caracteres 7 (Código IBGE).	
6	Idade do indivíduo				
		Numérico 3 dígitos		Numérico 3 dígitos*	
7	Ordem do parentesco com o chefe de família				
		Numérico 2 dígitos		Caracteres 2	
8	Identificação do domicilio				
		Identificação do setor	Numérico 8	Código domiciliar	Caractere 9
		Estrato de seleção	Numérico 8	Identificação da família	Caractere 15
		Número de questionário	Numérico 8	Identificação do domicílio.	Caractere 15
9	Endereço do domicílio				
		Endereço_c18a	Numérico 8	Tipo de Logradouro	Caractere *3*
		Endereço_c18b	Numérico 8	Nome de Logradouro	Caractere *50*
		Endereço_c18c	Numérico 8	Número de Logradouro	Caractere *15*
				Complemento	Caractere *53*

+ Caractere: é uma ordem de seqüências de símbolos. Estes símbolos são escolhidos de um conjunto pré-determinado ou do Alfabeto.
* Variável obtida com a diferença: (data referência da pesquisa – data de nascimento da base de CadÚnico).
Fonte: Informação tomadas dos dados da pesquisa AIBF e CadÚnico 2005

4.4.3 O fluxo do processo de relacionamento

Nesta seção, apresentam-se os passos que serão realizados no processo de relacionamento de dados determinístico e probabilístico e que serão descritos nas seções seguintes. A seguir, apresenta-se o diagrama de fluxo do processo de relacionamento de bases da pesquisa de campo AIBF e o registro administrativo CadÚnico:

QUADRO 4.2 – O diagrama de fluxo do processo de relacionamento: Determinístico e Probabilístico.



4.5 Padronização das variáveis.

Nos casos em que as bases não são padronizadas, existe a possibilidade que registros que são pares verdadeiros não sejam relacionados, porque variáveis comuns podem aparecer tão diferente que o peso pode mostrar-se menor ou negativo. Este processo é fundamental para os campos não estruturados como nome e sobrenome. Para o presente estudo, as variáveis que serão padronizadas são as seguintes:

A) Variáveis comuns com a mesma estrutura em ambas as bases de dados[33]

1. Número de identificação social (NIS): para padronizar esta variável utilizou-se uma regra prática, numérica de 11 dígitos.

Variável	Código*		Código padronizado
	AIBF	CadÚnico	
NIS	N11	C11	N11

*Daqui por diante considera-se à variável com formato Numérico como "N" e Caractere como "C"

2. Nome completo: para esta variável foram utilizados dois procedimentos de padronização. O primeiro consistia em manter a variável com 50 caracteres, mas eliminar todos os sinais de pontuações, cadeia de caracteres (de, dos, da etc.), espaços duplos e acentos. O segundo procedimento foi a subdivisão do nome que, além de realizar o mesmo procedimento anteriormente mencionado, criava automaticamente seis campos com nomes padrão:

Nome completo:	
FNOMEP	O primeiro nome
FNOMEU	O último nome
FNOMEI	As iniciais no meio do nome
FNOMEA	Os apêndices (Jr., Filho, Neto etc.).
PBLOCO	O primeiro nome formatado para a aplicação do código *Soundex* (modificações nas primeiras letras, para evitar problemas na utilização deste código).
UBLOCO	O último nome formatado para a aplicação do código *Soundex* (modificações nas primeiras letras, para evitar problemas na utilização deste código).

33 A padronização destas variáveis, denominadas como "comuns com a mesma estrutura" foi realizada utilizando as rotinas de relacionamento de "Padroniza" do programa RecLink II.

A seguir apresenta-se um exemplo desta padronização.

Nome completo: Waldemar Espinosa Melo Junior	
FNOMEP	WALDEMAR
FNOMEU	MELO
FNOMEI	E
FNOMEA	JUNIOR
PBLOCO	VALDEMAR
UBLOCO	MELO

Além disso, realizou-se uma padronização adicional do nome, na qual ao primeiro nome formato (PBLOCO) e ao último nome formato (UBLOCO), aplicou-se o código fonético de *Soundex* (Newcombe et al., 1988), útil para a blocagem[34].

3. Sexo: para esta variável utilizou como regra, utilizar o formato de um caractere com os seguintes códigos:

Variável	**Código**		***Label***	**Código padronizado**
	AIBF	**CadÚnico**		
Sexo da pessoa	1	F	Feminino	1
	2	M	Masculino	0

4. Data de nascimento: A variável foi convertida para 8 caracteres, eliminando pontuações e caracteres não alfanuméricos (/, -, .):

Variável	**Código**		**Código padronizado**
	AIBF	**CadÚnico**	
Data de nascimento	dd/mm/aaaa	C8	C8

Exemplo: 14/06/1980 passou para "14061980"

5. Código de município: Variável convertida para 7 caracteres:

Variável	**Código**		**Código padronizado**
	AIBF	**CadÚnico**	
Código IBGE	N7	C7	C7

[34] Como o *software* que foi utilizado permite realizar diretamente a padronização do código *Soundex*, esta opção da blocagem será detalhada na seção 4.6.2.

B) Variáveis comuns com diferentes estruturas em ambas as bases de dados:

1. Identificação do domicílio: Variável que é encontrada em ambas as bases de dados, mas com formatos e estrutura diferente. Desta forma, decidiu-se manter a variável com o formato original de sua respectiva base de dados, além disso, é uma variável que só será utilizada para uma revisão extra nos pares formados e definidos como indeterminados.

2. Idade da pessoa: Variável encontrada diretamente na base da pesquisa de campo AIBF, apresenta o formato numérico com 3 dígitos (N3). Para a base dos registros administrativos do CadÚnico, esta variável foi construída mantendo o formato de N3. Foi utilizada para uma revisão extra dos pares formados e definidos como indeterminados e sem data de nascimento na base AIBF e CadÚnico.

3. Endereço do domicílio: Variável que é encontrada em ambas as bases de dados, mas com estruturas diferentes. Manteve-se a variável com o formato original de sua respectiva base, porque esta variável será utilizada apenas para uma revisão extra nos pares formados considerados como indeterminado.

4.6 Relacionamento determinístico ou exato.

Nesta seção, o objetivo é descrever o método de relacionamento determinístico utilizado com a base da pesquisa de campo do AIBF e o registro administrativo do CadÚnico. Este método é escolhido quando existe um identificador único e a qualidade deste identificador é adequada. Além disso, o método confia na comparação da variável identificadora em ambas as bases de dados utilizadas.

4.6.1 Variável identificadora

O principal requerimento neste tipo de relacionamento é a disponibilidade de um identificador único, universal, fixo, de fácil registro e ao mesmo tempo rapidamente acessível e verificável. Neste trabalho, as bases de dados utilizadas apresentam uma variável que se aproximam às características deste tipo de identificador e que se denomina "Número de Identificador Social (NIS)". O NIS é um número que comprova a inscrição nos programas sociais do Governo Federal (tais como Bolsa Escola, Bolsa Alimentação,

Auxílio Gás ou Bolsa Família), designado à pessoa que realizou a inscrição para receber o beneficio. Embora seja um identificador único e intransferível, este apresentou problemas na declaração por parte das famílias entrevistadas (NIS com menos de 11 dígitos e inexistentes) e nos registros coletados a partir dos registros administrativos do CadÚnico (NIS com valor zero e duplicados).

4.6.2 Taxas de concordância encontradas para outras variáveis.

Depois de realizar o relacionamento exato, e conferir os resultados, foram realizadas comparações das outras variáveis que não foram utilizadas para o relacionamento exato, com o objetivo de reavaliar os pares verdadeiros formados. Além disso, estas comparações foram de importante utilidade para definir os parâmetros iniciais de concordância no relacionamento probabilístico tal como será visto na seção 4.7.5. Os resultados destas comparações são apresentados na TAB. 4.11.

TABELA 4.11 – Concordância das variáveis comuns entre os pares formados segundo o relacionamento determinístico. Brasil. 2006.

Região	Taxas de concordancia exata			
	Nome	Data Nascimento[1]	sexo	Codigo Municipo
Norte	58,8	86,74	88,15	96,4
Nordeste	59,1	89,81	88,75	95,1
Centro este	53,9	86,29	88,87	97,2
Sudeste	52,2	87,12	88,11	98,2
Sul	59,3	89,17	91,23	98,2

[1] Para a data de nascimento considero-se que o dia poderia ter até dois dias de difrença

Fonte: Dados encontrados com base ao relacionamento de base de dados da pesquisa de campos AIBF e CadÚnico.

4.6.3 Resultados de comparação determinística.

Como não existe incerteza no relacionamento exato, isto é, qualquer par de registros concorda ou não concorda, deve-se ter muito cuidado em conferir a qualidade da variável identificadora. Neste sentido, os registros escolhidos para este relacionamento foram aqueles que apresentaram o identificador NIS adequadamente preenchido ou que ao menos

garantem uma qualidade aceitável (o critério foi escolher somente aqueles que apresentaram 11 dígitos).

Os resultados do relacionamento determinístico depois de realizar uma revisão automática dos registros pareados são apresentados na TAB 5.4, e observa-se que 73,8% das pessoas ou registros que entraram para este relacionamento foram encontradas, os quais pertenciam a 35,24% do total das famílias entrevistadas na pesquisa de campo AIBF. Além disso, estes resultados indicam que, apesar de realizar uma avaliação da qualidade da informação do NIS, nem todas as pessoas ou registros foram considerados como pares verdadeiros.

TABELA 4.12 – Número de registros iniciais para o relacionamento determinístico* e resultados encontrados dos pares formados. Brasil. 2006.

Região	Pesquisa AIBF		CadÚnico	Pares verdadeiros encontrados	Famílias encontradas (***)	% em relação aos registros com NIS	% relação ao Total de famílias
	Pessoas	Famílias (**)					
Norte	1.440	1.236	5.581.690	1.063	930	75,26	38,07
Nordeste	3.308	2.758	28.075.258	2.355	2.056	74,54	40,26
Centro-Oeste	927	847	2.875.596	674	592	69,92	29,76
Sudeste	2.331	2.057	16.418.989	1.750	1.537	74,74	31,29
Sul	567	469	7.172.828	366	321	68,49	32,98
Total	8.573	7.367	60.124.361	6.208	**5.437**	73,80	35,24

* Neste caso, consideraram-se as pessoas que declararam ter número de NIS e que apresenta 11 dígitos.
** Considerou-se à família, quando ao menos um integrante da família declarou ter número de NIS.
*** Considerou-se família encontrada, quando menos um integrante da família foi encontrado.
Fonte: Dados elaborados a partir das bases da pesquisa de campo AIBF e registros administrativos do Cadúnico

A porcentagem de famílias encontradas com este método de relacionamento sugere a aplicação de outras metodologias de relacionamento, com o objetivo de incrementar o porcentual de famílias e realizar uma adequada nova alocação destas famílias nos grupos de comparação úteis na avaliação de impacto.

4.7 Relacionamento probabilístico.

4.7.1 Organização e tratamento das bases de dados para o relacionamento.

Antes de começar a delinear os passos utilizados para o processo do relacionamento probabilístico, ilustra-se o tratamento das bases de dados realizado. Como o objetivo do relacionamento de bases de dados neste trabalho é procurar o maior número de famílias entrevistadas na base de dados dos registros administrativos do CadÚnico, é necessário aplicar critérios factíveis para aproveitar a maior eficiência do relacionamento probabilístico. Ponderando este objetivo, o grande volume de informação dos registros administrativos do CadÚnico é um assunto que deve ser também considerado no tratamento dos dados e resultados.

Como conseqüência destes dois pontos, primeiro decidiu-se dividir a base de dados da pesquisa de campo AIBF em dois grupos: os beneficiários do PBF e os não beneficiários do PBF. Além disso, realizou-se o relacionamento por cada uma das cinco regiões do Brasil: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Em segundo lugar, como ainda se observou um grande volume de informação, o qual poderia complicar o processo de relacionamento (tanto no tempo e custo, como a capacidade das equipes de informática), o trabalho foi realizado em duas etapas. Na primeira etapa, considerou-se trabalhar apenas com os municípios que foram escolhidos na amostra da pesquisa de campo da Avaliação de Impacto do Programa Bolsa Família (269 municípios), mas, como ainda assim não se conseguiu atingir um número significativo de registros de pares verdadeiros, decidiu-se trabalhar em uma etapa seguinte com todos os municípios que foram encontrados no registro administrativo CadÚnico.

No quadro seguinte, resume-se a organização e tratamento realizado para as bases de dados do relacionamento probabilístico:

QUADRO 4.3 – Etapas utilizadas no relacionamento de base, segundo os grupos de população classificadas na base de dados da pesquisa AIBF e os registros Administrativos.

Considerando os municípios pesquisados na AIBF		
Etapa 1.	Beneficiários PBF	Norte Nordeste Centro – Oeste Sudeste Sul
Etapa 2	Não Beneficiários PBF	Norte Nordeste Centro – Oeste Sudeste Sul
Considerando todos os municípios encontrados no CadÚnico.		
Etapa 3	Beneficiários PBF	Norte Nordeste Centro – Oeste Sudeste Sul
Etapa 4	Não Beneficiários PBF	Norte Nordeste Centro – Oeste Sudeste Sul

4.7.2 Variáveis de blocagem

Na decisão das variáveis que serão utilizadas para a blocagem, dois critérios devem ser considerados: a confiabilidade e a discriminação. A confiabilidade objetiva diminuir os possíveis pares de registros perdidos, enquanto o critério de discriminação refere-se à procura por diminuição de custo e tempo de processamento (Gill, 2001). Desta forma, a escolha da melhor variável (ou variáveis) de blocagem implica a harmonia entre a confiabilidade e discriminação.

Data de eventos, data de nascimento, separado em meses, dias e anos; nome de batismo e sobrenome (ou seus correspondentes códigos fonéticos) são freqüentemente as melhores variáveis de blocagem.

Considerando que este processo pode apresentar também problemas de classificação, diferentes estudos, tais como Camargo e Coeli, 2002b; Jaro, 1989; Dean, 1996 e Machado, 2002 recomendam utilizar estratégias de blocagem em múltiplos passos. Esta estratégia

considera que os registros não pareados na primeira etapa do relacionamento serão novamente classificados no segundo passo ou etapa da estratégia de blocagem, a qual será realizada com uma nova chave.

Para nosso estudo, utilizou-se uma estratégia de blocagem em duas etapas, a primeira estratégia foi feita uma blocagem pela combinação dos códigos *Soundex* do último e do primeiro nome, sexo da pessoa e código de município; na estratégia seguinte foi realizada a blocagem pelo *Soundex* do primeiro e último nome, mais o sexo da pessoa (ver quadro 4.4). O objetivo era manter um tamanho menor de comparações, para evitar pares verdadeiros perdidos, a qual se pode conseguir com a estratégia e variáveis da blocagem utilizada:

QUADRO 4.4 – Estratégias de Blocagem utilizada para o relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico1.

Etapas	Chaves de Blocagem
B1[2]	*Soundex* (PBLOCO) + *Soundex* (UBLOCO) + MUNICIPIO DE RESIDENCIA + SEXO.
B2	*Soundex* (PBLOCO) + *Soundex* (UBLOCO) + SEXO.

[1] Para a aplicação das duas estratégias de blocagem empregou-se o programa *RecLink*.
[2] Esta blocagem foi utilizada considerando os municípios que foram escolhidos para a pesquisa de campo AIBF

4.7.3 Variáveis de relacionamento.

Uma opção para a escolha das variáveis de relacionamento é considerar todas as variáveis comuns nas bases de dados a relacionar. Mas, considerando que podem existir variáveis altamente correlacionadas entre si e que podem apresentar informação redundante do indivíduo, aumentando o peso e tempo de processamento, é preferível trabalhar apenas com um subconjunto de variáveis, porque podem contribuir muito mais ao poder de discriminação (GU, 2003).

Segundo Whalen et al (2001), as melhores variáveis de relacionamento são aquelas que possuem forte poder de discriminação e de identificação única, tal como "o número de documento de identidade nacional", que é único e aplicado a todos. Mas estas nem sempre estão disponíveis. Ante esta situação, é adequado procurar outras variáveis que tenham um poder similar ao mencionado. Uma dessas variáveis que também é um forte identificador

discriminatório e único é o nome completo da pessoa (admitindo que existam casos no qual o nome deixa de ser um único identificador). Outro exemplo de identificador é o sexo da pessoa, mas este é um identificador frágil, porque não provê um forte poder discriminatório como um único identificador, mas pode ser visto como uma variável de ajuda para a identificação do indivíduo. De igual forma, a data de nascimento pode ser utilizada como variável de ajuda para a identificação do indivíduo.

Gill (2001), em relação as variáveis de relacionamento, sugere utilizar um ou a combinação de variáveis que se encontram agrupadas nos seguintes grupos:

Grupo 1: Nomes próprios, os quais raramente mudam através do tempo (exceto o sobrenome das mulheres casadas)

Grupo 2: Características pessoais, que raras vezes mudam, tais como a data de nascimento e o sexo.

Grupo 3: Variáveis sócio-demográficas que podem ter variações severas durante o tempo, mas quando se relacionam bases que pertencem ao mesmo período do tempo podem ser utilizadas (endereço, estado civil).

Grupo 4: Variáveis coletadas para registros especiais, tal como ocupação, data de doença, diagnóstico, data de hospitalização, entre outros. Além disso, neste mesmo grupo, podem-se registrar as variáveis utilizadas para relacionamentos com fins familiares, tal como outros sobrenomes adicionais ao do grupo 1, peso ao nascer, genes, entre outras.

Grupo 5: Número de arbitrária alocação que identifica o indivíduo ou ente a ser relacionado.

As variáveis do grupo 1 e 2 são as que se utiliza comumente na prática quando estão presentes, mas, para serem utilizadas, é necessário realizar uma adequada edição e padronização destas. As variáveis do grupo 3 são utilizadas quando se deseja confirmar o par considerado como verdadeiro. Como conseqüência do descrito, pode-se dizer que a utilização nome, sexo e data de nascimento serão os identificadores que permitirão discriminar melhor os indivíduos. No trabalho, decidiu-se utilizar como variáveis de relacionamento: o nome completo e a data de nascimento; a variável sexo não foi utilizada, por estar incluída como variável de blocagem. Além disso, idade, endereço, código IBGE

de município (quando não estava incluído na blocagem) e a ordem das pessoas foram utilizados como variáveis para confirmar os pares verdadeiros.

4.7.4 Função de comparação para as variáveis de relacionamento.

Uma vez definidas as variáveis a serem utilizadas para o relacionamento, deve-se definir o peso de concordância e discordância de cada uma delas. O peso da variável será igual ao peso da concordância completa se a variável concorda completamente. Além disso, embora a variável concorde ou discorde, não necessariamente estas têm que ser exatas, desta forma, utilizando funções de comparação, a concordância completa, como também a concordância parcial é possível ser considerada. O *software* de relacionamento de dados utilizado neste trabalho, "*RecLink* II" apresenta as seguintes funções de comparação (CAMARGO e COELI, 2002b):

Aproximado: Realiza a comparação de seqüências de caracteres com base numa função determinada pela distância de Levenshtein[35]. Retorna valores entre 1 (correspondência total) e 0 (discordância total). É a função de comparação ideal para variáveis que guardam informações sobre nome.

Exato: Função que retorna 1 para pares exatos e 0 para pares discordantes (função que deve ser reservado para variáveis com apenas um caractere, nas quais a ocorrência de erros é pequena).

Caractere: Realiza comparações de seqüências de dígitos (ignorando separadores) compara pares de dígitos na mesma posição, retornando valores entre 1 para a correspondência total e 0 para a discordância total. Ë útil para variáveis que apresentam a data completa.

Diferença: Esta função calcula a diferença entre duas variáveis numéricas, considerando como par caso a diferença seja menor ou igual ao valor do parâmetro limiar aproximado. É utilizado para comparação de campos com informação ano, mês, dia.

---

[35] Chamada também de distância de edição, consiste no número mínimo de operações requeridas para transformar uma cadeia de caracteres em outra. Entende-se por "operação" a uma inserção, eliminação ou substituição de um caractere.

Para o presente trabalho, considerando as variáveis "nome completo" e "data de nascimento" como variáveis de comparação ou de relacionamento, devem-se utilizar as funções segundo foi indicado anteriormente. Assim, no Quadro 4.5 mostram-se as funções de comparação utilizadas.

QUADRO 4.5 – Função de comparação utilizada nas variáveis escolhidas para o relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

Variáveis Relacionamentos	Função de comparação	Concordância
Nome completo	Aproximado	Total ou parcial concordância
Data de nascimento	Caractere	Total ou parcial concordância

4.7.5 As probabilidades m_i e u_i.

A probabilidade "m" é a probabilidade que a variável concorde dado que o par de registros é um par verdadeiro. Isto pode ser interpretado como a confiabilidade de sua respectiva variável, dado que o calculo do "m" é igual a 1 menos a taxa de erro da variável. Como todas as variáveis não são igualmente confiáveis, espera-se que a probabilidade de "m" para diferentes variáveis pode variar.

Uma forma de encontrar a probabilidade "m" e a probabilidades "u" que é probabilidade da variável identificar um par de registros como verdadeiro, quando na realidade ele não é, é estimar estes valores através da teoria discutida por Fellegi e Sunter (1969) e Jaro (1989) tal como visto no capítulo 2. No entanto, existem formas práticas para encontrar os valores dos parâmetros, já que este procedimento formal é muito difícil e complicado de ser realizado.

Em qualquer situação, os parâmetros de relacionamento são usualmente estimados via um procedimento iterativo, o qual envolve uma revisão manual. Geralmente, a estimação dos parâmetros começa com um conjunto de parâmetros iniciais. Em seguida, a revisão de uma amostra de resultados de relacionamento e estimativas de parâmetros deve ser feita por meios de critérios *ad-hoc*. Finalmente, um processo de revisão e re-estimação deve ser repetido até que o relacionamento seja satisfeito de forma que os parâmetros e resultados não melhorem muito mais que o passo anterior. Um exemplo deste processo iterativo de

revisão e re-estimação encontra-se em Newcombe (1988), Estatísticas da Canadá (1983) e Jaro (1992) (WINKLER, 1993?).

De forma mais simples, também podem ser empregados valores previamente conhecidos pelo pesquisador ou de trabalhos realizados anteriormente. A seguir, apresentam-se algumas sugestões para os valores dos parâmetros de concordância e discordância.

QUADRO 4.6 – Probabilidade de concordância e discordância utilizadas ou sugeridas para algumas variáveis de relacionamento.

Autores	Valores para		Observações
	$m(\gamma)$	$u(\gamma)$	
Dean (1996)			
- Maioria das variáveis	0,90	10,0	Para a maioria dos campos, exceto para o caso do campo "sexo", no qual seria melhor empregar: $u(\gamma) = 0,5$.
Camargo e Coeli (2000)			
- Nome	0,92	1,0	Estes valores foram estimados pelos autores no relacionamento de arquivos de dados de Autorização de Internações Hospitalares e bases de mortalidade.
- Data de nascimento	0,90	5,0	
Data de integração das estatísticas de Nova Zelândia (2006)			
- Maioria das variáveis	0,90	-	As experiências nestes trabalhos mostraram que as variáveis padronizadas sexo, nome, sobrenome e data de nascimento têm bons valores de probabilidade "*m*". Para os valores de *u,* assume-se que os valores deste têm uma distribuição uniforme, e podem ser estimado por *1/n,* onde *"n"* é o número de valores da variável (ou categorias).
- Variável importante	0,99	-	
- variável moderadamente importante	0,95	-	
- Variável de pobre confiabilidade	<0,80	-	
Coeli CM et al. (2003)			
Primeiro nome	0,99	0,01	Avaliar os potenciais vantagens e limitações do uso das bases de dados dos formulários de Autorização de Internação Hospitalar e da metodologia do relacionamento probabilístico de registros, para a validação de relatos de utilização de serviços hospitalares durante inquéritos domiciliares.
Último nome	0,99	0,04	
Ano de nascimento	0,74	0,02	
Mês de nascimento	0,82	0,09	

Para o presente trabalho não se empregou um procedimento formal para a estimativa dos valores dos parâmetros. Estes foram escolhidos na combinação de: (i) valores iniciais tomados no relacionamento determinístico ou exato; (ii) testes com subconjuntos da base de dados; (iii) valores sugeridos pela bibliografia revisada (ver Quadro 4.6).

Como primeiro passo para definir os valores definitivos dos parâmetros que serão utilizados no relacionamento partiu-se de valores iniciais mostrados no Quadro 4.7.

QUADRO 4.7 – Parâmetros iniciais para o procedimento de definição dos parâmetros finais de *m* e *u* utilizadas para o relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

Variáveis	Probabilidades condicionais.			
	m	***u***	***1-m***	***1-u***
Nome completo	[0,80; 0,94]	[0,03;0,10]	[0,20, 0,06]	[0,97; 0,90]
Data de Nascimento	[0,81; 0,90]	[0,08; 0,15]	[0,24; 0,14]	[0,92; 0,85]

Logo depois de alguns testes com subconjuntos da base de dados que se está utilizando, os valores finais dos parâmetros são mostrados no quadro seguinte:

QUADRO 4.8 – Parâmetros e Funções de comparação utilizados para o relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

Campo	Função de comparação	$m(\gamma)$	$u(\gamma)$
Nome	Aproximado	90%	5%
Data de nascimento	Caractere	86%	10%

4.7.6 Pesos (w_i) e valores limiares.

Uma vez definido os valores de "*m*" e "*u*" o seguinte passo é calcular os pesos de cada variável que são construídos a partir de dois fatores de ponderação: posição de concordância e posição de discordância. O fator de concordância é calculado como,

$$w_c = \log_2\left(m/u\right) \qquad (4.1)$$

e o fator de discordância como,

$$w_d = \log_2\left(1-m/1-u\right). \qquad (4.2)$$

Em relação aos valores limiares, Fellegi e Sunter (1969) propuseram a definição do conceito destes com o objetivo de classificar os pares em três categorias: pares verdadeiros,

não pares e pares incertos. Isto é, os pares que apresentarem o escore acima de valor predeterminado (limiar superior) serão classificados como pares verdadeiros, enquanto aqueles que exibiram escore abaixo de um segundo valor também predeterminado (limiar inferior) serão considerados como não pares. Os registros pareados que apresentem valores de escore intermediários entre o limiar inferior e superior são registros pareados incertos e precisariam passar por um processo de revisão manual (CAMARGO e COELI, 2002). Os pesos calculados são apresentados no seguinte quadro:

Quadro 4.9 – Pesos e limiares para o relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

Campo	Probabilidades condicionais		Pesos de concordância	Pesos de discordância	Limiares
	$m(\gamma)$	$u(\gamma)$	(w_c)	(w_d)	
Nome	90%	6%	3,9069	-3,2327	85%
Data de nascimento	86%	10%	3,1043	-2,6845	84%

O escore total de um determinado registro pareado dentro de cada bloco é obtido a partir da soma dos fatores de ponderação atribuídos após a comparação de cada campo avaliado.

TABELA 4.13 – Poder de discriminação e pesos extremos encontrados no relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

Variável	Poder de discriminação da variável	Limiares extremos na escala do Escore (pesos extremos)	
		Inferior	Superior
Nome	7,1396	-5,9172	7,0112
Data de nascimento	5,7888		

Fonte: Dados encontrados com base ao relacionamento probabilístico da base de dados da pesquisa de campos AIBF e os registros administrativos do CadÚnico.

Como ilustração, na TAB. 4.13, mostra-se o poder discriminatório que apresenta cada variável de relacionamento utilizado e os escores extremos. Observa-se que o poder de discriminação é maior na variável "nome" como conseqüência de assinar a esta uma maior probabilidade de concordância (m) e menor probabilidade de discordância (u).

Os limiares extremos são também denominados pesos ou escore extremos, porque através destes pode-se aceitar como pares os registros pareados com valores de escores maior ou igual ao valor extremo superior, rejeitar os com valor menor ou igual ao extremo inferior e encaminhar para a revisão manual os registros pareados com valores intermediários de escore. O exemplo do escore extremo inferior da tabela anterior apresenta o caso na qual não houve concordância nos registros pareados nem na variável nome nem em data de nascimento; por outro lado, o escore extremo superior mostra os registros pareados em que houve concordância total em ambas as variáveis. No entanto, estes casos não são os únicos a ser encontrados no processo de comparação, dado que existem casos na qual a concordância não e total, mas sim parcial. Neste sentido, o *software Reclink* possui a capacidade de aplicar algoritmos mais complexos que permitem atribuir frações de pesos de concordâncias para variáveis que não necessariamente sejam iguais, mas similares (CAMARGo e COELI, 2002).

Considerando as concordâncias parciais, existe um trabalho de revisão manual dos escores associados aos registros relacionados, com o objetivo de explorar estes escores e definir os verdadeiros valores limiares, processo que será discutido na seção seguinte.

4.7.7 Revisão manual

O primeiro passo nesta parte do relacionamento é realizar uma revisão da distribuição de freqüências dos escores associados a cada par de registros relacionados. O objetivo neste primeiro passo é determinar os valores limiares que permitam reduzir a inspeção manual dos pares considerados incertos, economizando tempo na análise manual, porque os pares que não concordam em nenhuma das variáveis poderão ser sempre eliminados. Assim, será aceita como par verdadeiro os registros pareados com valores de escore maior ou igual a um valor do limiar superior, e rejeitar aqueles com valor menor ou igual ao limiar inferior e encaminhar para a revisão manual os registros pareados com valores intermediários de escore. Em nossa análise, utilizou-se uma decisão combinada entre os histogramas da distribuição dos pesos dos registros pareados no testes com subconjuntos da base de dados, e a inspeção dos registros pareados.

No segundo passo, uma vez definidas os limiares, foi realizada uma análise manual dos pares obtidos e considerados incertos, com o objetivo de determinar se a informação

refere-se à mesma pessoa. Para isto, em um primeiro momento, aplicou-se um procedimento automático para classificar os registros pareados segundo a data de nascimento nas seguintes categorias: acordo parcial e completo (exemplo: diferença de um ano e/ou um mês) e discordância. Quando os registros pareados ainda não poderiam ser classificados como par verdadeiro ou não, se utilizam outras variáveis auxiliares, como ordem da pessoa dentro da família, endereço do domicílio e código de município (quando este não foi utilizado na blocagem).

O processo de revisão manual neste trabalho foi rigoroso, porque não se desejava classificar como pares verdadeiros aqueles que não se referiam à mesma pessoa, por exemplo, nomes iguais, mas com algumas variáveis auxiliares diferentes eram descartados, a menos que os nomes completos fossem pouco comuns e a data de nascimento fosse próxima.

4.7.8 Concordância e discordância.

Quando os registros pareados de ambas as bases de dados são pares verdadeiros porque são identicamente iguais, é simples e fácil considerar que os registros “concordam totalmente” (no caso de serem completamente diferentes, “discordam totalmente”). No entanto, a concordância ou discordância total nem sempre podem ser vistas, e o problema da decisão da concordância ou discordância entre dois registros como par verdadeiro torna-se complicado. Isto porque existem registros pareados que podem ser considerados parcialmente concordantes ou discordantes, porque existem pequenas diferenças no nome da pessoa ou na data de nascimento. Segundo Jaro (1989), uma solução para os casos com discordância pequena é a atribuição de um fator de ponderação de concordância que contribui positivamente para o escore final de forma parcial, mas considerando que esta alocação deve ser menor do que aquela que seria utilizada no caso de concordância total. Além disso, definir a discordância parcial “aceitável” e que fator de ponderação de concordância deve ser utilizado é complicado, e como conseqüência a decisão se o registro pareado é um par verdadeiro nesses casos.

Exemplos da concordância total são apresentados na TAB 4.14, no qual se observam três casos em que não há problema no momento da decisão do par ser verdadeiro.

TABELA 4.14 – Casos práticos de concordância total encontrados no relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

	Bases	Casos	Nome	Data nascimento	sexo	Município (IBGE)
1	AIBF	1020	VALDEMAR SILVA OLIVEIRA	19051961	M	4319158
	CadÚnico	339087	VALDEMAR SILVA OLIVEIRA	19051961	M	4319158
2	AIBF	1022	ELIMAR MACIEL OLIVEIRA	2011997	F	4319158
	CadÚnico	339089	ELIMAR MACIEL OLIVEIRA	2011997	F	4319158
3	AIBF	1026	MERIANE BRAGA SOUZA	18111996	F	4319158
	CadÚnico	158806	MEIRIANE BRAGA SOUZA	18111996	F	4319158

Fonte: Dados encontrados no processo do relacionamento probabilístico da base de dados da pesquisa de campos AIBF e os registros administrativos do CadÚnico.

Na TAB. 4.15, mostra-se um caso prático encontrado no processo de relacionamento no qual um registro da base de dados da pesquisa de campo AIBF foi pareado com quatro registros da base de dados do registro administrativo do CadÚnico.

TABELA 4.15 – Caso prático de concordância parcial encontrados no relacionamento da base da pesquisa de Campo AIBF e Registros Administrativos CadÚnico.

	Bases	Casos	Nome	Data de nascimento	Sexo	Municipio (IBGE)
1	AIBF	24	ERICK VIDAL PINTO	17012003	M	4106902
	CadÚnico	303.637	ERIQUE ALISSON PINTO	05011997	M	4106902
2	**AIBF**	**24**	**ERICK VIDAL PINTO**	**17012003**	**M**	**4106902**
	CadÚnico	**348.780**	**ERICK VIDAL PINTO**	**18012003**	**M**	**4106902**
3	AIBF	24	ERICK VIDAL PINTO	17012003	M	4106902
	CadÚnico	355.632	ERIC LUIZ PINTO	4121984	M	4106902
4	AIBF	24	ERICK VIDAL PINTO	17012003	M	4106902
	CadÚnico	539.642	HERRIQUE EDUARDO PINTO	25111992	M	4106902

Fonte: Dados encontrados no processo do relacionamento probabilístico da base de dados da pesquisa de campos AIBF e os registros administrativos do CadÚnico.

Neste caso, observa-se que o registro da pesquisa AIBF não apresenta concordância total nas variáveis de relacionamento com os registros do CadÚnico, considerando este caso como de concordância parcial, a qual tem que ser definida mediante uma revisão manual. Neste sentido, analisando os pares formados na tabela, pode-se afirmar que o registro do CadÚnico que mais semelhança apresenta com o do AIBF, corresponde ao caso 2, isto é,

registro designado pelo número 24 no AIBF com o registro 348.780 do CadÚnico, portanto este registro pareado formará parte do grupo de pares verdadeiros, definidos através de uma concordância parcial.

4.7.9 Resumindo os passos de blocagem e variáveis de relacionamento utilizadas.

Depois de realizar todo o processo e etapas mencionadas, os pares formados pelo relacionamento de registros passaram a compor novos arquivos de dados para serem analisados segundo a proposta do estudo. A seguir, apresenta-se um quadro com a blocagem e as variáveis utilizadas no relacionamento, assim como as comparações utilizadas para se conseguir os pares verdadeiros.

QUADRO 4.10 – Variáveis utilizadas em cada passo do processo de relacionamento probabilístico e revisão manual.

Blocagem		
Etapa	**Famílias AIBF**	**Variáveis ordenadas**
1	Beneficiarias PBF	Soundex do primeiro nome + Soundex do primeiro nome + município de residência + sexo
2	Beneficiarias PBF	Soundex do primeiro nome + Soundex do primeiro nome + sexo
3	Não Beneficiarias PBF	Soundex do primeiro nome + Soundex do primeiro nome + município de residência + sexo
4	Não Beneficiarias PBF	Soundex do primeiro nome + Soundex do primeiro nome + sexo
Relacionamento		
Nome completo		Total ou parcial concordância
Data de nascimento		Total concordância
Revisão Manual		
Nome completo		Total ou parcial concordância
Data de nascimento ou idade		Total ou parcial concordância
Endereço ou (código de município IBGE)		Total ou parcial concordância
Ordem da pessoa		Total ou parcial concordância

4.7.10 Resultados do relacionamento probabilístico.

Antes de mostrar os resultados finais obtidos no relacionamento probabilístico, apresentam-se alguns resultados parciais que ilustram a aplicação deste relacionamento. Um primeiro resultado é mostrado na TAB 4.16, na qual se observa os pares que se

formariam com a não aplicação da Blocagem e os pares formados quando se aplicou a blocagem na Etapa 1. Comparando os pares que deveriam formar-se sem blocagem[36] e os formados com esta, ressalta-se a excelente redução dos pares formados conseguidos, redução que esteve em torno de 99%, para um grupo da amostra da pesquisa AIBF (beneficiários do PBF) e para todas as regiões do Brasil. Estes resultados foram conseguidos também nas outras etapas, regiões e subgrupos, resultados que são mostrados no APÊNDICE I.

TABELA. 4.16 – Número de registros iniciais para o relacionamento probabilístico e os pares formados. Brasil. 2006. Etapa 1.

Região	Pesquisa AIBF	CadÚnico	AxB	Pares formados segundo a Blocagem
	(A)	(B)		
Norte	2.108	1.378.954	2.906.835.032	682.417
Nordeste	4.445	4.216.672	18.743.107.040	390.882
Centro este	2.399	1.002.202	2.404.282.598	99.782
Sudeste	4.070	3.131.376	12.744.700.320	81.306
Sul	1.036	597.074	618.568.664	12.415
Total	14.058	10.326.278	37.417.493.654	1.266.802

Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. MDS. 2006

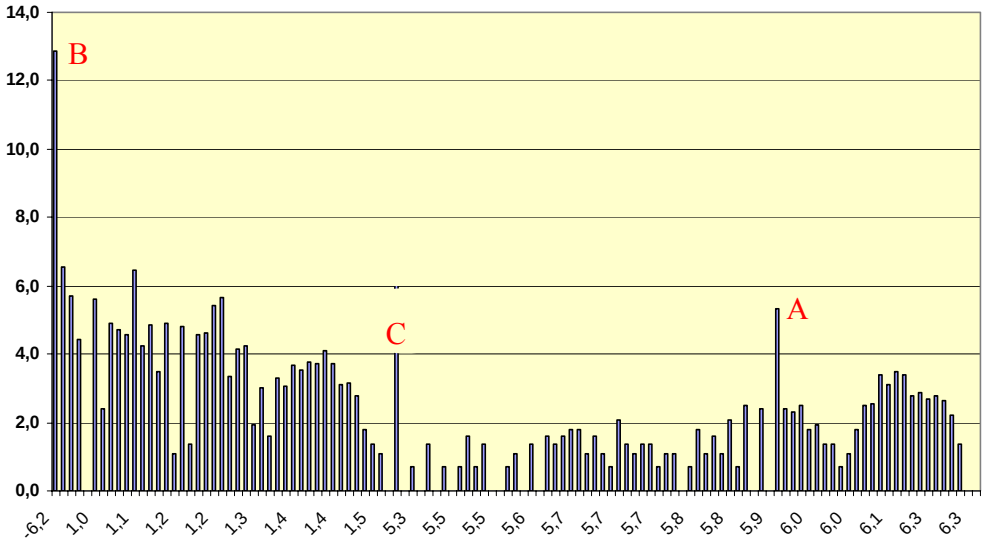
A importância de mostrar este exemplo é para avaliar a blocagem utilizada no trabalho, já que o êxito desta depende, em parte, da formação do menor número de pares de registros possíveis, para tornar o relacionamento mais eficaz e competente.

Um segundo resultado a ser mostrado encontra-se no GRAF 4.1, o qual representa a distribuição dos pesos (ou escores) para os registros considerados como pares e não pares para os beneficiários do PBF da região Sul e da ETAPA 1. Os escores à direita do gráfico representam os pares considerados como pares verdadeiros e tem sua maior freqüência no valor assinalado como "A". Em relação aos escores que estão na parte esquerda da

[36] Lembrar que para esta etapa a Blocagem foi "Soundex (PBLOCO) + Soundex (UBLOCO) + MUNICIPIO DE RESIDENCIA + SEXO".

distribuição, estes representam os registros pareados considerados como não-pares e tem sua maior freqüência no ponto assinalado com “B”. A maior freqüência do gráfico encontra-se à esquerda e é efetivamente o ponto assinalado com “B”, que é maior que a freqüência da direita identificada com “A”. Resultado importante a destacar, porque se confirma a hipóteses da configuração do histograma dos pesos ou escores vista na seção 2.3.2, na qual se afirma que existem mais registros pareados que são considerados como não pares. Além disso, pode-se observar outra freqüência relativamente maior no ponto “C”, que se encontra próximo ao centro do gráfico da distribuição dos pesos, freqüência que indica a área dos pesos dos registros pareados considerados como pares potenciais e que precisariam realizar uma revisão manual. A distribuição de todas as regiões e etapas realizadas pode ser encontrada no APÊNDICE II.

GRÁFICO 4.1 – Distribuição de freqüência dos pesos totais do relacionamento. Probabilístico. Região Sul. Brasil 2006. Beneficiários da Etapa 1.



Fonte: Dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico - MDS. 2006

Finalmente, depois de ilustrar alguns resultados parciais que foram encontrados durante o processo de relacionamento probabilístico, queda por apresentar os resultados finais da porcentagem de registros pareados que foram considerados como pares verdadeiros. Na TAB 4.17, apresentam-se os resultados finais obtidos ao realizar o relacionamento das bases de dado da pesquisa de campo AIBF com os registros administrativos do CadÚnico e utilizando o programa de *Reclink* II:

TABELA. 4.17 – Registros encontrados no método de relacionamento probabilístico nas regiões e etapas utilizadas. Brasil. 2006.

Região	Registros da pesquisa de campo AIBF*	Registros encontrados	% de registros encontrados
Norte	6.202	5.568	89,78
Nordeste	15.948	10.639	66,71
Centro-Oeste	5.347	3.987	74,57
Sudeste	13.371	9.575	71,61
Sul	2.445	2.059	84,21
Total	43.313	31.828	73,48

***** Neste caso foram considerados para o relacionamento probabilístico aquelas pessoas que não foram encontrados com o relacionamento determinístico ou exato.
Fonte: Dados elaborados a partir do relacionamento probabilístico da base da pesquisa de campos AIBF e os registros administrativos do CadÚnico.

Os resultados mostrados na TAB 4.17 nos indicam que a porcentagem de registros encontrados na aplicação deste relacionamento está em torno dos 73% para todo Brasil, porcentual que pode ser considerado significativo, já que apenas 27% não se consegue recuperar dos registros administrativos. Quando se revisou a porcentagem por região, observa-se que apenas o Nordeste apresentou um porcentual menor que 70%, enquanto a região que conseguiu maior registros pareados é a região Norte, atingindo um porcentual próximo a 90%. Encontrar uma a explicação técnica para estas diferenças pode ser trabalhoso, porque, quando se planejou o relacionamento probabilístico, considerou-se que todas as regiões partem com os mesmo supostos e parâmetros para o processo de relacionamento, portanto o tratamento durante o processo de relacionamento foi padrão para todas as regiões. No entanto, o estudo não escapa de alguns erros na medida em que as variáveis utilizadas para blocagem e relacionamento fossem mal preenchidas, ou mesmo não preenchidas, resultando impossível identificar um par verdadeiro.

4.8 Nova alocação das famílias nos grupos de comparação.

Uma das tarefas mais importante a realizar na tese é utilizar os registros administrativos do CadÚnico para alocar as famílias ao grupo de tratamento e comparação (grupos comparações), mas para isto, emprega-se o relacionamento de base entre a base da

pesquisa AIBF e do CadÚnico, para atingir a esse objetivo. A seguir, são mostrados os resultados das famílias que foram encontradas com ambos os métodos utilizados no relacionamento de base de dados, o processo e informação utilizados para a alocação destas famílias nos grupos comparações segundo CadÚnico e, por fim, os resultados da alocação das famílias nos grupos comparações utilizando ambas as fontes de dados.

4.8.1 Famílias encontradas depois do relacionamento determinístico e probabilístico.

Para um melhor entendimento e seguindo o tratamento dos dados realizados no processo de relacionamento, os resultados a seguir mostram-se por Regiões e por método de relacionamento utilizado. Assim, para Brasil todo, observa-se que do total de 15.426 famílias entrevistadas na pesquisa de campo AIBF, 5.437 famílias foram encontradas[37] com o relacionamento determinístico, isto é, 35% do total, enquanto que com o relacionamento probabilístico, foram encontradas 4.550 famílias que representam 30% do total da AIBF. Em conjunto a porcentagem de famílias encontradas para Brasil foi de 65% do total de famílias da pesquisa de campo AIBF. Quando se analisa o resultado por regiões, observa-se que a porcentagem de famílias encontradas é similar ao encontrado para todo Brasil, exceto para Região Nordeste na qual, em conjunto com ambos os métodos, aproximadamente 70% das famílias foram encontrados em relação às famílias entrevistadas na pesquisa de campo AIBF (ver TAB 4.18).

[37] Considerou-se "unida domiciliar da família encontrada", quando ao menos uma pessoa deste foi encontrada.

TABELA 4.18 – Famílias* encontradas nos dois métodos de relacionamento aplicados e nas etapas utilizadas. Brasil. 2006.

Região	Famílias a serem encontradas segundo AIBF	Famílias encontradas no processo de relacionamento com o CadÚnico			% de famílias encontradas
		Determinístico	Probabilístico	Total	
Norte	2.443	930	713	1.643	67,25
Nordeste	5.106	2.056	1.483	3.539	69,31
Centro-Oeste	1.990	592	567	1.159	58,24
Sudeste	4.913	1.537	1.472	3.009	61,25
Sul	974	321	316	637	65,40
Total	15.426	5.437	4.550	9.987	64,74

***** Considerou-se "família encontrada", quando ao menos uma pessoa deste foi encontrada.
Fonte: Dados elaborados a partir da base da pesquisa de campo AIBF e registros administrativos do Cadúnico

Analisando os resultados encontrados na TAB 4.18, pode-se dizer que, apesar da exclusão de registros com erros de preenchimento ou duplicidade da informação, o número de famílias que foi possível encontrar pode ser considerado ótimo, isto devido que, nem todas as famílias entrevistadas na pesquisa de campo AIBF estão no registro administrativo CadÚnico, porque segundo o plano amostral, na amostra AIBF existem famílias cadastradas no Cadastro Único, mas ainda não beneficiárias do Programa Bolsa Família (domicílios podem ser beneficiários de outros programas de transferência de renda, mas não do programa Bolsa Família); e famílias não cadastradas ou beneficiárias (Oliveira et *al*, 2007). Estas famílias que pertencem a estes últimos grupos da amostra mencionada talvez possam ser parte dos 35% das famílias não encontradas com a utilização dos métodos de relacionamento. Para poder ter mais argumentos sobre a qualidade dos resultados encontrados no relacionamento, será necessário cruzar algumas variáveis, como o benefício recebido por estas famílias entrevistadas segundo a pesquisa de campo AIBF com a variável do benefício recebido segundo os registros administrativos.

No trabalho de relacionamento entre a base da pesquisa AIBF e do CadÚnico, uma vez, que o par de registro relacionado era considerado como um par verdadeiro, o passo

seguinte foi recuperar a informação do "Número de Identificação Social - NIS" para as pessoas que não continham essa informação na base da pesquisa de campo do AIBF, informação útil para alocação das famílias entrevistadas nos grupos de comparação segundo o registro administrativo CadÚnico.

4.8.2 Procurando os grupos de comparação nos registros administrativos.

Uma vez que as pessoas encontradas em ambas das bases de dados contam com o NIS corretamente identificado, a seguir, realiza-se a alocação das famílias nos grupos de comparações segundo os registros administrativos do CadÚnico. No entanto, para levar a cabo esta alocação, precisa-se utilizar a informação das folhas de pagamento dos benefícios sociais de um mês anterior à data da pesquisa, isto é, folha de pagamento do dos benefícios sociais de outubro de 2005.

A Folha de Pagamento dos benefícios sociais, fornecida pela Gerência de Filial de Serviços Sociais (GISES/CT) da Caixa Econômica Federal (Brasil, 200-?f), possui importantes informações que servem para o controle dos beneficiários e facilitam o trabalho das prefeituras. A folha de pagamento funciona utilizando informações como nome do município, mês de referência, agência e superintendência da Caixa que atendem à região do município, informações da Regional da SETP (Secretaria Estadual de Trabalho, Emprego e Promoção Social) que atende ao município, bem como o portal da Caixa no site da SETP.

A Folha também apresenta a lista de todos os beneficiários por município, em ordem alfabética, contendo Código Domiciliar, NIS, Nome do Responsável Legal e valores disponibilizados ao beneficiário, por produto e total. As quais são importantes, porque a Folha de Pagamento lista as famílias por tipo de benefício que recebem e vice-versa. Também possibilita ao Gestor Municipal identificar visualmente se há nomes de Responsável Legal com duplicidade de benefícios, para aquelas pessoas que tiveram a atribuição de dois NIS em função de erro nos seus dados cadastrais (Brasil, 200-?f).

Considerando estas últimas informações, especificamente, o NIS do responsável legal integrante da família e o tipo de benefício que recebem as famílias, é possível que as famílias entrevistadas na pesquisa de campo AIBF e encontradas também no registro administrativo do CadÚnico, possam ser alocada nos grupos de comparações segundo este

registros. Como conseqüência da recuperação destas informações, aumenta a possibilidade de dispor de uma grande base de dados com informação completa das ambas as fontes de informação, tornando-se isto um fato.

4.8.3 Alocação das famílias nos grupos de comparação.

Finalmente, o resultado do relacionamento da base da pesquisa de campo AIBF e do registro administrativo CadÚnico possibilita estudar e analisar as mudanças ocorridas nas famílias quando são alocadas nos grupos de comparação, segundo cada uma das fontes de dados utilizadas. Estas informações são importantes porque também nos permitem observar a qualidade das informações pareadas no relacionamento, em especial das famílias beneficiárias do Programa Bolsa Família, que são o objetivo de nosso estudo.

Na TAB 4.19, mostram-se os resultados das famílias da Pesquisa AIBF segundo inserção em Programas de Transferência de Renda e alocadas segundo os Registros Administrativos (Folhas de Pagamento e Cadastro Único). Observa-se que, apesar de ter encontrado 65% das famílias da Pesquisa AIBF no CadÚnico, o número de famílias do PBF no processo de relacionamento atingiu quase 94% dos dados da base original AIBF. Em relação às famílias de outros benefícios da pesquisa de campo AIBF, a porcentagem de famílias encontradas em ambas as base de dados foi de 83%, nas famílias cadastradas sem benefício da pesquisa de campo AIBF, esta porcentagem caiu até 44%, enquanto que não cadastradas e sem benefícios foi de 19%.

TABELA 4.19 – Família* da Pesquisa AIBF segundo inserção em Programas de Transferência de Renda e Situação nos Registros Administrativos (Folhas de Pagamento e Cadastro Único). Brasil. 2006.

Pesquisa de Campo AIBF	Registros Administrativos				Total
	Bolsa Família	Outros benefícios	Cadastrados, mas não apresentam nenhum benefício.	Não encontrado	
Bolsa Família	4120 (87,57)	108 (2,30)	152 (3,23)	325 (6,09)	4.705
Outros benefícios	1167 (29,28)	2005 (50,3)	145 (3,64)	669 (16,78)	3.986
Cadastrados sem benefício	470 (10,62)	146 (3,30)	1250 (28,25)	2559 (57,83)	4.425
Não cadastrados e sem Benefícios	79 (3,42)	46 (1,99)	299 (12,94)	1886 (81,65)	2.310
Total	5.836	2.305	1.846	5.439	15.426

***** Considerou-se família encontrado, quando ao menos uma pessoa deste foi encontrada.
Fonte: Dados elaborados a partir da base da pesquisa de campo AIBF e registros administrativos do Cadúnico.

Analisando as redistribuições acontecidas como causa da alocação das famílias nos grupos segundo o CadÚnico, pode-se observar que 88% das família beneficiárias do PBF da pesquisa AIBF estão alocadas também no mesmo grupos segundo o CadÚnico, enquanto 2% recebem outros benefícios e 3% não recebem benefícios. Em relação às famílias de outros benefícios da pesquisa de campo AIBF, observa-se que apenas 50% destas estão alocadas também no mesmo programa segundo o CadÚnico e uma porcentagem significativa de 30% foi re-classificada no PBF, enquanto que só 4% não recebem benefícios. Observa-se que as famílias do grupo de cadastrados sem benefício e não cadastrados e sem benefícios segundo a pesquisa de campo AIBF, em conjunto, 14% alocam-se no PBF e 5% em outros programas, segundo os registros do CadÚnico. Contundo, os resultados da TAB 4.19, indicam que a alocação das famílias segundo o CadÚnico, as famílias beneficiárias do PBF incrementou em 24%, resultados que parecem ser alentadores, considerando que o propósito da tese é analisar as mudanças ocorridas nos resultados de impacto do PBF quando se utilizam registros administrativos ao alocar nos grupo de comparações, e obter maior número de beneficiários ou aumentar as amostra para a avaliação do impacto sempre gera maior robustez dos dados.

O número de famílias encontradas pode ser considerado bom, pois, deve-se ter em consideração que nem todas as famílias entrevistadas na pesquisa de campo AIBF devem estar no registro administrativo CadÚnico. Uma vez que, segundo a amostra AIBF, existem famílias cadastradas, mas ainda não beneficiárias do Programa Bolsa Família (domicílios beneficiários de outros programas de transferência de renda, mas não do PBF); e famílias não cadastradas ou beneficiárias (OLIVEIRA et *al*, 2007). As famílias que pertencem a estes últimos grupos da amostra mencionada talvez possam ser parte dos 35% das famílias não encontradas com a utilização de nenhum dos métodos de relacionamento.

Como não se conhece com exatidão em que medida o resultado obtido é influenciado pela qualidade das bases utilizadas (AIBF e CadÚnico) e a precisão do método de relacionamento probabilístico, é possível ter uma idéia da qualidade das informações pareadas no relacionamento e CadÚnico com a porcentagem das famílias beneficiárias do PBF da pesquisa de campo AIBF, que foram encontradas no CadÚnico. Isto informação é importante, porque as famílias do PBF têm maior probabilidade de estar registradas neste cadastro e, também porque nosso objetivo da tese tem como base estas famílias. Junto com este argumento, deve-se ter em conta que, na amostra AIBF existem famílias cadastradas no CadÚnico que ainda não são beneficiárias do PBF (domicílios podem ser beneficiários de outros programas de transferência de renda) e famílias não cadastradas ou beneficiárias, que realmente não podem ser encontradas no processo de relacionamento (Oliveira et al, 2007). Assim, o número de famílias que foi possível encontrar ou parear com o processo de relacionamento de bases de dados pode ser considerado suficiente para analisar as presumíveis variações ou sensibilidades dos resultados de impacto do PBF, quando se utilizam registros administrativos ao alocar às famílias nos grupo de comparações, mas sem desconsiderar os argumentos antes mencionados.

5 RESULTADOS DA AVALIAÇÃO DE IMPACTO DO PROGRAMA BOLSA FAMÍLIA NA EDUCAÇÃO

Neste capítulo, primeiro ilustram-se os dois tipos de alocação das famílias nos grupos de comparação utilizados para analisar os resultados de impacto nos indicadores da educação: alocação segundo o relacionamento de bases de dados com os registros administrativos do Cadastro Único (CadÚnico) e as diferenças que existem com a alocação dos grupos de comparação obtidos segundo a pesquisa de campo AIBF. A seguir, destaca-se a análise da sensibilidade dos resultados para avaliar as comparações dos indicadores de impacto na educação entre as duas alocações das famílias nos grupos de comparação utilizadas no presente trabalho. Na seção seguinte, descrevem-se os dados e as variáveis dependentes e independentes utilizados para a avaliação de impacto. As variáveis dependentes são os indicadores de impacto para avaliar os diferenciais do PBF na educação das crianças entre 7 e 14 anos de idade, enquanto, as variáveis independentes são aquelas características, utilizadas na especificação dos modelos equilibrados do escore de propensão e na regressão descontinua. Finalmente apresentam-se os resultados da aplicação do escore de propensão do método de *matching*, destacando a comparação dos resultados obtidos paras as alocações dos grupos de comparação utilizados. Além disso, como produto do pareamento das bases de dados da pesquisa de campo e dos registros administrativos, apresentam-se uns dos resultados obtidos com a aplicação do desenho S*harp* da regressão descontinua para os indicadores da educação.

5.1 Variável de identificação dos grupos recuperados para análise do impacto na educação e o termos relacionamento e pareamento (*matching*).

Com os dados do relacionamento das bases de dados obtidos no capítulo quatro, constituiu-se a nova alocação dos grupos de comparação a ser considerados para a análise de impacto na educação dos beneficiários do Programa Bolsa Família (PBF), considerando que existe já uma classificação dos grupos de comparação de acordo com as informações coletadas nos questionários, que foram ao campo, da pesquisa AIBF.

Os grupos de comparação definidos na pesquisa AIBF foram três. O primeiro deles chamado de "Tratamento" (T), é constituído pelos domicílios que declaram receber na data da pesquisa o benefício PBF. Os outros dois grupos, denominados grupos de comparação se subdividem em "Comparação 1" (C1) composto pelos domicílios que recebem na data da pesquisa outros benefícios; e o outro grupo de comparação, denominado "Comparação 2" (C2), que está composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público. Estes grupos como visto, dependem diretamente dos benefícios que a unidade domiciliar recebe ou declara receber por parte dos órgãos do Estado. Considerando a definição da distribuição dos grupos, a alocação alternativa proposta neste trabalho foi captada cruzando os dados encontrados no relacionamento com os registros administrativos do CadÚnico e uma base que continha os benefícios recebidos pela família no mês anterior à data da pesquisa (ver Capitulo 4). Além disso, para os objetivos deste trabalho, os grupos de comparação utilizados para analisar os resultados de impacto nos indicadores da educação são: "Tratamento" (T), e grupo denominado "Comparação 2" (C2).

A alocação que considera a classificação dos domicílios realizada com as informações coletadas nos questionários que foram a campo na pesquisa a AIBF em outubro de 2005 será denominada neste trabalho daqui por diante como, "Alocação segundo a pesquisa de campo AIBF", enquanto a alocação obtida como produto do relacionamento da base de dados AIBF com o registros administrativos do Cadastro Único será denominada "Alocação segundo os registros administrativos do CadÚnico" ou simplesmente "Alocação segundo o CadÚnico".

Além dos dois procedimentos de alocação dos grupos de comparação, ressalta-se que os domicílios serão classificados segundo critérios de elegibilidade de renda, considerando três cortes de renda domiciliar per capita: até R$200,00, até R$100,00 e até R$50,00. O restante da amostra, que é constituída por domicílios que já receberam algum tipo de benefício, mas que não o recebem mais, e de domicílios cuja renda domiciliar per capita é maior que R$200,00, será excluída da análise de avaliação de impacto, tal como indica Oliveira et al (2007).

Com a aplicação nas seguintes seções da técnica de *Matching* de Escore de Propensão (PSM), que compara resultados de famílias similares do grupo de Beneficiários do PBF com as famílias do grupo de comparação 2, deve-se trazer em consideração uma

observação feita no Capitulo 3, referente aos termos “Relacionamento” e “Pareamento” Assim, será apropriado avisar que, o termo “Relacionamento” será utilizado quando nós referimos à relacionamento das bases de dados realizados entre a base da pesquisa AIBF com os registros Administrativos do CadÚnico, enquanto que o termo de “Pareamento” será referido para a técnica utilizada na avaliação de impacto dos programas sociais com o escore de propensão, cujo objetivo é construir pares sobre as observações de controle e o tratamento que são similares em termos das características observáveis.

5.2 Utilizando a sensibilidade dos resultados para analisar a comparação dos dois tipos de alocações das famílias nos grupos de comparação.

Como mencionado na seção 2.5, o relacionamento de base de dados da pesquisa de campo AIBF e dos registros administrativos do Cadúnico possibilita aumentar a qualidade e quantidade de informação estatística sobre os dados trabalhados, porque embora a informação coletada na pesquisa de campo sobre o recebimento do benefício que as famílias entrevistadas, sejam consideradas adequadas para análise na AIBF; nas pesquisas de campo, é possível que as respostas estejam influenciadas por aspetos subjetivos, como opiniões ou atitudes das pessoas. Além disso, este relacionamento de dados permitirá avaliar a robustez de um novo procedimento para alocar às famílias nos grupos de comparação e que é alternativo à alocação utilizada com os dados da pesquisa de campo AIBF, com o objetivo de analisar o impacto dos resultados do programa Bolsa Família sobre os indicadores da educação das crianças de 7 a 14 anos. Assim, algumas variações ou diferenças de informação podem alterar a significância estatística dos impactos ou diferenciais dos resultados da avaliação.

Neste sentido, considerando que o trabalho compara os resultados provenientes de dois tipos de informação para alocar às famílias nos grupos de comparação (grupo de beneficiários PBF e grupo de comparação 2), é importante destacar a consideração que será utilizada para referir-se a tal análise comparativa. Analisar a sensibilidade dos resultados surge como uma opção importante para analisar os resultados de ambos os tipos de alocações. Para o presente trabalho, a sensibilidade consiste em determinar em que medida são sensíveis os resultados de impacto em relação, às mudanças que acontecem por parte das famílias, quando são alocadas nos grupos de comparação, segundo a base de dados da pesquisa de campo AIBF ou segundo a base de dados dos registros administrativos do

CadÚnico. O objetivo principal de analisar a sensibilidade dos resultados é estabelecer os diferenciais dos indicadores da educação das crianças de 7 a 14 anos, que modificam seu sentido ou direção, quando se utilizam os dois tipos de alocações das famílias.

Deve-se enfatizar que a análise comparativa da sensibilidade dos resultados partirá do pressuposto que tanto os dados da pesquisa de campo AIBF, como os registros administrativos do CadÚnico, apresentam informação fidedigna em relação aos grupos de comparação; mas considerando os viés, naturalmente aceitáveis, pela configuração como os dados foram coletados, é possível tomar como padrão, para analisar a sensibilidade dos resultados, a qualquer das alocações utilizadas.

Considerando que a análise de sensibilidade em qualquer método quantitativo consiste em, avaliar as mudanças dos dados ou métodos estudados para medir a incidência destes sobre os resultados, esta é utilizada nas conclusões finais dos trabalhos. A análise de sensibilidade neste estudo é parte integrante do planejamento do trabalho da análise comparativa de duas fontes de dados (AIBF e CadÚnico) e fornece informações sobre a importância de cada uma deles sobre os resultados. Com sua ajuda, é possível avaliar de que maneira incertezas nas fontes de dados influenciam sobre os resultados das avaliações de impacto efetuadas. Desta maneira, a análise de sensibilidade se torna ferramenta importante também para medir os possíveis erros que envolvem a utilização dos instrumentos de medidas para coletar a informação, da influencia dos operadores de coleta de dados e sobre o processo de medidas. Estes possíveis erros podem ser: sistemáticos, e ocorrem quando há problemas no método empregado; problemas com instrumentos de medidas; erros aleatórios que ocorrem quando há imperícia do operador; e erro de interpretação das informações. Em qualquer incerteza mencionada, para a presente tese é preciso adotar uma idéia substantiva que melhor represente a grandeza e uma margem de erro dentro da qual deve estar compreendido o valor de cada fontes de dados, neste caso é a análise da sensibilidade dos erros de medidas mencionados, que permite determinar em forma substantiva o valor e o seu respectivo desvio quando se compara ambos as fontes de dados utilizadas para alocar às famílias nos grupos de tratamento e comparação 2. Embora não se utilize uma medida estatística para medir a sensibilidade dos resultados e por conseguinte, os possíveis erros de medida, o trabalho realiza uma comparação dos resultados determinando as diferenças encontradas e discute a importância de utilizar dados de pesquisa de campo e registros administrativos para avaliar os resultados de

impacto dos programas sociais. Além disso, com a análise da sensibilidade dos resultados, as possíveis variações serão analisadas individualmente para cada indicador e região determinada, isto é, analisa-se a sensibilidade dos resultados devido à modificação dos dados assumindo que todos os outros indicadores e regiões permanecem sem alteração alguma.

Além disso, continuando com a análise comparativa dos resultados utilizando dois tipos de fonte de informação, realiza-se uma das aplicações mencionadas na seção 2.5.3, associada às técnicas. Uma dessas técnicas não-experimentais útil para avaliar os resultados de impacto do PBF, é aplicação da regressão descontinua *Sharp* (RD), que utiliza as descontinuidades no processo de alocação ao programa para identificar o efeito causal, em que se supõe uma variável continua pré-tratamento influi nas variáveis resultados assim como na variável que define a participação no programa, a qual afeta também ao resultado. A aplicação desta técnica somente será possível, quando se utiliza a variável continua, que para este estudo é a "renda familiar", dos registros administrativos do CadÚnico, que se pressupõe é pré-tratamento e não está influenciada pelos renda que recebem os beneficiários, mas que influiria nos resultados de impacto do PBF e na participação das famílias beneficiaria neste programa.

5.3 Variáveis e indicadores utilizados para a avaliação dos impactos na educação do PBF.

Avaliar os efeitos do programa de transferência de renda Bolsa Família sobre a variável de resultado Y (indicadores da educação)[38] segundo os dois tipos de alocações das famílias é nosso objetivo. Supondo que esta variável dependa de um conjunto de variáveis exógenas, *X*, e de uma variável de tratamento, *D,* então o problema da avaliação será dado por:

$$Y = \alpha + \beta D + \theta X + \mu \qquad (5.1)$$

em que, D_i representa uma variável *dummy* para tratamento, que assume o valor 1 se a família recebe o Bolsa-Família e 0 caso contrário, ao α, β e θ são parâmetros; *X* representa

[38] No trabalho consideramos como resultado às variáveis dependentes de educação que desejamos avaliar.

as variáveis de controle, enquanto μ_i é o termo de erro, o valor estimado de β fornece o impacto do tratamento.

Na equação 5.1, o lado direito é constituído por uma série de características individuais, familiares e domiciliares.Os métodos utilizados para avaliar o efeito do PBF sobre a educação das crianças de 7 a 14 anos, são: primeiro, o método de *matching* por escore de propensão (PSM) para as famílias alocadas nos grupos de comparação segundo a pesquisa de campo AIBF e estimações de PSM para os grupos alocados segundo o registro administrativos CadÚnico; e segundo, como alternativa considerando-se os resultados do relacionamento ou pareamento das bases de dados, estima-se o desenho *Sharp* da regressão descontinua (RD).

5.3.1 As variáveis dependentes

Considerando que, no componente educacional do programa Bolsa-Família, há uma condicionalidade de que as crianças entre 6 e 15 anos freqüentem regularmente a escola, espera-se que os beneficiários do programa apresentem efeitos positivos sobres os indicadores da educação.

Para Schultz (2000), existem dois pontos que tornam importante a educação; o primeiro está relacionado com o arcabouço do capital humano, no qual se considera que a educação é custeada pelas famílias para aumentar a produtividade futura do estudante; e segundo, as famílias pobres têm mais restrições para investir na escolaridade de suas crianças em um nível socialmente desejável devido a limitações de crédito e informação. Desta forma, o programa Bolsa Família visa compensar estas limitações, transferindo recursos públicos diretamente às famílias pobres (OLIVEIRA et al, 2007).

Muitos estudos têm analisado importância dos antecedentes familiares na determinação dos resultados educativos dos adolescentes. Behrman, Duryea E Székely (1999) analisam a influência do background familiar de forma direita sobre os ganhos educativos do adolescente. Sobre a produção familiar do capital humano, Gary Becker (1993) foi um dos primeiros em destacar que as mercadorias domiciliares são produzidas por uma combinação de bens e trabalho doméstico. Assim, revela que os investimentos de recursos humano na nutrição, saúde e educação refletem decisões de comportamento do nível familiar. Uma das evidências que explicam esta relação são os resultados do PROGRESA ,

nos quais os estudantes em idade escolar beneficiários deste programa têm diminuído as taxas de evasão e taxa de repetência, e melhoraram o grau de progressão e de re-iniciação de estudos entre aqueles que deixaram os estudos (Behrman et al., 2001).

No caso das famílias com crianças em idade escolar, os diferenciais do PBF podem ser mensurados pelas variáveis de desempenho escolar destacado por Berhman et al (2001), que a seguir, apresentam-se na TAB 5.1. no qual se mostram os indicadores para avaliar os diferenciais do PBF na educação das crianças de 7 a 14 anos:

TABELA 5.1 – Variáveis dependentes: Indicadores para avaliar os diferenciais do PBF na educação. (crianças entre 7 e 14 anos de idade).

Variáveis	Descrição
Não deixaram de ir à escola no último mês (ou o complemento deste)	Proporção de meninas e meninos no domicílio que não deixaram de ir à escola no último mês.
Evasão ou abandono	Proporção de meninas e meninos no domicílio que evadiram do sistema de ensino entre 2004 e 2005.
Progressão	Proporção de meninas e meninos no domicílio que foram aprovados entre 2004 e 2005.
Alocação entre trabalho e estudo	Proporção de meninas e meninos no domicílio que declararam só estudar atualmente, vis-à-vis aqueles que declararam só trabalhar, trabalhar e estudar e não trabalhar nem estudar.
Retenção	Proporção de meninas e meninos que foram reprovados entre 2004 e 2005.
Fonte: Dados tomados a partir do Oliveira, *et al*, 2007	

5.3.2 Variáveis Independentes

Tal como foi explicado no capitulo 3, quando se estima o escore de propensão através do modelo probit, o cálculo deve incluir variáveis preditoras que influenciam a participação no programa. Além disso, as variáveis utilizadas devem ter uma estreita relação com a elegibilidade da pessoa ou família para participar do programa e com as variáveis de educação das crianças de 7 a 14 anos, porque através destes, seus valores médios são "balanceados" entre os grupos de tratamento e controle dentro da cada bloco de famílias.

TABELA 5.2 – Variáveis independentes: variáveis utilizadas na especificação dos modelos equilibrados do Escore de propensão e na Regressão descontínua, para avaliar os diferenciais do PBF na educação.

Atributos do chefe de família:	
Raça do chefe de família	Branca Não Branca
Sexo do chefe de família	Masculino Feminino
Escolaridade do chefe de família	Até 3 anos de estudos* Até 4 anos de estudos* Até 7 anos de estudos*
Idade do chefe de família	Menor e igual há 50 anos Mais que 50 anos
Altura em metros do chefe de família	Medida em metros (mts)
Escolaridade da mãe do chefe de família	Mãe alfabetizada Mãe não alfabetizada
Tempo de permanência do chefe de família no município	Menos de 10 anos* Menos de 5 anos*
Tempo de permanência do chefe de família na área rural.	Viveu até os 14 anos Não viveu até os 14 anos
Características da família:	
Número de membros da família	Número de membros no domicilio
Crianças entre 0 a 3 anos de idade	Proporção de crianças de 0 a 3 anos
Crianças entre 0 a 6 anos de idade	Proporção de crianças de 0 a 6 anos
Crianças mulheres 7a14/ criança 0 a 14 anos	Proporção crianças mulheres 7 a 14/ crianças 0 a 14
Casal com filhos até 14 anos	O Casal tem filhos até 14 anos O Casal não tem filhos até 14 anos
Presença de pessoas de 60 anos ou mais	Há pessoa de 60 anos e mais no domicílio Há pessoa menor de 60 anos no domicílio.
Características do domicilio:	
Qualidade de domicilio[1]	Qualidade inferior* Qualidade media*
Área de residência do domicilio	Urbana Rural
Região de residência do domicílio	Nordeste* Norte – Centro Oeste*

* Para cada um destas categorias criara-se uma variável *dummy*
[1] Esta variável foi gerada através do método Grade of Membership (GOM), com três categorias para a qualidade das condições dos domicílios, classificadas em: muito boa, regular e ruim
Fonte: Dados tomados a partir do Oliveira *et al*, 2007.

No nosso caso, o escore de propensão estimado forneceu a probabilidade estimada de participação no Programa Bolsa Família (PBF) de uma determinada família, e utilizando estes valores foi realizado o *matching* entre os grupos de tratamento e comparação. A

inclusão de variáveis individuais, familiares e domiciliares garante o suposto de ortogonalidade ao tratamento, e que o *matching* das famílias as torne comparável em termos das características observáveis.

5.4 Descrição dos dados e das variáveis incluídas no modelo

A seguir realiza-se a descrição dos grupos de tratamento e comparação, considerando as duas fontes utilizadas para alocar às famílias nos grupos.

Entre as famílias elegíveis e não elegíveis, a amostra é constituída de 15.426 domicílios. Excluindo os não elegíveis para análise de impacto na educação, a amostra é de 12.514 domicílios[39]. Segundo a pesquisa AIBF, estes estão distribuídos em 35% como beneficiários do PBF (Tratamento), 28% beneficiários de outros programas (Comparação 1) e 37% não são beneficiários (comparação 2). A distribuição segundo a alocação com o CadÚnico, foi de 43% beneficiários do PBF, 16% beneficiários de outros programas e 41% não são beneficiários ou não cadastrados (ou não encontrados no cadastro). Analisando por região, Nordeste, Norte – Centro-Oeste, e Sul – Sudeste, observa-se que a distribuição dos domicílios por grupos de comparação é similar ao como um todo Brasil, tanto para os grupos obtidos pela alocação segundo a pesquisa de campo AIBF, como a alocação segundo o relacionamento com o CadÚnico.

[39] Excluíram-se domicílios que já receberam qualquer benefício, mas não recebem mais, e domicílio cuja renda domiciliar per capita líquida dos valores recebidos das transferências é maior que R$200,00 (duzentos reais). Este corte de renda, acima do limite máximo de elegibilidade oficial, foi utilizado para garantir a representatividade amostral em todos os grupos, inclusive o de tratamento.

TABELA 5.3 – Distribuição de famílias, segundo grupos de comparação Brasil e Regiões, 2005.

Regiões	Tratamento [2]		Comparação 1 [2]		Comparação 2 [2]		Total[3]
	AIBF	CadÚnico	AIBF	CadÚnico	AIBF	CadÚnico	
Brasil [1]	4.375 (34,96%)	5.361 (42,84%)	3.450 (27,57%)	1.967 (15,72%)	4.689 (37,47%)	5.186 (41,44%)	12.514
Norte e Centro-Oeste [1]	1.221 (33,26%)	1.586 (43,2%)	1.050 (28,6%)	583 (15,88%)	1.400 (38,14%)	1.502 (40,92%)	3.671
Nordeste [1]	1.616 (36,77%)	1.900 (43,23%)	1.214 (27,62%)	760 (17,29%)	1.565 (35,61%)	1.735 (39,48%)	4.395
Sudeste e Sul [1]	1.538 (34,58%)	1.875 (42,15%)	1.186 (26,66%)	624 (14,03%)	1.724 (38,76%)	1.949 (43,82%)	4.448

Fonte: AIBF, 2005 e CadÚnico 2005.
1)Corte de renda domiciliar per capita considerado como critério de elegibilidade até R$ 200.00
2) O grupo Tratamento corresponde aos beneficiários do programa Bolsa Família; o grupo Comparação 1 corresponde aos beneficiários de outros programas sociais (exclusive o Bolsa Família); e o grupo Comparação 2 corresponde àqueles que não recebem nenhum tipo de programa de transferência de renda.
3) No Total estão incluídos todos os domicílios elegíveis

Os resultados da distribuição das famílias nos grupos de comparação utilizando o CadÚnico, indicam que as famílias beneficiárias do PBF (tratamento) incrementou-se como produto desta alocação, quando se compara com a alocação segundo a pesquisa de campo AIBF. Isto procede dos resultados observados no capitulo 4, em que as famílias que antes pertenciam ao grupo de outros benefícios e cadastrados sem benefício foram re-classificados no grupo de beneficiários do PBF utilizando o CadÚnico. Como conseqüência destes resultados, o grupo denominado Comparação 1 diminui o número de casos. Contudo, sugere-se que as famílias beneficiários dos PBF segundo os dados do CadÚnico, é mais robustos, considerando que é possível encontrar maior confiabilidade nos registros administrativos em relação ao beneficio que recebem as famílias, comparando com as declarações que as famílias entrevistadas informaram na pesquisa de campo AIBF.

Por outro lado, da amostra de 12.514 domicílios elegíveis para medir o impacto do PBF na educação, selecionou-se apenas os famílias com membros crianças de 7 a 14 anos, resultando em 8.407, distribuídos segundo a pesquisa AIBF, em 38,6% famílias beneficiárias do Programa Bolsa Família – PBF (Tratamento), 30,8% beneficiárias de outros programas (Comparação 1) e 30,6% não são beneficiárias (comparação 2). Da mesma forma, a alocação das famílias nos grupos segundo o relacionamento com o CadÚnico, indicou 47,4% famílias beneficiárias do PBF, 17,4% beneficiárias de outros programas e 35,2% são não-beneficiárias ou não-cadastrados. Estes resultados mostram também que utilizando o CadÚnico para alocar às famílias, incrementa-se o número de

caso, no grupo de tratamento (9%) e comparação 2 (5%), enquanto, o grupo de comparação 1, diminui em 13%.

Com o objetivo de medir e comparar os resultados de impacto na educação das crianças de 7 a 14 anos que pertencem aos domicílios beneficiários do PBF, só foi utilizado o grupo dos domicílios que não recebem nenhum benefício, isto é, comparam-se crianças pertencentes a dois domicílios do grupo de tratamento e comparação 2. Assim, analisam-se de forma comparativa as variáveis de impacto na educação entre os grupos de beneficiários e não-beneficiários do PBF, considerando-se, a alocação dos grupos obtidos diretamente da pesquisa de campo AIBF e o do relacionamento com os registros administrativos do CadÚnico.

A tabela 5.4. descreve as variáveis de impacto. Observa-se que no Brasil 88,27% das crianças de 7 a 14 anos, não deixaram de ir à escola ou creche em outubro de 2005. Comparando os resultados segundo as alocações utilizadas, os resultados deste indicador, não apresentam grandes diferenças, apenas uma pequena diferença nas crianças femininas de Brasil para o grupo de Tratamento e Comparação 2. Analisando comparativamente os grupos de comparação: tratamento e comparação 2, os resultados indicam maior porcentagem para o grupo de tratamento, tanto nos alocados segundo a pesquisa de campo AIBF, como os alocados com o CadÚnico. Além disso, a diferenças que existem entre ambos os grupos é estatisticamente significativa para Brasil como um todo e para as crianças femininas também para os dois procedimentos de alocação utilizados. Os resultados sobre as crianças que não deixaram de ir a escola indicam, que utilizando ambas as alocações este indicador distribui-se de forma similar, sugerindo robustez à utilização de duas fontes de informação para alocar às famílias e descrever as características de crianças em relação a este item.

TABELA 5.4 – Indicadores para avaliar os diferenciais do PBF na educação de crianças de 7 a 14 anos, segundo grupos de comparação, Brasil e Regiões, 2005 (em%).

Variáveis de Impacto	Grupos AIBF		P-value	Grupos CadÚnico		P-value	Total
	Tratamento	Comparação 2		Tratamento	Comparação 2		
Não deixo de ir à escola no ultimo mês							
Brasil	89,73	86,01	<0,01	89,52	87,70	<0,01	88,27
Homem	89,14	89,07	NS	90,24	90,56	NS	88,78
Mulher	90,38	83,12	<0,01	88,74	85,11	<0,01	87,70
Evasão ou abandono							
Brasil	1,05	2,12	<0,01	1,22	2,35	<0,01	1,59
Homem	0,84	2,48	<0,01	0,94	2,51	<0,01	1,35
Mulher	1,27	1,79	NS	1,53	2,22	<0,10	1,85
Progressão							
Brasil	82,81	87,33	<0,01	83,58	86,59	<0,01	86,46
Homem	80,00	86,59	<0,01	80,59	84,90	<0,01	85,16
Mulher	85,90	87,98	<0,10	86,77	88,07	NS	87,88
Alocação entre trabalho e estudo							
Brasil	91,87	95,06	<0,01	92,37	94,23	<0,01	94,15
Homem	90,71	93,75	<0,01	91,53	92,38	NS	93,44
Mulher	93,14	96,30	<0,01	93,29	95,93	<0,01	94,94
Repetência							
Brasil	16,01	11,19	<0,01	15,01	12,14	<0,01	12,22
Homem	19,16	12,50	<0,01	18,41	14,10	<0,01	13,93
Mulher	12,54	10,05	NS	11,39	10,43	NS	10,37

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: A coluna Total refere-se a valores para todos os domicílios com crianças de 7 a 14 anos.
O grupo Tratamento corresponde aos beneficiários do programa Bolsa Família e o grupo Comparação 2 corresponde àqueles que não recebem nenhum tipo de programa de transferência de renda.
p-value: é a probabilidade de se observar um resultado tão ou mais extremo que o da amostra, supondo que a hipótese nula seja verdadeira.
NS: Não significante.

Em relação à evasão escolar entre 2004 e 2005, observa-se que, aproximadamente 2% das crianças de 7 a 14 anos abandonaram a escola em 2005. Considerando os resultados segundo alocação utilizada, encontram-se pequenas diferenças entre os resultados, sendo que, a maior diferença está entre as crianças femininas do grupo de comparação 2 (aproximadamente de 0,4%). Analisando o diferencial de impacto entre as crianças do grupo de Tratamento e comparação 2, observa-se diferenças mais acentuadas, para as criancas masculinas de Brasil, isto nos dois tipos de alocações utilizadas. No entanto,

segundo a significância estatística esta é diferente para Brasil como um todo e para as crianças masculinas, em ambas as alocações. Por outro lado, no caso das meninas, apesar, de que as diferenças entre o Tratamento e comparação 2, sejam similares para ambas das alocações, esta só apresenta diferença significativa para os alocados segundo o CadÚnico.

A progressão indicou que 86,46% dos alunos de 7 a 14 anos de idades foram aprovados em 2005 em todo Brasil. Os resultados comparando as alocações utilizadas indicaram uma diferença mais acentuada entre as crianças pertencentes do grupo de Comparação 2 (1,7%), nos outros grupos as diferenças foram mínimas e nem atingiram o 1%. Considerando a diferença entre o resultado do grupo de tratamento e comparação 2, a as maiores diferenças foram observadas Brasil como um todo e os meninos homens, isto para as duas alocações utilizadas, sendo maior para os meninos e com alocação segundo a pesquisa de campo AIBF (6,6%). Além disso, os resultados considerando a hipótese da diferença mostram significância estatística para Brasil como um tudo e para os meninos. No referente às meninas, a diferença é maior para alocação segundo pesquisa de campo AIBF, resultado refletido na significância estatística, á qual só observa-se diferença estatísticas significativa com esta alocação.

Na Alocação entre trabalho e estudo, o percentual das crianças de 7 a 14 anos que estavam só estudando situa-se acima de 90%, enquanto as crianças que apenas trabalhavam está formada por uma pequena parcela. Comparando os resultados segundo alocação das famílias, observam-se diferenças não acentuadas, no entanto é possível distinguir uma diferença de até 1,4% entre as crianças masculinas do grupo de “Comparação 2”. Em relação aos resultados entre o grupo de tratamento e comparação 2 observa-se que o percentual de crianças de 7 a 14 anos que apenas estudava no grupo de tratamento é menor, isto para ambas das alocações, apresentando maior diferença entre esses grupos, com os resultados obtidos da alocação com os dados da pesquisa de campo AIBF. No entanto, as diferenças encontradas foram estatisticamente significativas, para os dois procedimentos de alocação de grupos utilizados.

O ultimo indicador de impacto refere-se à retenção escolar, no qual se observa que apenas 12,22% dos alunos repetiram o ano escolar em Brasil. Considerando os resultados segundo alocação utilizada, observa-se que a diferença mais destacável entre as crianças masculinas do grupo de “Comparação 2” (1,6%), nos outros grupos a diferença está em torno de 1% ou menos, a qual pode ser considerada aceitável. Analisando comparativamente os

resultados dos grupos de Tratamento e Comparação 2, o percentual da retenção escolar é maior no grupo de tratamento nos dois tipos de alocação utilizada. Além disso, as maiores diferenças comparando esses grupos, observa-se nos resultados do produto da alocação segundo pesquisa de campo e, sobretudo nas crianças do sexo masculinos (6%). No entanto, as diferenças encontradas entre os grupos de comparação são estatisticamente significativas no Brasil como um todo e nas crianças masculinas. No caso das crianças femininas, embora se observe diferenças, estas não são confirmadas com o teste de hipótese, nem uma das alocações utilizadas.

A seguir serão descritas as variáveis independentes da especificação dos modelos equilibrados do Escore de Propensão e da Regressão descontínua. Na TAB 5.5 encontram-se os resultados para Brasil e para os grupos de tratamento e comparação 2.

TABELA 5.5 – Variáveis independentes para a especificação dos modelos equilibrados do Escore de Propensão e na Regressão descontínua para avaliar os diferenciais do PBF na educação de crianças de 7 a 14 anos, segundo grupos de comparação, Brasil. 2005.

(continua)

Variáveis de Impacto	Grupos AIBF		*p-value*	Grupos CadÚnico		*p-value*	Total[1]
	Tratamento	Comparação 2		Tratamento	Comparação 2		
Medias							
Altura em metros da mulher responsável	1,55	1,54	<0,10	1,55	1,54	<0,10	1,55
Altura em metros do homem responsável***	1,34	1,28	<0,01	1,33	1,26	<0,01	1,31
Membros do domicílio***	4,93	4,43	<0,01	4,83	4,44	<0,01	4,49
Porcentagens (%)							
Chefe não-branco***	64,84	49,51	<0,01	61,68	49,78	<0,01	51,21
Chefe mulher	37,25	34,89	<0,10	36,06	35,30	NS	33,85
Chefe com até 3 anos de estudos***	48,24	31,21	<0,01	44,19	34,69	<0,01	34,07
Chefe com até 4 anos de estudos***	66,69	48,36	<0,01	62,49	53,03	<0,01	50,64
Chefe com até 7 anos de estudos***	81,76	65,41	<0,01	80,00	68,31	<0,01	66,5
Chefe com menos de 50 anos***	84,39	75,83	<0,01	82,95	76,23	<0,01	77,43
Chefe menos de 10 anos no município***	13,71	16,55	<0,01	14,00	17,37	<0,01	14,12
Chefe menos de 5 anos no município**	8,92	8,79	NS	7,46	10,49	<0,01	8,13

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: [1] A coluna Total refere-se a valores para toda a população, incluindo os não elegíveis.
O grupo Tratamento corresponde aos beneficiários do programa Bolsa Família e o grupo Comparação 2 corresponde àqueles que não recebem nenhum tipo de programa de transferência de renda.
p-value: é a probabilidade de se observar um resultado tão ou mais extremo que o da amostra, supondo que a hipótese nula seja verdadeira.
NS: Não significante.

Considerando as alocações utilizadas, ressalta-se que as informações descritas a seguir, sobre as variáveis independentes, indicam resultados similares para os dois procedimentos de alocação utilizados. Isto é importante, porque agora é possível dizer que tanto as variáveis dependentes, como independentes não mostraram grandes diferenças entre as alocações utilizadas, mostrando que os resultados descritivos são robustos aos tipos de fontes de informação utilizada: pesquisa de campo e registros administrativos.

Analisando as variáveis cujo indicador é a media, observa-se que a altura média da mulher responsável e o número de membros do domicílio é similar para o grupo de tratamento e comparação 2, já a altura média do homem responsável é maior nos tratados em relação ao grupo “Comparação 2”. A diferença estatística das médias destas variáveis entre os grupos de comparação é significativa para a altura em metros do homem responsável e para os membros médios no domicílio. Resultados observados para os dois procedimentos de alocação.

TABELA 5.5 – Variáveis independentes para a especificação dos modelos equilibrados do Escore de Propensão e na Regressão descontínua para avaliar os diferenciais do PBF na educação de crianças de 7 a 14 anos, segundo grupos de comparação, Brasil. 2005.

(fim)

Variáveis de Impacto	Grupos AIBF		*p-value*	Grupos "CadÚnico"		*p-value*	Total[1]
	Tratamento	Comparação 2		Tratamento	Comparação 2		
Porcentagens							
Chefe viveu até os 14 anos em área rural	54,19	39,5	<0,01	52,81	39,97	<0,01	41,86
Mãe de chefe alfabetizada	47,37	55,93	<0,01	48,13	53	<0,01	54,31
Mulher responsável presente	99,22	97,99	<0,01	98,97	98,08	<0,01	98,55
Homem responsável presente	79,74	75,59	<0,01	79,19	74,81	<0,01	77,56
Proporção de crianças entre 0 e 6 anos de idade	13,89	11,45	<0,01	13,26	11,07	<0,01	9,97
Razão: Crianças mulh. 7 -14 / Criança 0 – 14	47,36	52	<0,01	47,87	53,15	<0,01	46,78
Presença de pessoa de 60 anos ou mais	7,88	11,35	<0,01	9,05	11,67	<0,01	12,79
Casal com filhos até 14 anos	72,47	60,83	<0,01	70,79	59,57	<0,01	64,61
Domicílio de qualidade inferior	35,92	19,21	<0,01	33,86	19,73	<0,01	26,51
Domicílio de qualidade média	19,76	18,03	<0,10	19,96	18,30	<0,10	19,52
Domicílio em área urbana***	75,25	82,93	<0,01	76,68	82,00	<0,01	82,04
Região Nordeste***	41,07	23,43	<0,01	37,21	24,58	<0,01	27,35
Região Norte ou Centro-Oeste***	12,74	17,22	<0,01	14,03	18,04	<0,01	14,93

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: [1] A coluna Total refere-se a valores para toda a população, incluindo os não elegíveis.
O grupo Tratamento corresponde aos beneficiários do programa Bolsa Família e o grupo Comparação 2 corresponde àqueles que não recebem nenhum tipo de programa de transferência de renda.
p-value: é a probabilidade de se observar um resultado tão ou mais extremo que o da amostra, supondo que a hipótese nula seja verdadeira.
NS: Não significante.

Entre as variáveis descritas pela proporção, as que apresentam percentual similar no grupo de “Tratamento” e “comparação 2”, foram: mulher como chefe de domicílio, chefe domicílio menos de 5 anos no município, mulher responsável presente no domicílio, proporção de crianças entre 0 e 6 anos de idade e domicílio de qualidade média. Nas variáveis que apresentam percentual maior no grupo de “Tratamento” em relação ao grupo “Comparação 2”, foram: chefe de domicílio não-branco, com até 3, 4 e 7 anos de estudos,

com menos de 50 anos e que viveu até os 14 anos em área rural; o homem responsável do domicílio; crianças entre 0 e 13 anos de idade presente no domicílio; casal com filhos até 14 anos; domicílio de qualidade inferior; e domicílio de região Nordeste". Por outro lado, nas variáveis no qual o percentual é menor no grupo de tratamento comparado como o de comparação 2, foram, chefe de domicílio menos de 10 anos no município; mãe de chefe alfabetizada; razão de crianças mulheres 7 a 14 por criança 0 a 14; presença de pessoa de 60 anos ou mais no domicílio; e domicílio em área urbana e na região Norte ou Centro-Oeste. Estes resultados foram os mesmos nos grupos alocados segundo a pesquisa de campo AIBF, como nos alocados segundo o relacionamento com o CadÚnico.

As variáveis, na qual os resultados são diferentes entre as alocações utilizadas, foram: chefe do domicílio não branco, com até 3 e 4 anos de estudos; e domicílios na Região Nordeste, variáveis nas qual a diferença atingem em torno de 4%, sendo maior no grupo de tratamento obtido da alocação segundo a pesquisa de campo AIBF.

Uma forma de confirmar se as diferenças são significativas para ambas as alocações utilizadas, é realizar o teste de hipóteses das diferenças de médias ou percentuais entre os grupos de comparação. Os resultados deste teste indicam diferença estatisticamente significativa para a maioria das variáveis, e para ambos os procedimentos de alocações utilizadas, exceto na variável chefe de domicílio menos de 5 anos no município, a qual é estatisticamente diferentes, apenas nos grupos alocados segundo o paramento com o CadÚnico.

Considerando os resultados dos grupos de Tratamento e Comparação 2, produtos da alocação dos grupos da pesquisa de campo AIBF e a alocação de grupos segundo paramento com CadÚnico, estes indicariam que não existem diferenças acentuadas nem contraditórias entre as duas fontes de obtenção dos grupos, resultado que sugere, que as informações obtidas diretamente da alocação da pesquisa de campo AIBF são compatíveis com as informação obtidas da alocação dos grupos do relacionamento com os registros administrativos, e por conseguinte, robustas as dois tipos de fonte de dados utilizados, embora estes sejam dados descritivos, mas que evidenciam a boa qualidade das dados.

5.5 Resultados da aplicação do modelo de impacto na educação do PBF.

Nesta seção para avaliar o efeito do PBF sobre a educação das crianças de 7 a 14 anos, considera-se primeiro, os resultados da aplicação do método *matching* de escore de propensão (PSM) para as famílias alocadas nos grupos de comparação; e segundo, os resultados correspondente à estimação do desenho *Sharp* da regressão descontínua.

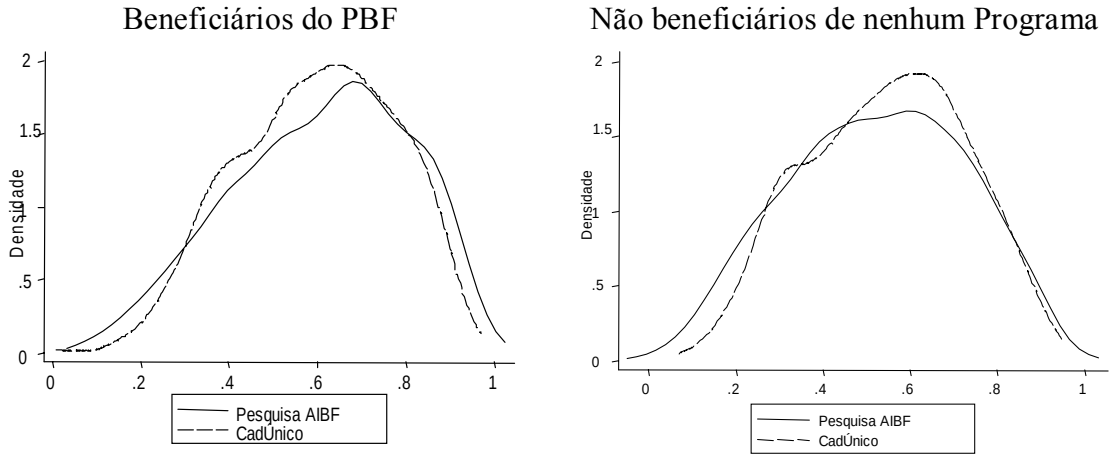
5.5.1 Resultados do método de pareamento por escore de propensão.

5.5.1.1 Análise do balanceamento com o método pareamento por escore de propensão.

Um primeiro passo em uma análise estatística consiste em descrever, a distribuição das variáveis estudadas e em particular, dos dados que definem as diferenças e similitudes quando se compara dois tipos de informação. Uma informação importante para uma inicial avaliação dos dois tipos de alocação utilizada, é o pareamento por escore de propensão (PSM), que é a probabilidade condicional de um indivíduo participar do programa dadas as suas características individuais ou domiciliáreis[40]. O PSM é um resultado importante para avaliar o balanceamento dos domicílios nos grupos de comparação de tratamento e comparação 2, quando a participação dos indivíduos ou famílias não foi alocada de forma aleatória em um programa. Com base neste resultado, a seguir, compara-se a distribuição de densidade do PSM com as famílias alocadas segundo a pesquisa de Campo AIBF e registros administrativos CadÚnico, com o objetivo de verificar se existem algumas diferenças sistemáticas entre os dois tipos de alocações. Além disso, a comparação e análise das distribuições, diferenciam-se para ambos os grupos de comparação utilizadas para o estudo: tratamento (beneficiários do PBF) e comparação 2 (não beneficiários de nenhum programa).

[40] No APÊNDICE III mostram-se as variáveis utilizadas na especificação dos modelos equilibrados do escore de propensão.

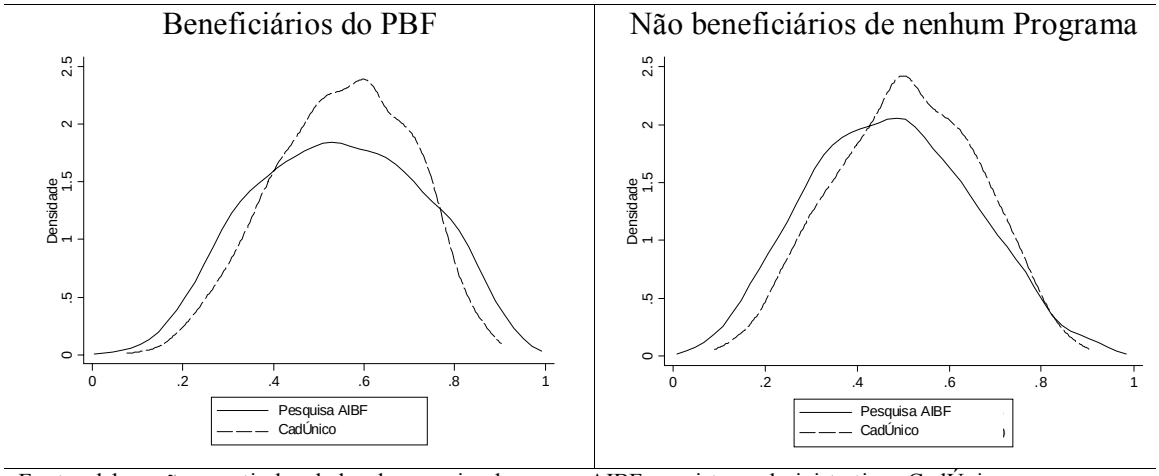
GRAFICO 5.1 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Corte de renda até R$50,00. Brasil. 2006



Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.

No GRAF 5.1 observa-se a distribuição de densidade da estimação do PSM para as famílias de todo Brasil com corte de renda até R$50,00. O comportamento da distribuição para esta população indica pequenas diferenças enquanto a distribuição dos PSM, isto é, diferenças não sistemática, são observadas, mas algumas variações produzidas pela sensibilidade dos resultados apresentam-se nos extremos da distribuição, as quais são advertidas pela utilização das diferentes alocações. Estes resultados são observados, tanto entre o grupo de Tratamento, como no grupo de Comparação 2.

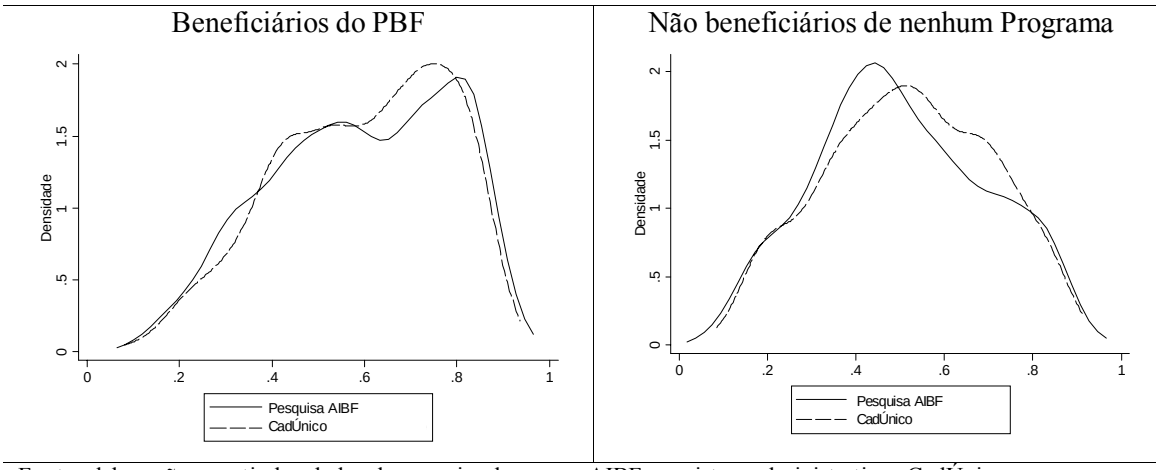
GRAFICO 5.2 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Corte de renda até R$100,00. Brasil. 2006.



Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.

Em relação à distribuição de densidade da estimação do PSM para os domicílios de todo Brasil com corte de renda até R$100,00 (Ver GRAF 5.2), algumas diferenças são observadas entre as distribuições e em ambas as alocações utilizadas. No que se refere ao grupo de Tratamento, a maior diferença observa-se entre os extremos e na posição central dos dados, assim, parece ser que distribuição segundo a alocação com o CadÚnico, é mais concentrado e, portanto, as diferenças são apresentadas com maior intensidade no que se refere ao nível da estimação do PSM, sendo maior com os alocados segundo o CadÚnico; sobretudo no grupo de tratamento ou dos beneficiários do PBF.

GRAFICO 5.3 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Corte de renda até R$200,00. Brasil. 2006.



Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.

O último grupo de população, são as famílias de todo Brasil com corte de renda até R$200,00, pequenas diferenças são observadas quando se analisa a distribuição para os dois tipos de alocação de famílias utilizadas. Neste caso, os extremos apresentam comportamentos parecidos, e apenas observam-se diferenças nos pontos mais altos da curva, mas que não mudam a configuração da curva. Este é valido para os dois grupos de comparação, tratamento ou beneficiários do PBF e comparação 2 ou não beneficiários (Ver TAB 5.3).

Considerando as distribuições apresentadas acima, é possível dizer que as variações observadas na comparação da distribuição da estimação do PSM utilizando a alocação segundo a pesquisa de campo AIBF e registros administrativos CadÚnico, não são fortemente afetadas no seu comportamento, assim, é possível afirmar que, apesar da

existência de algumas variações nos extremos das curvas, a configuração da distribuição de densidade é mantida para ambas as alocações em cada grupo de comparação e corte de renda estudada. A distribuição das estimações do PSM para as outras regiões e corte de renda, não são apresentadas nesta seção, mas estes podem ser encontradas no APÊNDICE IV. Os resultados da distribuição para estes casos apresentam também distribuições similares no que se refere às duas alocações de famílias utilizadas, e diferenças sistemáticas e acentuadas não são encontradas.

5.5.1.2 Análise e discussão dos resultados dos indicadores de impacto na educação

Neste estudo o escore de propensão foi estimado utilizando um modelo paramétrico de escolha binária, um modelo probit. Como visto no capitulo 3 o calculo do escore de propensão, será realizado utilizando um conjunto de variáveis explicativas obedecendo à condição de equilíbrio (Oliveira et al, 2007). Estas variáveis procuram caracterizar as condições do domicílio em termos da elegibilidade ao programa e em alguns casos servir de controle para o cálculo dos efeitos do tratamento sobre os tratados (Ver TAB. 5.2). Para a técnica de pareamento, serão utilizados os 3 métodos com maior robustez: o método do vizinho mais próximo (*Nearest Neighbour Matching* – NNM) com reposição, o do raio (*Radius Matching* - RM) e estratificado (SM), mas reportamos apenas os diferenciais considerando a técnica do NNM com reposição[41] e quando ao menos em 2 dos 3 métodos utilizados é estatisticamente significativo.

Para a análise dos ATT's considerou-se a magnitude do valor estimado, o sinal e a significância estatística, lembrando que o método Escore de propensão de *matching* calcula o ATT subtraindo o valor médio do resultado para os não tratados do valor estimado para os tratados. Como a grande contribuição ao presente estudo será conferida a análise comparativa dos resultados de impacto das variáveis utilizadas na educação nos grupos alocados segundo a pesquisa de campo AIBF e segundo o relacionamento de bases de dados com o CadÚnico, tal como se explicou na seção 5.1.

---

41 Isto é, porque o método tem: maior facilidade de interpretação dos resultados, utilizar maior número possível de observações do tratamento uma vez que a amostra dos grupos de comparação é menor que a de tratamento e por ser arbitrário na escolha do parâmetro da distância.

Os resultados são apresentados comparando os indicadores de Educação expressados em percentuais observados para cada grupo de comparação e cada procedimento de alocação. Neste caso, as medidas comparativas são diferenças entre os percentuais do grupo beneficiário do PBF e os não beneficiários de nenhum programa, os quais podem ser denominados "diferenciais do indicador".

Como algumas diferenças entre os resultados dos indicadores de impacto da educação entre os dois tipos de alocações utilizadas, serão observadas, alguns argumentos poderiam explicar estas diferenças, as quais podem ser colocadas apenas, como suposições. Mas, considera-se relevante comentá-los, tendo em vista a importância para o entendimento dos resultados do trabalho. Estes possíveis argumentos discutidos a seguir referem-se aos resultados dos indicadores de impacto, em função dos diferenciais entre grupos de comparações (grupo de beneficiários PBF e grupo de comparação 2):

1) A re-distribuição dos grupos observadas na seção 4.6.3, que é decorrente da alocação das famílias pelo CadÚnico, mostrou que uma porcentagem de famílias , que segundo alocação da pesquisa AIBF não pertenciam ao grupo de beneficiários do PBF, passaram a fazer parte deste grupo como a alocação do CadÚnico. Este resultado, por um lado, pode evidenciar diferenciais que não apareceram como significativos através da alocação das famílias pela pesquisa AIBF, mas sim com a nova alocação. Por outro lado, devido à redistribuição das famílias com a alocação com os dados do CadÚnico, alguns diferenciais podem também não ser significativos, pela variabilidade observada nos diferenciais de cada indicador. Em ambos dos casos, os resultados dependeram das características educativas das crianças pertencentes às famílias que foram alocadas em outros grupos de comparação segundo o CadÚnico.

2) Também se deve levar em conta, que o registro administrativo CadÚnico parece ser confiável em termos de grupo de comparação, pois essa variável é utilizada para definir quem recebe ou não o beneficio do PBF. Em contrapartida a variável renda deste registro administrativo CadÚnico podem estar apresentando dados menores, porque as famílias, para garantir o recebimento do beneficio, declaram ter menos renda, tal como é sugerido por Ramos e Santana, 2002. Desta forma para focar-se na sensibilidade originada ao comparar os registros administrativos e dados de pesquisa, decidiu-se utilizar, para esta análise, a renda declarada pelas famílias na pesquisa AIBF, no entanto, não pode ser ignorado o viés decorrente da declaração da renda, sobretudo no corte de renda de

R$200,00, corte que esta acima do limite máximo de elegibilidade oficial, e que foi utilizando apenas para garantir a representatividade amostral em todos os grupos, inclusive o de tratamento (OLIVEIRA et al, 2006).

Estes dois elementos apresentados são válidos para todos os casos, e o comportamento dependerá especificamente do indicador, região e corte de renda analisado e que será explicado em cada caso.

a) Evasão

Na TAB 5.6 apresentam-se o efeito do PBF sobre evasão escolar no último ano, os efeitos foram calculados para crianças masculinas e femininas individualmente e para cada região. Os diferenciais que são estatisticamente significativos são favoráveis ao programa, na medida em que são negativos, indicando uma menor evasão dos beneficiários PBF, em relação ao grupo de comparação 2. Esta menor evasão nos beneficiários do PBF conferem-se tanto nos grupos alocados segundo a pesquisa de campo AIBF, quanto nos grupos alocados segundo o paramento com o CadÚnico, embora, existem mais diferenças significativas no segundo tipo de alocação dos grupos.

Assim, no caso dos grupos alocados segundo a pesquisa de campo AIBF, as diferenças significativas são observadas nos domicílios com corte de renda domiciliar per capita até R$50,00 para as crianças do Brasil e especificamente nas crianças masculinas do Brasil, como nas crianças do Nordeste. Nos domicílios com corte de renda domiciliar per capita até R$200,00 as diferenças significativas foram para as crianças do Nordeste e as crianças femininas desta região. Para os grupos alocados segundo o CadÚnico, encontraram-se diferenciais significativos nas famílias com corte de renda domiciliar per capita até R$50,00 nas crianças como um todo do Brasil, Nordeste, além disso, crianças masculinas do Brasil, Nordeste e Sul/Sudeste; resultados observados também com corte de renda domiciliar per capita até R$100,00.

TABELA 5.6 – Diferenciais significativos entre os grupos de comparação "Tratamento e Comparação 2", sobre a proporção de crianças que evadiram a escola em 2004.

Corte de elegibilidade até	AIBF			CadÚnico		
	R$200,00	R$100,00	R$50,00	R$200,00	R$100,00	R$50,00
Brasil						
Total			-0,020***	-0,008*	-0,011***	-0,024**
Homens			-0,024**	-0,014**	-0,014**	-0,029**
Mulheres						
Nordeste						
Total	-0,016**		-0,023*	-0,021**	-0,030***	-0,050***
Homens				-0,022*	-0,051***	-0,067***
Mulheres	-0,020*					
Norte/C.Oeste						
Total						
Homens						
Mulheres						
Sudeste/Sul						
Total				-0,012*	-0,010*	
Homens				-0,009*	-0,010**	-0,036*
Mulheres						

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Segundo os resultados observados, os grupos de comparação alocados segundo o relacionamento com CadÚnico, apresenta maior quantidade de diferenciais estatisticamente significativa. Considerando os argumentos apresentados sobre as diferenças encontradas entre os resultados dos diferencias da evasão entre os grupos de comparação, pode-se supor que os diferenciais significativos da evasão encontradas com a alocação do paramento, mas, não com a alocação da pesquisa de campo AIBF, devem-se, a que as crianças das famílias que passaram a ser parte do grupo de tratamento com a alocação segundo o CadÚnico, são mais parecidos ao comportamento da evasão dos beneficiários do PBF, em tal sentido, neste caso as crianças que mudaram para o grupo de tratamento apresentaram menor evasão, e como conseqüência os diferenciais encontrados se incrementam e são significativos para a alocação segundo o relacionamento com o

CadÚnico. Este contexto confere-se entre as crianças totais e masculinas de todo Brasil, região Nordeste, Sudeste/Sul, e cortes de renda R$200,00 e R$100,00[42].

No caso das crianças femininas da região nordeste com corte de renda até R$200,00, em que os resultados dos diferenciais da evasão foram significativos apenas na alocação segundo a pesquisa AIBF, supõe-se que, este resultado está influenciado pela porcentagem das crianças que mudaram para o grupo de tratamento ou beneficiário do PBF e que evadiram da escola entre 2004 e 2005, como também pela porcentagem das crianças que mudaram para o grupo de comparação 2 (o caso inverso) e que evadiram da escola (no primeiro a porcentagem esta em torno de 3% e neste ultimo caso 1%) . Diante disto, o diferencial encontrado diminuirá e não será significativo para a alocação segundo o relacionamento com o CadÚnico. Como este resultado apresenta-se no corte de renda R$200,00, deve-se também considerar que, neste grupo encontram-se famílias com renda acima do limite máximo de elegibilidade oficial, e os resultados podem ser tomados com cuidado, porque, se espera que, neste caso as famílias estejam em melhores condições educacionais e portanto, os diferenciais podem ou não ser significativas.

Finalmente, os resultados segundo a alocação proveniente do paramento com os dados dos registros administrativos CadÚnico, confirmam alguns resultados da evasão, evidenciam outros resultados que não foram observados com a alocação dos grupos segundo a pesquisa de campo AIBF, mas também demonstra à sensibilidade dos resultados em relação à distribuição de cada alocação dos grupos de comparação utilizada, tal como se observa com o resultados das crianças femininas da região nordeste com corte de renda até R$200,00.

b) Aprovação

A análise de impacto da aprovação escolar das crianças de 7 a 14 anos de idade entre 2004 e 2005 são apresentadas na TAB 5.7. Estes resultados comparam a progressão do último ano, dos crianças masculinas e femininas no sistema escolar que freqüentaram. O impacto do programa procura encontrar diferenças positivas, considerando que existe uma maior aprovação das crianças provenientes de famílias beneficiarias do PBF. No entanto, no

[42] Assim, temos que o total das crianças masculinas de Brasil que mudaram ao grupo de comparação de tratamento, apenas 1% evadiram a escola entre 2004 a 2005.

primeiro momento, o fato das crianças beneficiárias do programa permanecerem mais no sistema escolar de um ano para o outro, pode levar a uma diminuição da aprovação ou progressão (OLIVEIRA et al, 2006).

Na TAB 5.7 as diferenças positivas sugerem um efeito potencial do PBF sobre a aprovação das crianças beneficiarias deste programa e são observados apenas para as crianças em total de homens da região nordeste com corte de renda até R$50,00 e para as mulheres da região sul/sudeste com corte de renda até R$200,00, resultados que são observados somente para os grupos de comparação alocados segundo o relacionamento com CadÚnico. Por outro lado, os diferenciais negativos que sugerem efeito inverso do PBF sobre a aprovação das crianças beneficiárias, são observados para as crianças de todo Brasil com corte de renda até R$200,00 e para as mulheres da região sul/sudeste com corte de renda até R$200,00, resultados encontrados com os grupos de comparação alocados segundo a pesquisa de campo AIBF. Da mesma forma diferenciais negativos conseguidas com a alocação segundo o relacionamento com o CadÚnico, são observados para as crianças da região Norte/Centro-Oeste com corte de renda até R$100,00 e mulheres da região Norte/Centro-Oeste com corte de renda até R$200,00 e R$100,00.

Comparando os resultados dos diferenciais significativos obtidos com os grupos de comparação da pesquisa de campo da AIBF e alocação segundo o paramento com o CadÚnico, observa-se que os resultados coincidem somente para as crianças em geral da região Norte/Centro-Oeste. Os demais diferenciais significativos obtêm-se resultados diferentes para ambas os tipos de alocações utilizadas.

TABELA 5.7 – Diferenciais significativos entre os grupos de comparação "Tratamento e Comparação 2", sobre a proporção de crianças que foram aprovados na escola entre 2004 e 2005.

Corte de elegibilidade até	AIBF			CadÚnico		
	R$200,00	R$100,00	R$50,00	R$200,00	R$100,00	R$50,00
Brasil						
Total	-0,020**					
Homens						
Mulheres						
Nordeste						
Total						0,108***
Homens						0,169***
Mulheres						
Norte/C.Oeste						
Total	-0,025*	-0,050**			-0,071***	
Homens	-0,043**		-0,133**			
Mulheres				-0,050**	-0,088***	
Sudeste/Sul						
Total			-0,070*			
Homens						
Mulheres				0,044*		

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Analisando os resultados que diferem nas alocações utilizadas, temos os diferenciais significativos da aprovação escolar encontradas com a alocação do relacionamento, mas, não com a alocação da pesquisa de campo AIBF, resultados que indicam que as crianças que foram alocadas no grupo de tratamento segundo o CadÚnico, apresentam maior aprovação escolar entre 2004 e 2005, determinando que o diferencial seja positivo e significativo para esses grupos de crianças. Nos resultados em que os diferenciais da aprovação escolar são significativos, apenas com a alocação segundo a pesquisa de campo AIBF, supõe-se que para os resultados com corte de renda até R$200,00, como as famílias apresentam renda acima do limite máximo de elegibilidade oficial, espera-se que estas estejam com melhores condições educacionais e portanto, os resultados dos diferenciais podem ou não ser significativos com a nova alocação, porque como as rendas das famílias do CadÚnico é menor então as maiorias das famílias que recebem o benefício estão abaixo dessa renda limite. No caso dos resultados com corte de renda até R$50,00, parecem ser que os resultados são influenciados pela porcentagem das crianças que mudaram decorrente da nova alocação, que passaram do grupo de comparação 2 para o tratamento,

crianças que parecem ser mais parecidos com o grupo de comparação 2, determinando uma diminuição do diferencial da aprovação das crianças entre os grupos de comparação e que resulta numa diferença não significativa. Assim, novamente observa-se a sensibilidade de utilizar uma nova alocação das famílias considerando os registros administrativos do CadÚnico, em comparação com a alocação segundo a pesquisa de campo AIBF.

c) Repetência

A repetência escolar é uma das manifestações perceptíveis da inadequação dos sistemas escolares contemporâneos às condições e possibilidades concretas e diferenciadas da população, e em particular dos alunos provenientes dos setores sociais menos favorecidos pelo desenvolvimento. (UNESCO, 1996).

Nos resultados das repetências, espera-se encontrar diferenças negativas, os quais sugerem potencial efeito positivo do programa, pela menor reprovação dos beneficiários do Bolsa Família. No entanto, da mesma forma que a progressão, o impacto não é tão óbvio nem imediato, pois a própria redução da evasão pode levar em um primeiro momento a uma maior repetência e diferenças positivas podem ser encontradas neste momento (OLIVEIRA et al, 2006).

Na TAB 5.8 apresentam-se os resultados para as crianças femininas e masculinas que foram reprovados entre 2004 e 2005. Comparando os resultados obtidos com a alocação dos grupos de comparação segundo a pesquisa de campo AIBF e alocados segundo o relacionamento com o CadÚnico, observa-se que, os resultados que são similares para ambas alocações encontra-se entre as crianças mulheres da região Nordeste com corte de renda até R$200,00, além disso, os resultados são similares também, na região Norte/Centro-Oeste, entre as crianças total com corte de renda até R$200,00 e R$100,00, crianças homens para os três cortes de renda consideradas e entre as crianças mulheres com corte de renda até R$100,00. Resultados diferentes entre ambos dos tipos de alocação são encontrados entre as crianças totais do Brasil para os três cortes de renda consideradas, na qual, apresentam diferenciais significativos somente para a alocação dos grupos de comparação segundo a pesquisa de Campo AIBF.

TABELA 5.8 – Diferenciais significativos entre os grupos de comparação "Tratamento e Comparação 2", sobre a proporção de crianças que repetiram a escola entre 2004 e 2005. Brasil e Regiões, 2005.

Corte de elegibilidade até	AIBF			CadÚnico		
	R$200,00	R$100,00	R$50,00	R$200,00	R$100,00	R$50,00
Brasil						
Total	0,020**	0,018*	0,040**			
Homens	0,034**					
Mulheres						
Nordeste						
Total						
Homens						
Mulheres	0,055*			0,052*		
Norte/C.Oeste						
Total	0,041*	0,045*	0,100*	0,036**	0,075***	
Homens	0,052**	0,061*	0,162***	0,051*	0,083**	0,081*
Mulheres		0,043**			0,069**	
Sudeste/Sul						
Total						
Homens						
Mulheres						

Fonte: AIBF, 2005 e CadÚnico 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Os resultados na TAB 5.8, mostram que todo os diferenciais significativos do Programa Bolsa Família são positivos, tanto na alocação dos grupos segundo a pesquisa de campo AIBF, como nos grupos obtidos pelo relacionamento com o CadÚnico. Estes resultados indicam uma maior reprovação dos beneficiários do Programa Bolsa Família em relação ao grupo de comparação 2, mas como mencionamos anteriormente deve-se ter cautela ao interpretar os resultados, por tratar-se de um indicador influenciado pela imediata redução da evasão e um acompanhamento e a avaliação em pontos subseqüentes no tempo podem mostrar evidências diferentes.

Comparando os resultados obtidos por ambos os tipos de alocações das famílias nos grupos de comparação, observa-se que os resultados diferem, para as crianças totais e homens de todo Brasil, em que os diferenciais da repetência escolar é significativos, somente, para a alocação das famílias segundo a pesquisa de campo AIBF. Ao analisar estas diferenças, devem-se levar em conta novamente as características das crianças que mudaram de grupo de comparação. No caso da repetência escolar, considera-se que, crianças pertencentes ao

grupo de comparação 2 e com a nova alocação passaram a ser parte do grupo de tratamento, em sua maioria não repetiu. Mas, por outro lado, aqueles que mudaram de forma inversa nos grupos de comparações (de grupo de tratamento pra comparação 2), uma porcentagem significativa repetiu, mas não foi maior que a proporção de repetência do grupo de tratamento[43]. Estes dois eventos determinaram que os diferenciais da repetência escolar, com a nova alocação das famílias, diminuíssem e, por conseguinte não fossem significativas. Como antes explicamos, estes resultados são melhores explicados no corte de renda até R$100,00 e R$50,00.

Os resultados para a repetência escolar continuam advertindo a sensibilidade da informação que se utiliza para alocação das famílias, além disso, as famílias com corte de renda acima de R$200,00, mantém um comportamento distinguível em relação às significância dos diferenciais dos indicadores de impacto. Finalmente, os resultados utilizando a alocação dos registros administrativos do CadÚnico, não variam, mas se confirmam os resultados obtidos com os grupos alocados com os dados da pesquisa de campo do AIBF, exceto para o total de Brasil, produto da sensibilidade da alocação e do corte de renda.

d) Deixou de ir à escola

A freqüência das crianças aos cursos básicos, concede aos alunos uma perspectiva de atingir, os conhecimentos para desenvolver-se na sociedade, os quais são importantes, porque, através destes têm possibilidades de abrir espaços a outros niveles educativos ou sociais. Os pais têm um papel importante para a freqüência de seus filhos à escola, no entanto, filhos de famílias pobres, muitas vezes deixam de freqüentar à escola, devido a restrições no investimento escolar (SCHUTZ, 2000). Além disso, é importante incentivar a freqüência escolar, porque esta contribui para a diminuição do abandono e a evasão escolar, por parte das crianças.

Os resultados da proporção de crianças femininas e masculinas no domicílio que deixaram de ir à escola no último mês são apresentados na TAB 5.9. Os resultados esperados para este indicador são diferenças negativas, que indicam uma maior freqüência dos

[43] No caso das crianças masculinas do Brasil com corte de renda até R$100, 00, os incrementos observados para o grupo do PBF e de comparação 2 foram, -6% e 8% respectivamente.

beneficiários do PBF em relação aos não beneficiários de nenhum programa social. No entanto utilizando a alocação dos grupos de comparação segundo o relacionamento com o CadÚnico, algumas diferenças positivas significativas são encontradas entre os homens e mulheres da região Norte/Centro-Oeste com corte de renda até R$100,00, o qual indica diferenças favoráveis aos não-beneficiários, isto é, as crianças provenientes de famílias que não recebem benefício do PBF, apresentam um efeito mais consistente em relação aos beneficiários.

TABELA 5.9 – Diferenciais significativos entre os grupos de comparação "Tratamento e Comparação 2", sobre a proporção de crianças que deixaram de ir à escola no último mês. Brasil e Regiões, 2005.

Corte de elegibilidade até	AIBF			CadÚnico		
	R$200,00	R$100,00	R$50,00	R$200,00	R$100,00	R$50,00
Brasil						
Total	-0,032***			-0,018**		
Homens						
Mulheres	-0,052***	-0,036***		-0,015***		
Nordeste						
Total						
Homens						
Mulheres	-0,065**	-0,031*				
Norte/C.Oeste						
Total					0,018**	
Homens					0,033**	
Mulheres						
Sudeste/Sul						
Total	-0,061***			-0,047**		
Homens						
Mulheres	-0,093***	-0,026***		-0,022**	-0,050**	

Fonte: AIBF, 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Embora tenham se observado alguns diferenciais positivos significativos neste indicador, à maioria dos diferenciais significativos é negativo, resultados que indica uma maior freqüência dos beneficiários do Programa Bolsa Família em relação ao grupo de não-beneficiários. Estes resultados trabalhando com os grupos de comparação obtidos segundo a pesquisa de campo AIBF mostram que, existe uma maior freqüência dos beneficiários, em relação ao grupo de não-beneficiários entre: as crianças como um todo com corte de

renda até R$200,00 e crianças mulheres com corte de renda de até R$200,00 e R$100,00 para Brasil; mulheres com corte de renda de até R$200,00 e R$100,00 para a região Nordeste; e total de crianças com corte de renda até R$200,00 e crianças mulheres com corte de renda de até R$200,00 e R$100,00 para a região Sudeste/Sul. Considerando os resultados obtidos com famílias alocadas nos grupos de comparação segundo o relacionamento com o CadÚnico, observa-se diferenciais positivos, entre as crianças como um todo e crianças mulheres com corte de renda até R$200,00 para Brasil, e as crianças como um todo com corte de renda até R$200,00 e crianças mulheres com corte de renda de até R$200,00 e R$100,00 para a região Sudeste/Sul.

Comparando os resultados entre ambos os tipos de alocação, apresentam-se diferenças apenas entre os resultados das mulheres da região Nordeste, e entre as crianças da região Norte/Centro-Oeste. Considerando o resultado, em que o diferencial do indicador é significativo, apenas para a alocação das famílias segundo o relacionamento com o CadÚnico, pode-se encontrar a explicação deste desempenho entra as criança que na nova alocação mudaram de grupo de comparação, por um lado, as crianças que mudaram de grupo de comparação 2 para o grupo de tratamento, apresentaram maior proporção de crianças que deixaram de ir à escola no último mês, e por outro lado, o grupo de comparação 2 com a nova alocação, ficaram em menor proporção de crianças que deixou de ir à escola no último mês, portanto, ambas as mudanças, ocasionou que, o comportamento das crianças do grupo de tratamento é mais diferente do que o grupo de comparação 2, em relação a este indicador, e que determina diferenciais positivos e significativos para este grupo de crianças.

Uma explicação disto pode ser encontrada a partir das crianças que mudaram de grupo com a nova alocação, mostrando que, entre as crianças antes pertencentes ao grupo de comparação 2, e que passaram a ser parte do tratamento, existem mais crianças que deixaram de ir à escola, e em contrapartida, no grupo de comparação 2, com a nova alocação, a proporção de crianças que deixaram ir à escola foi menor. Assim, o comportamento das crianças do grupo de tratamento é similar ao comportamento do grupo de comparação 2, em relação a este indicador. Disto resulta que os diferenciais entre os grupos de comparação com a nova alocação diminuíram e por tanto este não seja significativo.

Se bem que os resultados, utilizando a alocação segundo os registros administrativos do CadÚnico, não indicaram substancial diferença entre os indicadores de impacto para as crianças que deixaram ir à escola no ultimo mês, confirmam a sensibilidade dos resultados, influenciados pelas alocação das famílias segundo os dados da pesquisa de campo do AIBF e os registros administrativos do CadÚnico.

e) Alocação de trabalha e estuda

Uns dos motivos principais para que as crianças não freqüentarem a escola ou creche é o fato de que estas estão trabalhando ou tomando providência para trabalhar com o objetivo de conseguir dinheiro para as despesas familiares. Assim, muitos adolescentes, forçados por necessidades econômicas impostergáveis de seus grupos familiares, procuram ingressar prematuramente no mercado de trabalho com competências mínimas, sem ter adquirido as habilidades essenciais requeridas pelos postos de trabalho, com insuficientes e frágeis redes de relações sociais (Schutz, 2000). Desta forma os programas sociais, assim como o PBF, visam promover o abandono das crianças ao trabalho infantil para voltar ou manter-se na escola.

Na TAB 5.10 apresentam-se a proporção de crianças masculinas e femininas no domicílio que declararam só estudar atualmente, vis-à-vis aqueles que declararam só trabalhar, trabalhar e estudar e não trabalhar nem estudar. Estes resultados mostram apenas um diferencial negativo significativo entre as crianças femininas de 7 a 14 anos de idade da região Sudeste/Sul com corte de renda até R$100,00 e nos grupos de comparação alocados segundo a pesquisa de campo AIBF. Estes diferenciais não implicam necessariamente, uma menor freqüência à escola, mas pode ser reflexo da conciliação entre trabalho e estudo (OLIVEIRA et al, 2006).

Entre os outros resultados significativos diferenciais positivos são encontrados, o que indicam uma maior alocação do tempo para o estudo às crianças provenientes de famílias beneficiárias do PBF, em comparação ao grupo não–beneficiário, resultados observados em ambas as alocações dos grupos de comparação utilizados. Assim, os diferenciais positivos significativos encontram-se entre o total das crianças com corte de renda até R$50,00 e crianças mulheres com corte de renda de até R$50,00 para Brasil; e total de crianças para os 3 cortes renda, crianças homem com corte de renda R$100,00 e mulheres com corte de renda de até R$50,00 para a região Norte/Centro-Oeste, estes resultados

confirmam-se para ambas as alocações de grupos de comparação utilizados Os diferenciais positivos encontrados, sugerem uma diferença favorável aos beneficiários do Programa Bolsa Família, como conseqüência de que os benefícios do Programa Bolsa Família estejam, associados a famílias com filhos, que devem freqüentar a escola, implicando que o valor do tempo dos filhos no trabalho devam reduzir, e conseqüentemente sua participação na força de trabalho tende a diminuir (OLIVEIRA et al, 2006).

TABELA 5.10 – Diferenciais significativos entre os grupos de comparação "Tratamento e Comparação 2", sobre a proporção de crianças que são estudavam em 2005.

Corte de elegibilidade até	AIBF			CadÚnico		
	R$200,00	R$100,00	R$50,00	R$200,00	R$100,00	R$50,00
Brasil						
Total			0,022**			0,011**
Homens						
Mulheres			0,029**			0,023**
Nordeste						
Total						
Homens						
Mulheres						
Norte/C.Oeste						
Total	0,022*	0,022**	0,058**	0,020*	0,034**	0,073***
Homens		0,018**		0,034**	0,057**	0,082**
Mulheres			0,087**			0,091**
Sudeste/Sul						
Total						
Homens						
Mulheres		-0,015*				

Fonte: AIBF, 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Comparando os resultados obtidos pelos dois métodos de alocação de grupos de comparação, observa-se que a diferença mais ressaltante, está entre as crianças mulheres da região Sul/Sudeste com corte de renda até R$100,00, em que o diferencial além de ser negativo é significativo, mas unicamente com a alocação das famílias segundo a pesquisa de campo AIBF. Este resultado indicaria que nesta primeira a alocação do comportamento em relação às crianças que não estudavam é mais diferente entre os grupos de comparação, mas com proporção maior no grupo de comparação 2; no entanto, com a nova alocação eles tornaram-se mais similares, resultado do comportamento das crianças que mudaram do

grupo de comparação 2 ao tratamento, a qual mostrou uma proporção de crianças que só estudavam, incrementando a proporção de crianças dedicadas ao estudo no grupo de tratamento, mas não de forma tal que este seja maior que no grupo de comparação 2, grupo o qual apresentou menor proporção de crianças dedicadas a estudar. Estas mudanças determinaram que o diferencial entre os grupos de comparação, diminuísse, por conseguinte, deixa de ser significativo.

Desta forma os resultados utilizando o CadÚnico para alocar as famílias nos grupos, confirmam os resultado encontrados com a alocação das famílias com dados da pesquisa de campo do AIBF, mas também com esta última alocação, não foi possível ressaltar, a diferença observada entre as crianças mulheres da região Sul/Sudeste com corte de renda até R$100.

5.5.2 Resultados da aplicação da Regressão Descontínua (RD).

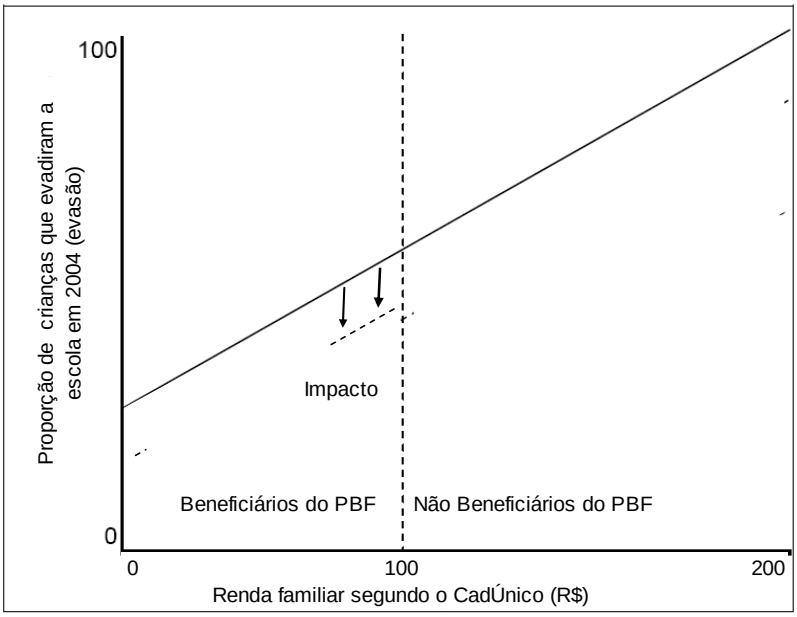
a) Generalidades da aplicação do modelo.

Uma alternativa para estimar o impacto do programa Bolsa Família sobre a educação das crianças de 7 a 14 anos é utilizar a técnica de Regressão Descontinua (RD). A aplicação da avaliação de programas sociais e políticas públicas utilizando este método consistem, de forma geral, na relação existente entre a variável que determina o tratamento e outras variáveis que indicam os impactos do mesmo.

O desenho da RD que será aplicada neste trabalho supõe que em principio, existem uma relação continua ou "suave" entre a renda das famílias do registro administrativo do Cadastro Único para Programas Sociais (CadÚnico) e a variável de impacto, isto é, indicadores de impacto para avaliar os diferenciais do PBF na educação das crianças entre 7 e 14 anos. No entanto, como para fazer parte do PBF, às famílias precisavam ter renda mensal de até R$100,00 (cem reais) por pessoa devidamente cadastrada no CadÚnico, esta renda classifica às famílias que recebem o beneficio do PBF e os que não recebem. Neste sentido, existe um ponto "definido" que separa estes dois tipos de famílias, e que pode ser considerada a renda mensal de até R$100,00. Com base nesta idéia, espera-se que a relação "suave" da variável renda mensal familiar apresente uma descontinuidade no ponto corte ou separação (R$100,00). Esta descontinuidade será explicada pelo fato de que as famílias que receberam os benefícios do PBF teriam melhores indicadores da educação, supondo

que os benefícios do programa tiveram o impacto esperado. Assim, por exemplo, um impacto positivo do PBF sobre a evasão escolar, mostrar-se-ia intuitivamente como um deslocamento até abaixo da linha que indica a relação entre ambas as variáveis, precisamente antes do ponto de corte que separa às famílias como beneficiárias ou não do PBF, tal como se mostra na seguinte figura:

FIGURA 5.1 – Esquema da descontinuidade da renda familiar do CadÚnico, em relação ao impacto da proporção das crianças que evadiram a escola em 2004. Brasil. 2005.



Na realização da análise da descontinuidade deve-se considerar às famílias que se encontram na vizinhança do umbral ou ponto de corte, surgindo o problema de como definir essa vizinhança. Quanto a vizinhança define-se de forma muito ampla ─ por exemplo, considerando praticamente a todas as famílias que são consideradas no estudo ─ então as estimações ganham em termos de poder estatístico, mas perdem no sentido de que os grupos em cada lado contém famílias mais heterogêneas e, por conseguinte mais difícil de comparar entre si. Quando a vizinhança define-se de forma estreita, então ocorre exatamente o contrário. Para este trabalho, o tamanho da vizinhança, definiu-se de tal forma que se obtenha uma amostra suficiente para ter poder estatístico nas estimações. No entanto, com o objetivo de verificar se os resultados são sensíveis ao tamanho da vizinhança selecionada definiu-se mais uma vizinhança, mas apenas como teste.

O método de estimação de RD que é utilizado neste trabalho é conhecido como estimadores não-paramétricos e dependem da escolha da função de *Kernel* e da *bandwith*. Em nosso estudo escolheu-se para reportar os estimadores RD obtidos com um *bandwith* de 50. A função de *Kernel* utilizada é a uniforme (ou retangular), que assina igual peso para todas as observações que caem dentro da banda de +/-50 pontos discriminantes a longo da região especificada pelo valor de corte da descontinuidade e peso zero para as observações fora da banda (isto é, menos escores ou mais que 50 pontos longe da região especificada pelo ponto de corte)

Considerando estas características utilizamos uma notação formal para modelar os indicadores de impacto (variável resultado) do Programa Bolsa Família sobre a educação nas crianças de 7 a 14 anos, através da seguinte equação:

$$Y_{ir} = \beta_0 + \beta_1 TRAT_{ir} + \delta(r) + \sum_{j=1}^{J} \theta_j X_{ij} + \varepsilon_{ir} \qquad (5.2)$$

em que, Y_{ir} é a variável resultado para a criança i com renda familiar r. O efeito da renda familiar segundo CadÚnico sobre a variável é capturada pela função $\delta(r)$, enquanto $TRAT_{ir}$ é uma variável *dummy* que indica se a criança provêem de uma família beneficiaria do PBF, que neste caso é expressada, através da renda familiar segundo o CadÚnico até os R$100,00 reais, e que pode ser definida como:

$$TRAT_{ir} = \begin{cases} 0 \; se \; r \leq 100{,}00 \\ 1 \; se \; r > 100{,}00 \end{cases},$$

além disso, na equação também encontramos X_{ij} que representa o vetor de *J-th* variáveis de controle ou de equilíbrios, que consiste em variáveis individuais familiares e locais (ver TAB. 5.2). Um último termo é encontrado na equação, o ε_{ir}, que representa o resumo das influencias dos erros aleatórios.

Da equação 5.2, o coeficiente importante é o parâmetro β_1—relacionado à variável dicotômica que indica se o PBF influi ou não—isto é, se o PBF foi eficaz para melhorar os indicadores da educação das crianças de 7 a 14 anos. Desta forma, quando o coeficiente é negativo (ou positivo) e estatisticamente significativo então se pode falar que o PBF foi eficaz na educação das crianças de 7 a 14 anos das famílias beneficiarias.

Quando se realiza a aplicação do modelo *RD*, a literatura sobre estes modelos distingue dois tipos de desenho de *RD*: o chamado *Sharp* e o *Fuzzy*[44]. Para este trabalho, o desenho S*harp* será utilizado, uma vez que, o tratamento *TRAT* é conhecido e supõe-se que depende de uma forma determinística de $\delta(r)$.

Uma observação que deve ser realizada antes de apresentar os resultados está relacionado ao ponto de corte, isto é, sabe-se que o Programa Bolsa Família em 2006 apresentou dois tipos de benefícios: o básico e variável. O benefício básico, de R$ 50,00 (cinqüenta reais), era pago às famílias consideradas extremamente pobres, aquelas com renda mensal de até R$ 50,00 (cinqüenta reais) por pessoa (pago às famílias mesmo que elas não tenham crianças, adolescentes ou jovens). Por sua parte, o benefício variável, de R$ 20,00 (vinte reais), era pago às famílias pobres, aquelas com renda mensal de até R$ 100,00 (cem reais) por pessoa desde que tenham crianças e adolescentes de até 15 anos (MDS, 2006). Disto, destaca-se que para questões de análises, as estimações do modelo de RD serão realizadas para dois pontos de cortes ou umbral: renda familiar até R$50,00 e até R$100,00.

b) Resultado e discussão da estimação do modelo.

Agora Na TAB 5.11 apresentam os resultados da estimação do modelo de RD para os diferenciais na educação das crianças de 7 a 14 anos do PBF, reportando-se, apenas, os coeficientes da variável que indica um diferencial estatisticamente significativo.

Considerando a descontinuidade no ponto de corte até R$100,00, temos que, para as crianças masculinas de todo Brasil e crianças totais da região Nordeste que evadiram da escola em 2004, têm, diferenciais significativos, e são favoráveis às famílias com renda abaixo de R$100,00, na medida em que são negativos. Da mesma forma, considerando a descontinuidade em R$50,00, a evasão das crianças masculinas de todo Brasil e total de crianças da região Norte/Centro-Oeste, que evadiram da escola em 2004, apresenta diferenciais significativas e favoráveis às famílias com renda abaixo de R$50,00 reais, porque os diferenciais são negativos. Diante estes resultados, é possível supor que, existe uma diferença favorável aos beneficiários do Programa Bolsa Família em relação às

[44] Por exemplo, pode revisar o Capitulo 3 desta tese ou Van der Klaauw (2002) and Hahn et al (2001) para uma discussão detalhada do desenho *Sharp* e *Fuzzy*.

crianças pertencentes de domicílios que não participam de nenhum programa, resultado que favorece aos objetivos do programa nessas regiões e grupos de crianças.

TABELA 5.11 – Estimação da regressão descontínua dos indicadores para avaliar os diferenciais do PBF na educação de crianças de 7 a 14 anos. Brasil e Regiões, 2005.

Variáveis/Regiões	Ponto de corte ou umbral até					
	R$100,00			R$50,00		
	Total	Homens	Mulheres	Total	Homens	Mulheres
a) Crianças que evadiram a escola em 2004 (evasão)						
Brasil		-0,015**			-0,017*	
Nordeste	-0,026*					
Norte/Centro-Oeste				-0,023*		
b) Crianças que foram aprovados a escola entre 2004 e 2005						
Brasil						
Nordeste					0,283*	
Norte/Centro-Oeste						
c) Crianças que repetiram a escola entre 2004 e 2005 (repetência)						
Brasil						
Nordeste	-0,097*				-0,290*	
Norte/Centro-Oeste						
d) Crianças que são estudavam em 2005						
Brasil				-0,218***		
Nordeste		-0,134**				
Norte/Centro-Oeste						

Fonte: AIBF, 2005.
Nota: * valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.
O grupo de Tratamento é constituído pelos domicílios que recebem atualmente o benefício do Bolsa Família. O grupo de Comparação 2 é composto pelos domicílios que declararam nunca ter recebido nenhum tipo de benefício, independentemente de serem cadastrados em algum programa público.

Em relação aos resultados em termos da proporção de aprovados entre 2004 e 2005, diferenciais significativos do PBF possuem diferença positiva, para a descontinuidade em R$50,00 e nas crianças masculinas da região nordeste, este resultado indica uma maior aprovação das famílias com renda abaixo de R$50,00 reais. Disto, supõe-se que, como as famílias que recebem o benefício do PBF são aqueles abaixo deste corte de renda, então se sugere um potencial efeito positivo para os beneficiários deste programa, em relação ao grupo de não-beneficiários.

Para a variável de repetência escolar entre 2004 e 2005, encontraram-se diferenciais significativos e negativos, para o corte de descontinuidade de R$100,00 entre o total de

crianças na região Nordeste e para o umbral de descontinuidade de R$50,00 entre as crianças masculinas na região Nordeste. Estes resultados poderiam ser interpretados como favoráveis às famílias com renda abaixo desses cortes de renda especificados, famílias que possivelmente recebem os benefícios do PBF e, portanto, supõe-se que há uma diferença favorável aos beneficiários do PBF em relação às crianças em domicílios que não participam de nenhum programa.

Considerando a proporção de crianças que trabalham vis-à-vis aqueles que só estudam, ou não trabalham nem estudam, diferenciais significativos e negativos são encontrados, para a descontinuidade em R$100,00 entre meninos da região Nordeste, e para a descontinuidade de R$50,00 entre crianças de todo Brasil. Estes resultados indicam uma maior participação na força de trabalho entre as crianças com renda familiar abaixo dos cortes de renda e regiões consideradas, em comparação ao grupo de famílias não-beneficiárias. Resultados diferentes poderiam ser esperados com este indicador, porque famílias abaixo desses cortes de rendas, supõem-se que recebem o beneficio do PBF, mas considerando que, é possível existir ainda, uma maior participação na força de trabalho independente da freqüência à escola por parte das crianças, o qual poder ser o reflexo da conciliação entre trabalho e estudo que ainda, não tem conseguido ser diminuída ou eliminada, mas para futuras medições espera-se resultados diferentes (OLIVEIRA et al, 2006).

Finalmente, deve-se destacar que a variável de não deixar de ir à escola no ultimo mês, não foi mostrada na tabela, porque nenhum diferencial foi significativo. Além disso, a presença de poucos diferenciais significativos para todos os indicadores da educação e regiões estudadas, pode ser interpretado como resultado da configuração do modelo de RD. O modelo de RD considera que, as famílias no entorno reduzido da vizinhança do umbral ou ponto de corte é descontínua em relação a uma variável exógena aos resultados potenciais do impacto, que para nosso caso é a renda familiar dos registros administrativos do CadÚnico (corte de renda de R$100,00 e R$50,00). Deste modo, as famílias que estão nos extremos ou com renda distante dos pontos de corte, não serão explicitamente representadas, famílias que em sua maioria estão na extrema pobreza, e para as quais se supõem que os benefícios do PBF atingem em melhor medida.

6 CONSIDERAÇÕES FINAIS

Este trabalho investigativo explorar as possibilidades únicas que são abertas pelo relacionamento de bases de dados para analisar a sensibilidade dos resultados de impacto dos programas sociais de transferência de renda, quando se utiliza dois tipos de fontes de informação para a alocação das famílias nos grupos de tratamento e comparação 2. Para tal análise, realizou-se a aplicação específica da avaliação de impacto do Programa Bolsa Família nos indicadores da educação para as crianças de 7 e 14 anos, utilizando a alocação das famílias nos grupos de comparação, segundo a pesquisa de campo AIBF e os registros administrativos CadÚnico.

Para utilizar duas fontes de informação que permita alocar as famílias e proceder à análise comparativa dos resultados de impacto, foi preciso realizar o relacionamento de bases de dados ou *record linkage*, das bases obtidas da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. Nesse sentido, precisou-se, primeiro recuperar informação do Número de Identificação Social (NIS) dos integrantes das famílias entrevistadas na pesquisa de campo do AIBF, e seguidamente, re-alocar estas famílias com a variável que define os benefícios que recebiam no mês da pesquisa de campo segundo o CadÚnico.

Para avaliar os efeitos do PBF sobre os indicadores da educação das crianças de 7 a 14 anos, conforme os dois tipos de alocações das famílias utilizadas adotaram-se as técnicas econométricas Pareamento por Escore de Propensão (PSM) e Regressão Descontínua (RD). A primeira técnica consiste em atribuir mais peso na análise para quem tem mais probabilidade de ser selecionado para o PBF. A segunda técnica consiste em comparar as famílias que estão no limite de elegibilidade do Programa. Dada às restrições dos dados, a utilização destas técnicas parecem ser as metodologias mais indicadas, sendo que o primeiro método permite parear os indivíduos do grupo de tratamento e controle através das características observáveis, reduzindo assim o viés de seleção, enquanto o segundo método é uma aplicação como resultado direto do relacionamento de bases de dados que supõe em princípio que existe uma relação continua ou “suave” entre a renda das famílias do CadÚnico e a variável de impacto, isto é, indicadores de impacto para avaliar os diferenciais do PBF na educação das crianças entre 7 e 14 anos.

De acordo aos resultados obtidos ressalta-se que a metodologia de relacionamento de bases de dados é de relevante importância para a aplicação de outras técnicas não-experimentais, úteis para avaliar os resultados de impacto de programs sociais, possibilitando um olhar integrado sobre as informações disponíveis em várias fontes de informações e permitindo uma análise comparativa. Isto é pertinente, porque diversas áreas aplicaram o relacionamento de base de dados, como ferramenta para melhorar a quantidade e qualidade das informações necessárias para uma pesquisa (GILL, 2001).

Considerando os objetivos de construir uma base de dados com informações combinadas da base da pesquisa AIBF e CadÚnico, os resultados realmente permitiram conhecer informação adicional das famílias entrevistadas na pesquisa de campo AIBF, comparar com informação do CadÚnico e aplicar a RD, exercícios que não seriam viáveis usando apenas uma única fonte de informação. Diante da necessidade de incrementar informação aos dados obtidos da pesquisa de campo, os resultados do relacionamento mostraram-se bastante representativos e precisos, sobretudo considerando a apropriada porcentagem das famílias beneficiárias do PBF entrevistadas na pesquisa de campo AIBF e que foram encontrados no CadÚnico, decorrência importante, porque estas famílias têm maior probabilidade de estar registradas neste cadastro. Assim, o número de famílias que foi possível encontrar ou parear com o processo de relacionamento de bases de dados pode ser considerado satisfatório para analisar as presumíveis variações ou sensibilidades dos resultados de impacto do PBF, quando se utilizam registros administrativo para alocar às famílias nos grupo de tratamento e comparação 2. Há que se considerar também, que estes resultados são representativos, já que na amostra AIBF existem famílias cadastradas no CadÚnico que ainda não são beneficiárias do PBF (domicílios podem ser beneficiários de outros programas de transferência de renda), bem como famílias não cadastradas ou beneficiárias (OLIVEIRA et al, 2007).

De acordo com o objetivo, de análisar comparativamente os resultados dos indicadores de impacto da educação entre os dois tipos de alocações utilizadas, os resultados sugerem que os argumentos que poderiam explicar as diferenças encontradas podem ser colocadas apenas como suposições, mas com caráter relevante. Assim, a re-distribuição dos grupos decorrente da alocação das famílias pelo CadÚnico mostrou que um porcentagem de

famílias[45], que pela alocação da pesquisa AIBF não pertenciam ao grupo de beneficiários do PBF, passaram a ser parte desse grupo com a alocação do CadÚnico. Além disso, o CadÚnico parece ser confiável em termos de grupo de comparação, pois essa variável é utilizada para definir quem recebe o beneficio do PBF. Em contrapartida a variável renda do CadÚnico pode estar apresentando valores subestimados, já que as famílias, para garantir o recebimento do beneficio, declaram ter menos renda (RAMOS e SANTANA, 2002). Para minimizar tal problema e focar-se na sensibilidade originada ao comparar os registros administrativo e dados de pesquisa, utilizou-se para a aplicação do PSM a renda declarada pelas famílias na pesquisa AIBF. Ressalta-se, pois, que não pode ser ignorado o viés decorrente da declaração da renda, sobretudo no corte de renda de R$200,00 - corte que esta acima do limite máximo de elegibilidade oficial e que foi utilizando apenas para garantir a representatividade amostral em todos os grupos, inclusive o de tratamento (OLIVEIRA et al, 2006).

Os resultados da análise comparativa apresentados no capitulo 5 evidenciam diferenciais que não são relevantes considerando-se a alocação das famílias pela pesquisa de campo AIBF, mas que se tornam significativos com a alocação decorrentes do CadÚnico, sobretudo na proporção de crianças que evadiram a escola em 2004, demonstrando a sensibilidade dos resultados em relação à distribuição de cada alocação dos grupos de comparação. Nesse caso, verificou-se que as crianças de famílias que passaram a ser parte do grupo de tratamento com a alocação segundo o CadÚnico, são mais parecidos quanto ao comportamento de evasão dos beneficiários do PBF, ou seja, com a aplicação do CadÚnico as crianças que mudaram de grupo de tratamento apresentaram menor evasão, e conseqüentemente os diferenciais encontrados se incrementaram e foram significativos. Por outro lado, existem situações em que a significância não se verifica, isto é, diferenciais que são expressivos através da alocação das famílias pela pesquisa de campo AIBF, não agregam novos dados do CadÚnico, este é o caso da retenção escolar: quando as crianças pertencentes ao grupo de comparação 2 foram re-alocados confirmaram a situação de reprovação.

Por outro lado, aqueles que mudaram de forma inversa nos grupos de comparações (de grupo de tratamento pra comparação 2), uma porcentagem significativa repetiu, mas não

[45] O 16% das famílias que pertenciam a outros grupos de comparação passaram a ser parte do grupo de tratamento ou dos beneficiários do PBF, segundo a alocação com o CadÚnico.

foi maior que a proporção de repetência do grupo de tratamento, determinando que as diferenciais da retenção escolar para este grupo de alunos com a alocação segundo CadÚnico, diminuíssem e, por conseguinte não fossem significativas. Os resultados mencionados a maneira de exemplo, advertem a sensibilidade da informação que se utiliza segundo a alocação das famílias. Além disso, para o caso das famílias com corte de renda acima de R$200,00 espera-se um comportamento distinguível mantido em relação à significância dos diferenciais dos indicadores de impacto, isto é, supõe-se que os resultados para corte de renda até R$200,00, sendo renda acima do limite máximo de elegibilidade oficial, encontrem melhores condições educacionais e portanto, os resultados dos diferenciais podem ou não ser significativas com a nova alocação, pois, uma vez que a renda das famílias registradas no CadÚnico é menor, a maioria das famílias que recebem o beneficio estão abaixo dessa renda limite.

Recorrendo a uma forma particular de identificar os grupos potencialmente beneficiários e não-beneficiários do PBF, através da aplicação do método de Regressão Descontinua *Sharp* (RD), verifica-se a presença de poucos resultados significativos para os indicadores da educação e regiões estudadas. É possível que tais resultados tenham sido influenciados pela configuração do modelo, que considera apenas as famílias que estão no entorno de uma vizinhança reduzida do umbral ou no ponto de corte da descontinuidade. Isto é, as famílias que estão nos extremos ou com renda distante dos pontos de corte, não serão explicitamente representadas, famílias que em sua maioria estão em situação de extrema pobreza e para as quais se supõe que os benefícios do PBF atingem em melhor medida. No entanto, os resultados expressivos que foram encontrados com a RD confirmaram alguns resultados encontrados com a aplicação do PSM.

Avaliando a sensibilidade dos resultados de impacto da educação, observa-se que, utilizando as duas fontes de informação os resultados apresentam algumas alterações, sugerindo a existências de diferenças quando se utilizam diferentes fontes de dados na alocação das famílias nos grupos de comparação. No entanto, há indícios que os resultados encontrados na análise comparativa dependam das características próprias educativas das crianças pertencentes às famílias que foram alocadas em outros grupos de comparação. Esses sinais evidenciam-se, em maior medida, quando os resultados para ambas as fontes de informações mostram-se compatíveis. Assim os resultados dos indicadores de impacto da educação utilizando a alocação dos registros administrativos do CadÚnico não

invalidariam as conclusões sobre o impacto do PBF na educação com os dados da pesquisa AIBF; confirmando-se estes resultados, mas também evidenciando que a aplicação dos métodos não-experimentais, utilizando a alocação segundo os dados do CadÚnico, incrementam a robustez dos resultados e portanto a validação destes.

Embora os resultados de impacto neste trabalho esteja mais relacionado à comparação dos diferenciais de impacto das duas fontes de informação para alocar as famílias, a relevância das avaliações de impacto é direta, pois os efeitos indicam que os resultados podem ser associados ao PBF ou às melhorias em programas existentes para o atingir os objetivos da política social.

Certamente a análise comparativa proposta neste trabalho leva a uma reflexão sobre as fontes de informação, a metodologias de avaliação e a importância que estas têm na execução da avaliação das políticas públicas mais eficientes. Pontua-se aqui que outras avaliações de programas com formatos bastante similares ao PBF, como o Progresa no México (atualmente Oportunidades) e Familias en Acción na Colômbia, já utilizaram para a sua avaliação registros administrativos e dados de pesquisas de campo, como também diferentes (ou combinação) técnicas metodológicas para focalizar e avaliar. O objetivo da avaliação foi encontrar resultados mais robustos, porque à medida que eles permanecem inalteráveis expressivamente, ainda que sob a utilização de diferentes técnicas e fontes de informação, pode-se assegurar sua validez e eficiência dos resultados.

A sugestão de uma agenda de pesquisa imersa na análise de avaliação de impacto está baseada na utilização ou combinação de métodos e fontes de informações disponíveis, não apenas na avaliação de impacto, mas talvez na implementação de programas de transferências condicionadas à renda. No Brasil grandes bases de dados de produção de serviços e de abrangência nacional, como também, pesquisas nacionais baseadas na coleta de dados primários com objetivos específicos, podem ser integradas com o objetivo de contribuir para a melhoria da qualidade dos dados registrados, do seguimento longitudinal e da ampliação do escopo de perguntas a serem respondidas. Assim, a integração de bases de dados de naturezas diversas permitirá aperfeiçoar o planejamento, análise, avaliação e posterior implementação de políticas públicas que permitam o desenvolvimento da plena convivência social, política e econômica dos diversos atores que participam na formação de um Estado.

REFERÊNCIAS BIBLIOGRÁFICAS

ABADIE, A. Semiparametric difference-in-differences estimators. Review of Economic Studies, Cambridge, v. 72, n. 1, p 1–19, Jan. 2005.

ADATO, M. E.; ROOPNARAINE, T. Sistema de evaluación de la red de protección social de Nicaragua: un análisis social de la "Red de Protección Social" (RPS) en Nicaragua. Washington, DC: International Food Policy Research Institute. 2004. Informe final.

ALFONSO, J. La importancia social de la información. Journal of the National Center of Information on Medical Sciences, La Habana, v. 9, n. 3, p. 221-223, sep./dic. 2001.

ALMEIDA, M. F.; JORGE, M. H. de M. O uso da técnica de "Linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. Revista de Saúde Pública, São Paulo, v. 30, n. 2, p. 141 - 147, abr.1996.

ARBACHE, J. Pobreza e mercados no Brasil. In: COMISSÃO ECONÔMICA PARA AMÉRICA LATINA E O CARIBE Pobreza e mercados no Brasil: uma análise de iniciativas de políticas públicas. Brasília, 2003.

ATHEY, S., IMBENS, G. W. Identification and inference in nonlinear deifference-in-differences models. Stanford: National Bureau of Economic Research, 2002. (NBER Technical Working Paper, 0280).

ATTANASIO, O. et al. Baseline report on the evaluation of familias en accion. Bogota: Centre for the Evaluation of Development Policies, 2002.

AVILEZ M. J. Recolección de datos. [2007?] Disponível em: <http://www.monografias.com/trabajos12/recoldat/recoldat.shtml>. Acesso em: 15 jul. 2007.

BAKER, Y. L. Evaluación del impacto de los proyectos de desarrollo en la pobreza. Washington DC: Banco Mundial, 2000. Disponível em:<http://www.worldbank.org>. Acesso em: 22 ago. 2007.

BANCO MUNDIAL. The contribution of social protection to the milennium goals. Washington, DC: 2003a. Disponível em: http://siteresources.worldbank.org/ SOCIALPROTECTION/Publications/20847137/SPMDGs.pdf >. Acesso em: 19 abr. 2008.

BANCO MUNDIAL. Development Research Group. Evaluating anti-poverty programs. In: EVENSON, R. E; SCHULTZ, T. P. (Ed.). Handbook of development economics. Amsterdam, North-Holland, 2003b. v.4.

BANCO MUNDIAL. Hunger zero project. [2001?] Disponível em: <http://www.fomezero.gov.br/publicacoes/arquivos/programa_fz_ingles.pdf>. Acesso em: 23 mar. 2007.

BARROS, R.; CARVALHO, M.; MENDONÇA, R. Sobre as utilidades do Cadastro Único. Niterói: Universidade Federal Fluminense, Faculdade de Economia, 2008. (Texto para Discussão, 244).

BARTHOLOMEW, D. J. A method of allowing for 'not-at-home' bias in sample surveys, applied statistics. A Journal of the Royal Statistical Society, London, v. 10, n.1, p. 52-59, Mar. 1961.

BECKER, G. S. Human capital: a theorical and empirical analisis, with special reference to education. London: The University of Chicago Press, 1993.

BECKER, S.O.; ICHINO, A. Estimation of average treatment effects based on propensity score. Stata Journal, v. 2, n. 4, p. 358-377, Nov. 2002.

BEHRMAN, J. R.; DURYEA, S.; SZÉKELY.; M. Schooling investments and aggregate conditions: a household-survey-based approach for Latin America and the Caribbean. Washington, DC: Inter-American Development Bank, 1999. Não publicado.

BEHRMAN, J.; SENGUPTA, P.; TODD, P. Progressing through PROGRESA: an impact assessment of a school subsidy experiment. Washington, D.C: International Food Policy Research Institute, 2001.

BERTRAND, M.; DUFLO, E.; MULLAINATHAN, S. How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, Cambridge, v. 119, n. 1, p 249-275, Feb. 2004.

BLACK, D.; GALDO, J.; SMITH, J. Evaluating the regression discontinuity design using experimental data. Minchigan: University of Michigan, 2005.

BLUNDELL, R.; COSTA, D. M. Evaluation methods for non-experimental data. Fiscal Studies, London, v. 21, n. 4, p. 427–468, Jan. 2000.

BOHLAND, A. K. Óbitos de mulheres em idade fértil em Aracaju (SE): estratégias para melhorar a qualidade da informação. 115 f. Tese (Doutorado em Epidemiologia) - Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo, 2003.

BOUSSY, C. A.; SCOTT, K. G. Use of data base linkage methodology in epidemiological studies of mental retardation. International Review of Research in Mental Retardation, San Diego, v. 19, p. 135-161, 1993.

BRASIL. Ministério da Saúde. Departamento de Gestão de Políticas Estratégicas: Secretaria de Políticas de Saúde. Programa de Saúde da Criança Governo federal lança programa de combate à desnutrição. Brasília, DF, [200-?a]. Disponível em: <http://www.rebidia.org.br/novida/bolsa_alim.htm#ATOS%20DO%20PODER>. Acesso em: 20 mar. 2007.

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. Benefício de prestação continuada de assistência social (BPC) Brasília, DF, [200-?b] Disponível em: <http://www.mds.gov.br/programas/rede-suas/protecao-social-basica/beneficio-de-prestacao-continuada-bpc>. Acesso em: 25 abr. 2007.

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. O Programa Bolsa Família. Brasília, DF, [200-?c]. Disponível em <http://www.mds.gov.br/bolsafamilia/o_Programa_bolsa_familia>. Acesso em: 29 mar. 2007.

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. Programa de erradicação do trabalho infantil (PETI). Brasília, DF, [200-?d]. Disponível em: http://www.mds.gov.br/programas/rede-suas/protecao-social-especial/programa-de-erradicacao-do-trabalho-infantil-peti. Acesso em: 25 mar. 2007.

BRASIL. Ministério de Desenvolvimento Social e Combate à Fome. Projeto Agente Jovem de Desenvolvimento Humano. Brasília, DF, [200-?e] Disponível em: http://www.mds.gov.br/programas/rede-suas/protecao-social-basica/servicos-e-usuarios

/concessao-de-bolsa-para-jovens-em-situacao-de-vulnerabilidade-socia/projeto-agente-jovem-de-desenvolvimento-humano>. Acesso em: 25 set. 2007.

BRASIL. Gerência de Filial de Serviços Sociais (GISES) – Caixa Econômica Federal Brasília, DF, [200-?f]. Disponível em: <http://www.quatrobarrasparana.com.br/acaosocial/Inf_F_Pagamento.pdf>. Acesso em: 12 jul. 2008.

BROUSSEAU, R.; MONTALVÁN, G. Curso de monitoreo y evaluación de proyectos. Banco Interamericano de Desarrollo. [2007?] Disponível em:< http://www.iadb.org/int/rtc/ecourses/esp>. Acesso em: 26 jun. 2008.

BUDDELMEYER, H.; SKOUFIAS, E. An evaluation of the performance of regression discontinuity design on PROGRESA. Washington, DC: World Bank, 2004. (Policy Research Working Paper , 3386).

BURTLESS, G. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives, Washintong, D.C, v. 9, n. 2, p. 63-84, Spring. 1995.

CAIXA ECONÔMICA FEDERAL. Transferência de benefícios: CAIXA vai pagar auxílio-gás ainda em fevereiro. 2002. Disponível em: <http://www1.caixa.gov.br/imprensa/imprensa_release.asp?codigo=300822&tipo_noticia=0>. Acesso em: 20 abr. 2007.

CAMARGO JR., K. R.; COELI, C. M. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Revista Brasileira de Epidemiologia, São Paulo, v. 5, n. 2, ago. 2002a.

CAMARGO JR., K. R; COELI, C. M. Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. Cadernos de Saúde Pública, Rio de Janeiro: v. 16, n. 2, p. 439-47. abr./jun.. 2000.

CAMARGO JR., K. R.; COELI, C. M. Reclink II: guia do usuário. Rio de Janeiro, 2002b. Disponível em: http://paginas.terra.com.br/educacao/kencamargo/RecLinkdl.html. Acesso em: 02 mar. 2007.

CEPAL. Directorios estadísticos de empresas elaborados a partir de registros administrativos. In: CONFERENCIA ESTADÍSTICA DE LAS AMÉRICAS DE LA CEPAL, 2., 2003, Santiago de Chile. Informe. Santiago de Chile, 2003a.

CEPAL. Panorama social de América Latina. Santiago de Chile, 2004. Documento informativo.

CEPAL. Registros administrativos, calidad de los datos y credibilidad pública. In: CONFERENCIA ESTADÍSTICA DE LAS AMÉRICAS DE LA CEPAL, 2., 2003, Santiago de Chile. Informe. Santiago de Chile, 2003b.

CERVO, A. R.; BERVIAN, P. A. Metodologia científica. 5 ed. São Paulo: Prentice Hall, 2002.

CHRISTEN, P.; CHURCHES, T. Secure health data linkage and geocoding: current approaches and research directions. In: NATIONAL E-HEALTH PRIVACY AND SECURITY SYMPOSIUM, Brisbane, 2006. Proceedings... [2006?].

COADY, D.; GROSH, M.; HODDINOTT, J. Targeting of transfers in developing countries: review of lessons and experience. Washington; World Bank, 2004.

COELI, C. M. et al. Probabilistic linkage in household survey on hospital care usage. Revista de Saúde Pública, São Paulo, v. 27, n. 1, p. 91 - 99, fev. 2003.

COHEN, E., et. al. Los desafíos de la reforma del estado en los programas sociales: tres estudios de caso. Santiago de Chile: CEPAL, 2001. (Serie de Políticas Sociales, 45)

COHEN, E.; FRANCO, R. Evaluación de proyectos sociales. Santiago de Chile: Instituto Latinoamericano y del Caribe de Planificación Económica y Social, 1988.

COLOMBIA. Departamento Nacional de Planeación Dirección de Evaluación de Políticas Públicas. Programa Familias en Acción: impactos en capital humano y evaluación beneficio - costo del programa. Bogotá, 2007.

DIAZ, J. J.; HANDA, S. An assessment of propensity score matching as a nonexperimental impact estimator: evidence from a mexican poverty program. Carolina do Norte: Office of Evaluation and Oversight, 2005. (Working Paper: OVE/WP, 04/05).

DU BOIS, D. N. S. A solution to the problem of linking multivariate documents. Journal of the American Statistical Association, Virginia, v. 64, n. 33, p. 163-174. Mar. 1969.

DUNN, H. L. Record linkage. American Journal of Public Health, Washington, D.C, v. 36 n. 12, p. 1412-1416, Dec., 1946.

DURÁN, C. Evaluación microeconométrica de las políticas públicas de empleo: aspectos metodológicos. Hacienda Pública Española. Revista de Economía Pública, Madrid, v. 170, n. 3, p.107-133, set. 2004.

EZEMINARI, K.; RUDQVIST, A.; SUBBARAO, K. Impact evaluation concepts and methods. En evaluation and poverty reduction. Washington, D.C: World Bank, 2002.

FAIR, M. Fetal-infant mortality study group of the canadian perinatal surveillance system. validation study for a record linkage of births and infant deaths in Canada. Ottawa: Statistics Canada, 1999. (Catalogue, 84F0013-XIE).

FAIR, M. E. Recent developments at statistics Canada in the linking of complex health files. In: FCSM RESEARCH CONFERENCE PAPERS, 1999. Session IX-A. [1999]. Disponível em: <http://www.fcsm.gov/99papers>. Acesso em: 15 out. 2006.

FELLEGI, I. P.; SUNTER, A. A theory of record linkage. Journal of the American Statistical Association, New York,v. 64, n. 328, p. 1183-1210, Dec. 1969.

FERNANDES, D. M. Concatenamento de informações sobre óbitos e nascimentos: uma experiência metodológica do Distrito Federal 1989. 1991. 71f. Tese (Doutorado em Demografia) – Centro de Desenvolvimento e Planejamento Regional, Universidade Federal de Minas Gerais, Belo Horizonte, 1997.

FLORIDI, L. Is semantic information meaningful data? Philosophy and Phenomenological Research, Oxford, v. 70, n. 2, Mar. 2005.

FOWLER, A. F. Assessing NGO performance: difficulties, dilemmas and a way ahead. In: EDWARDS, M.; HULME, D. Beyond the magic bullet: NGO performance and accountability in the post – cold war world. Connecticut: Kumarian, 1996.

FREEMAN, H.; ROSSI, P. Y.; WRIGHT, S. Evaluating social projects in developing countries. Paris: Development Centre/Organisation for Economic Co-operation and Development, 1980.

GALASSO, E.; RAVALLION, M.; SALVIA, A. Assisting the transition from workfare to work: a randomized experiment. Industrial and Labor Relations Review, v. 58, n. 1, p. 128-142, Oct. 2004.

GILL, L. E. E.; BALDWIN, J. A. Methods and technology of record linkage: some practical considerations. In: ACHESON, E. D.; GRAHAM, W. J. Textbook of medical record linkage. Oxford: Oxford University, 1987. p.39-54.

GILL, L. Methods for automatic record matching and linking in their use in national statistics. London: Office for National Statistics, 2001. (National Statistics Methodological Series, 25)

GOLDACRE, M. J. Implications of record linkage for health services management. In: BALDWIN, J. A.; ACHESON, E. D.; GRAHAM, W. J. Textbook of medical record linkage. Oxford: Oxford University, 1987. p.305-317.

GOMATAM, S.; CARTER, R. A computerized stepwise deterministic strategy for linkage. Gainesville: University of Florida, Department of Statistics, 1999. Technical Report.

GÓMEZ, L. C.; MURGUEITIO, C.; RODRIGUEZ, M. Evaluación de impacto del programa familias en acción. Bogota: Unión Temporal IFS, Econometría s.a. SEI s.a, 2006. Informe Final.

GU, L. Record linkage: current practice and future directions. Canberra: CSIRO, Mathematical and Information Sciences, 1983. (Technical Report, 03-83).

HECKMAN, J. et al. Characterizing selection bias. Using experimental data. Econometrica, Chicago, v. 66, n. 5, p. 1017-1089, Sept. 1998.

HECKMAN, J.; HOTZ, J. Choosing among alternative non experimental methods for estimating the impact of social programs: the case of manpower training. Journal of the American Statistical Association, Chicago, v. 84, n. 408, p. 862-880, Dec; 1989.

HECKMAN, J.; ICHIMURA, H.; TODD, P. Matching as an econometric evaluation. estimator: evidence from evaluating a job training program. Review of Economic Studies, Oxford, v. 64, n. 4, p. 605-654, Oct. 1997.

HECKMAN, J.; LALONDE, R.; SMITH, J. The economics and econometrics of active. labor market programs. In: ASHENFELTER, O.; CARD, D. (Ed.) The handbook of labor economics. Amsterdam: North Holland, 1999. v.3a, pt.6, cap.31, p.1865-2097.

HECKMAN, J. Randomization and social policy evaluation. In: MANSKI, C.; GARFINKEL, I. (Ed.). Evaluating welfare andtraining programs. Cambridge: Harvard University Press, 1992.

HECKMAN, J.; VYTLACIL, E. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings of the National Academy of Sciences of the United States of America, Chicago, v. 96, n. 8, p. 4730-4734, Apr. 1999.

HECKMAN J.; VYTLACIl, E. Structural equations, treatment effects and econometric policy evaluation. Cambridge: National Bureau of Economic Research, 2005. (NBER Technical Working Paper, 306).

HODDINOTT, J.; SKOUFIAS, E.; WASHBURN, R. The impact of PROGRESA on consumption: a final report. Washington, D.C.; International Food Policy Research Institute, 2000.

HOWE, G. R.; LINDSAY, J. A generalized iterative record linkage computer system for use in medical follow-up studies. Computers and Biomedical Research, Arlington, v. 14, n. 4, p 327-340, Aug. 1981.

HOWE, G. R Use of computerized record linkage in cohort studies. Epidemiologic Reviews, New York, v. 20, n. 1, p. 112-21, 1998.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Síntese de indicadores sociais: uma análise das condições de vida da população brasileira 2007. [2007?] Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/ condição devida/ indicadoresminimos/sinteseindicsociais2007/indic_sociais2007.pdf> Acesso em: 09 abr. 2008.

JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, Florida, v. 84, n. 406, p. 414-420, June. 1989.

JENSEN, K., P. Probabilistic methodology for record linkage determining robustness of weights. 2004. A project submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Master of Science

KASSOUF A. L. Análise das políticas e programas sociais no Brasil. Brasília: OIT/Programa IPEC América do Sul, 2004. 108 p. (Documentos de Trabajo, 182).

KIRKENDALL, N. J. Weights in computer matching: applications and an information theoretic point of view. In: KILSS, B.; ALVEY, W. (Ed.). Record linkage techniques: proceedings o the Workshop o Exact Mactching Methodogies, Arlington, Virginia, 1985. 1985. p. 189-196. Disponível em: <http://www.fcsm.gov/working-papers/1367.pdf>. Acesso em: 15 maio 2008.

KNUTH, D. E. The art of computer programming. 2nd ed. Massachusetts: Addison-Wesley, 1973. v. 1, cap. 2.

LaLONDE, R. Evaluating the econometric evaluation of training programs with experimental data. The American Economics Review, Nashville, v. 76, n. 4, p. 604-620, Sept 1986.

MACHADO, C. J. Early infant morbidity and infant mortality in the city of São Paulo, Brazil: a probabilistic approach. 336 f. Tese (Doutorado em Filosofia) – Johns Hopkins University, Baltimore. 2002.

MACHADO C. J. A literatura review of record linkage procedures focusing on infant health outcomes. Cadernos de Saúde Pública, Rio de Janeiro, v. 20, n. 2, p. 362-371, abr. 2004.

MALUCCIO, J. A.; FLORES, R. Impact evaluation of a conditional cash transfer program: the Nicaraguan Washington, DC.: International Food Policy Research Institute, 2005. (Red de Protección Social Research Report, 141)

MALUCCIO, J. A. Nicaragua: Red de protección social — Mi familia rompiendo el ciclo de pobreza. Washington, DC.: International Food Policy Research Institute, 2005.

MARCONI, M. A.; LAKATOS, E. M. Fundamentos de metodologia científica. 5. ed.. São Paulo: Atlas, 2003.

MELLO, A. L. C ; COUTINHO, E S. F.; COELI, C. M. Prevalência de casos de acidente vascular encefálico, município do Rio de Janeiro – 1998. Cadernos Saúde Coletiva, Rio de Janeiro, v. 14, n. 2, p. 345-360, abr./jun. 2006.

MOFFITT, R. A. The role of randomized field trials in social science research: a perspective from evaluations of reforms of social welfare programs. London: Institute for Research on Poverty, 2003. (Discussion Paper, 1264-03)

MOHR, L. Impact analysis for program evaluation. Ann Arbor: University of Michigan, 1988.

MOHR, L. The qualitative method of impact analysis. American Journal of Evaluation, Ann Arbor, v. 20, n. 1, p.69-84, 1999.

NAVARRO, H. Manual para la evaluación de impacto de proyectos y programas de lucha contra la pobreza. Santiago de Chile: Instituto Latinoamericano y del Caribe de Planificación Económica y Social, 2005.

NEWCOMBE, H. B. Automatic linkage of vital records. Science, Washington, D.C., v. 30, n. 130, p. 954-959, Oct 1959.

NEWCOMBE, H. B.; FAIR, M. E.;, LALONDE, P. Discriminating powers of partial agreements of names for linking personal records. Methods of Information in Medicine, Silver Spring, v. 28, n. 2, p. 86-91, Apr. 1989.

NEWCOMBE, H. B. Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford; Oxford University Press, 1988.

NORONHA, C. P.; SILVA, R. I.; THEME FILHA, M. M. Concordância de dados das declarações de óbitos e de nascidos vivos para a mortalidade neonatal no município do Rio de Janeiro. Informe Epidemiológico do SUS, Brasília, v. 4, n. 4, p. 57-65, 1997.

OLIVEIRA, A. et al. Primeiros resultados da análise da linha de base da pesquisa de avaliação de impacto do programa bolsa família. In: VAITSMAN, J.; SOUSA, R. P. Avaliação de políticas e programas do mds –Resultados: Bolsa Família e Assistência Social. Brasília, DF: Ministério do Desenvolvimento e Combate a Fome, Secretaria de Avaliação e Gestão da Informação, 2007. v.2

PATTON, M. Qualitative research & evalaution methods. 3rd ed. Thousand Oaks: SAGE,. 2002.

RAMOS, C. E.; SANTANA, R. Os pobres que levantem a mão (mas será que são mesmo pobres?). Uma tentativa de validar o cadastro único. Brasília: Universidade de Brasília, 2002.

RAVALLION, M. How can qualitative methods help in measuring poverty? Banco Mundial, 2002.

RAVALLION, M. The mystery of the vanishing benefits: Ms Speedy Analyst's introduction to evaluation. Washington, D.C.: Banco Mundial, 2001.

RAWLINGS, L. B. E.; RUBIO, G. M. Evaluación del impacto de los programas de transferencias condicionadas en efectivo: lecciones desde América Latina. México: Secretaría de Desarollo Social, 2003. (Cuadernos de Desarrollo Humano, 10).

RICHARDSON, R. J. Pesquisa social: métodos e técnicas. 3. ed. São Paulo: Atlas, 1999.

ROOS, L. L.; WAJDA, A. Record linkage strategies. Methods of Information in Medicine, Silver Spring, v. 30, n. 2, p. 117–123, Apr. 1991.

ROSENBAUM, P.; RUBIN, D. The central role of the propensity score in observational studies for causal effects. Biometrika, London, v.70, n. 1, p. 41-55, Apr. 1983.

SCHEUREN, F.; WINKLER, W. E. Regression analysis of data files that are computer matched – Part II. 1997. Disponível em: <http://www.fcsm.gov/working-papers/scheuren_part2.pdf>. Acesso em: 10 mar. 2008.

SCHUTT, R.I Investigating the social world: the process and practice of research. Thousand Oaks: Pine Forge Press, 2001.

SILVA, M. O. Os programas de transferência de renda enquanto estratégia de enfrentamento à pobreza no Brasil: possibilidades e limites. [2006?]. Trabalho apresentado ao 32nd International Conference on Social Welfare, Brasília, julho de 2006.

SKOUFIAS, E.; DAVIS, E.; VEGA, S. de la. Focalización de los pobres en méxico: evaluación de la selección de hogares que participan en progresa. Washington, D.C.; International Food Policy Research Institute, 1999.

SKOUFIAS, E. PROGRESA y su efecto sobre el bienestar de las familias rurales de México. Washington, DC.: International Food Policy Research Institute, 2006. (Informe de Investigación, 139).

SMITH, M. E. Record – keeping and data preparation practices to facilitate record linkages. In: KILSS, B.; ALVEY, W. (Ed.). Record linkage techniques: proceedings o the Workshop o Exact Mactching Methodogies, Arlington, Virginia, 1985. 1985. p. 321-326. Disponível em: <http://www.fcsm.gov/working-papers/1367.pdf>. Acesso em: 15 maio 2008.

TINTÓ, M. La difusión de y el acceso a la información administrativa electrónica. Trabalho apresentado ao II Congreso Online Del Observatorio Para la Cibersociedad, Cornellà de Llobregat, Barcelona. [2004?]. Disponível em: <http://www.cibersociedad.net/congres2004/grups/fitxacom_publica2.php?idioma=pt&id=654&grup=56> Acesso em: 10 mar.2008.

VACCARISI, M. E. Asistencia social y políticas alimentarías: tensión entre legitimación y control social. In: SUJETOS sociales y políticas: historia reciente de la Norpatagonia Argentina. Neuquén: Universidad Nacional del Comahue, Centro de Estudios Historicos de Estado, Politica y Cultura, 2005. cap. 6.

VAUS, D. de. Surveys in social research. Journal of Sociology, London, v. 22, n. 3, p. 492-494, 1986.

VÉLEZ, C.; CASTAÑO, E.; DEUTSCH, R. An economic interpretation of targeting systems for social programs. Washington, D.C.: SISBEN, 1999.

WEBER, G. I.. Achieving a patient unit record within electronic record systems. In MEDICAL RECORDS INSTITUTE (Ed.). Toward an electronic patient record. Newton, Ma, 1995. p. 126-134.

WEISS, C. H. Evaluation: methods for studying programs & policies. 2.nd. ed. Englewood Cliffs: Prentice Hall, 1998.

WHALEN D. et al. Linking client records from substance abuse, mental health and medicaid state agencies. Rockville: U.S. Department Of Health And Human Services, 2001.

WHITE, H. Combining qualitative and quantitative approaches in poverty análisis. Brighton: Institute of Development Studies, 2002.

WINKLER, W. E. Advanced methods for record linkage. Washington, DC.: Statistical Research Division, U.S. Bureau of the Census, 1994. p. 467-472. Technical Report Disponível em: <http://www.amstat.org/Sections/Srms/Proceedings/papers/1994_077.pdf> Acesso em: 14 out. 2007.

WINKLER, W. E. Improved decision rules in the fellegi-sunter model of record linkage. [1993?]. Disponível em: <http://www.census.gov/srd/papers/pdf/rr93-12.pdf> Acesso em: 17 out. 2007.

WINKLER, W. E. Near automatic weight computation in the fellegi-sunter model of record linkage, In: ANNUAL RESEARCH CONFERENCE, 5, 1989, Washington, DC. Proceedings... Washington, D.C.; Bureau of the Census, 1989.

WINKLER, W. E.; THIBAUDEAU, Y. An application of the Fellegi-Sunter model of record linkage to the 1990 U. S. decimal census. [1991?]. Disponível em: http://www.census.gov/srd/papers/pdf/rr91-9.pdf>. Acesso em: 17 out. 2007.

WODON, Q. et al. PROGRESA de Mexico: focalización innovadora, centrado en el género y sus efectos en el bienestar social. En breve: una serie regular de notas destacando las lecciones recientes del programa operacional y analítico de la región de América Latina y el Caribe, del Banco Mundial, Washington, n. 17, ene. 2003.

ZEPEDA, E. Transferências condicionadas de renda (tcr) reduzem a pobreza? One Pager, Brasília, n. 21, abr. 2008.

ANEXO I: Evidências de relacionamento de bases de dados nos paises desenvolvidos

A seguir, são apresentados alguns trabalhos que tratam do relacionamento de bases de dados:

• Nos Estados Unidos um número de seguro social foi criado em 1935, mas sua utilização limitou-se ao no programa de seguros. O não uso deste número social tornou o relacionamento de dados é uma tarefa difícil. Muitas bases de dados dos países desenvolvidos contêm um identificador único que é utilizado para integrar grandes bases provenientes de diferentes fontes de dados, no entanto, este identificador único nem sempre é utilizado ou atualizado para toda a população. Dessa forma, quando fez-se necessário pesquisar as características genealógicas das pessoas nos registros dos censos, foram propostos métodos de relacionamento probabilístico para de determinar a viabilidade de unir as pessoas valendo-se dos anos de coleta de dados do censo. Com a utilização desta metodologia pode-se diminuir ou eliminar a necessidade de realizar uma revisão manual em grandes números de registros. Os dados utilizados para o relacionamento probabilísticos nesse trabalho, correspondem a uma amostra do censo dos anos 1910 e 1920, e ilustram os benefícios de uma aproximação automatizada de relacionar registros provenientes de censo de população (JENSEN, 2004).

• Também nos Estados Unidos tem sido realizado um trabalho para avaliar o risco de emergirem arquivos denominados confidenciais, conhecidos como Arquivos de Uso de Público (PUF), tais documentos contêm dados sintéticos, criados a partir de um arquivo de dados confidenciais. Para avaliar o risco de descoberta destes arquivos tem-se utilizado o relacionamento de bases de dados automático. O procedimento utilizado relaciona os arquivos PUF aos dados de arquivos confidenciais do qual eles foram construídos originalmente. Este trabalho está vinculado ao projeto longitudinal da dinâmica empregador-empregado (LEHD), que são dados em desenvolvimento de arquivos que contêm informação combinadas, relacionando o trabalhador selecionado e o empregador registrado em uma pesquisa estatística (WALA, STINSON, ABOWD, 2005).

• Uma das aplicações mais difundidas de relacionamentos de dados computacionais no contexto de estudos de coorte nos Estados Unidos é provida pelo Índice de Morte Nacional (NDI). O NDI é um registro de todas as mortes que acontecem nos Estados Unidos e é administrada pelo Centro Nacional de Estatísticas de Saúde (Bilgrad, 1990). Os registros de mortes computados são providos ao NDI pelos escritórios de estatísticas vitais em todos os 50 estados, no distrito de Columbia, em Porto Rico, e nas Ilhas de Virgem, dentro de 12 meses da conclusão de cada ano civil. O NDI contém mortes desde 1979, com aproximadamente 2 milhões de mortes somadas em cada ano. O NDI pode ser usado por epidemiologistas e outros especialistas para agregar os dados inicialmente de coortes, podendo identificar data, fato e estado onde a morte aconteceu e a coorte em que morreu o indivíduo. A combinação dos registros do NDI, junto com o relacionamento de dados probabilístico, provê um recurso excelente para administrar estudos de coorte nos Estados Unidos no qual a morte é o ponto final (HOWE, 1998, BUEHLER JW, PRAGER K, HOGUE CJ, 2000).

• Outra aplicação adicional realizada nos Estados Unidos é o relacionamento de dados que integra a informação sobre o serviço de saúde mental (MH) e de álcool/drogas (AOD). Patrocinado pelo Centro de Serviços Administrativos de Abusos de Sustâncias e Saúde Mental (SAMHSA) para o tratamento de Abuso de Substância proibido (CSAT) e o Centro para Saúde Mental Conserta (CMHS), o Grupo de MEDSTAT construiu bases de dados como parte de um convênio (270-96-0007), cada Base de Dados Integrada (IDB) inclui informação de interesses para as agências estatais de MH e AOD, como também Agências de Medicina para três estados: Delaware, Oklahoma, e Washington (WHALEN et al., 2001).

• Na Canadá, segundo Fair (1999?), o relacionamento probabilístico é considerado o método de relacionamento preferível. A evidência desta afirmação é mostrada em um estudo que relaciona registros de nascimentos com os de mortalidade infantil em Nova Scotia e Alberta, mostrando que 99% de mortes infantis dos dados locais de Nova Scotia foram localizadas com êxito nas Estatísticas de Saúde Pública do arquivo do Canadá. Um dos objetivos do relacionamento das informações é analisar entre todas as variáveis, a idade gestacional e o peso ao nascer das crianças (FAIR, 1999?; MACHADO, 2002).

• Outro estudo no Canadá sobre os registros de nascimento e mortes foi realizado com um objetivo definido. Neste caso foram considerados os nascimentos vivos durante os anos

1985-1994 selecionados da Base de Dados de Nascimentos Canadense (CBDB). Todos os registros de nascimentos durante esses anos foram incluídos no relacionamento, e foram feitas exclusões necessárias posteriores devido a identificadores perdidos. Por outro lado, foram selecionados registros desde a Base de Dados de Mortalidade Canadense CMDB durante os anos 1985-1995 para crianças nascidas nos anos 1985-1995. Para assegurar que todas as mortes infantis realmente haviam sido incluídas, foram também selecionados registros de mortes codificados com causa de morte perinatal ou por anomalia congênita. Dados Geográficos (como, código postal e subdivisão de censo) e peso de nascimento foram acrescentados aos registros de morte utilizando o relacionamento de dados. No relacionamento de nascimento e mortes, formatos de nome estavam separados por partes de componente. Além do formato, havia também o problema de apelidos, títulos, pseudônimos, nomes múltiplos, só iniciais, sufixos (como Jr.), sinônimos por perda (bebê, gêmeo, etc.). Para controlar problemas como esses foram desenvolvidos programas de computação (FAIR, 1999?).

• Outro problema abordado com o relacionamento de dados é a estimação da subcobertura ou sub-registros e da sobre-cobertura nos censos populacionais, como é o caso do Canadá, que não tem uma pesquisa de enumeração posterior para mensurar a subestimação e super-estimação dos censos. O primeiro estudo de cobertura primário dos censos no Canadá é o *Reverse Record Check* (RRC) em 1966. O propósito do RRC é calcular os erros de cobertura da população e das moradias privadas, além disso, procura analisar as características das pessoas que não foram enumerados ou foram enumerados mais de uma vez no momento do censo. O RRC utiliza uma amostra independente de pessoas que deveriam ter sido enumeradas no momento do censo. Entre as informações extraídas dos registros se for possível o RRC obtém os endereços das pessoas selecionadas e dos membros familiares atualizados através dos registros administrativos. Operações de recuperação de registros são levadas a cabo por meio de entrevistas, com o objetivo de contatar à pessoa selecionada a quem se direciona o questionário. Estas operações de recuperação são conferidas com os questionários e a base de dados do censo para determinar quantas vezes a pessoa selecionada é enumerada. O relacionamento probabilístico é usado no procedimento de atualização dos endereços, e tendo sido constituído em duas fases. Primeiramente, o relacionamento probabilístico une o arquivo RRC com uma primeira base de *Revenue Canada* (RCT), que apresenta informação do rendimento da pessoa. Uma vez que este relacionamento é completado, o Número de

Seguro Social (SIN) da pessoa selecionada ou de um membro da família é obtido. Na segunda fase, uma partida exata é feita entre o RRC e a base mais recente do RCT para obter o endereço mais recente disponível nesses arquivos (BERNIER, 1997)

• Ainda no Canadá, em um estudo similar ao anterior, foi feito um estudo de relacionamento automatizado (AMS) com o objetivo de estimar a sobre-cobertura das moradias privadas no censo. O AMS é uma série de programas computacionais que identificam pares de moradias que são "similares", definidos em termos do número de membros das duas moradias e da proximidade geográfica relativa entre elas. Os pares de moradias identificados constituem uma amostra do *survey*, a qual são conferidos com uma amostra dos questionários do censo para determinar em quanto se estima a sobre-cobertura. (HA, MAYDA e TOURIGNY, 1998).

• Na Austrália, uma Pesquisa de pós-enumeração (PES), administrada independentemente do censo utiliza o sistema de estimação dual (DSE) para calcular a subestimação (Dunstan et al, 1999). Uma amostra é extraída da Pesquisa da Força de trabalho ABS. O PES coleta as informações por uma entrevista face a face, solicitando aos entrevistados o endereço onde eles possam ter sido incluídos no formulário do censo. Os visitantes são registrados no *survey* PES e fornecem o endereço de residência habitual. O PES coleciona o nome, sexo e data de nascimento ou idade para facilitar um relacionamento mais preciso. Executa-se uma revisão manual de indivíduos utilizando o censo físico e formulários do PES. Finalmente são relacionadas visitas que podem ter sido enumeradas em domicílios não amostrados, para buscar os endereços onde se percebem que as famílias podem ter sido enumeradas. Estas respostas são utilizadas para determinar o número de vezes que cada entrevistado foi incluído no censo (WOOLFORD, 2001).

• Os procedimentos em Nova Zelândia para medir e ajustar a subestimação são semelhantes aos adotados na Austrália. A pesquisa pós-enumeração de Nova Zelândia (PES-NZ) é uma amostra de unidades domésticas extraídas da pesquisa de domicílios da força de trabalho (SNZ) (Dunstan et *a*l, 1999). A amostra cobre aproximadamente 0,8% das moradias privadas totais do País. A informação coletada no PES é semelhante ao PES australiano. O relacionamento de indivíduos é um exercício manual, utilizando o formulário físico do PES e as imagens dos formulários do Censo. Quando a informação de endereços é insuficiente, o status de pareamento foi imputado (WOOLFORD, 2001).

• Na Escócia o Serviço Nacional de Saúde (NHS) executou mais de 150 exercícios de relacionamento até 1997. Esses exercícios envolveram, principalmente, relacionar conjunto de dados externos (i.e. dados de *surveys*, auditoria clínica, entre outros) para serem centralizados nos registros de saúde (Kendrick, 1997). Provavelmente o trabalho de maior pertinência para o *One Number Censu* (ONC) foi o relacionamento do Índice de Saúde de Comunidade (CHI) e dados do Registro Central de Serviços de Saúde Nacional (NHSCR). Este relacionamento combinou estratégias determinísticos e probabilísticas. A comparação de registros de CHI com registros de NHSCR foi realizada em três fases. Foram localizados resultados para cada registro de CHI com o identificador único do número de NHSCR correspondente. Os registros de CHI continham "a data de aceitação através de prática de GP (General Practitioners)" que poderia ser comparado com "a data de transferência para da tabela da saúde atual" no registro de NHSCR. Assim, números de NHS ficaram disponíveis em todos os registros NHSCR e na maioria dos registros de CHI (WOOLFORD, 2001).

ANEXO II: Métodos de estimação de impacto para desenhos não experimentais

1. Método diferença em diferença ou diferença dupla.

A estimação Diferença em Diferença (DD) tem crescido nos últimos anos como o método mais popular para estimar relações causais. Este método consiste em comparar um grupo de tratamento e um de controle antes (primeira diferença) e depois de um programa (segunda diferença) (HECKMAN, ICHIMURA, SMITH e TODD, 1998; MORDUCH, 1999; BLUNDELL e COSTA DIAS, 2002; AGHION e MURDOCH, 2005; CALIENDO e HUJER, 2005).

As comparações simples dos resultados pré-tratamento e pos-tratamento, para os indivíduos expostos, provavelmente o tratamento será contaminado pelas tendências temporais na variável de resultado ou pelo efeito do evento. Diferentemente de outros tratamentos, isto acontece entre ambos os períodos. Porém quando só uma parte da população é exposta ao tratamento, um grupo de comparação controle pode ser utilizado para identificar a variação temporal no resultado que não é devido à exposição ao tratamento (ABADIE, 2003).

O estimador de DD é baseado numa idéia simples. Card e Krueger (1994) ajustaram o efeito do emprego no aumento do salário mínimo em New Jersey utilizando o estado vizinho, Pennsylvania, para identificar a variação do emprego em New Jersey que deveria ter experimentado na ausência do aumento do salário mínimo. Outro estudo de aplicações do DD inclui efeito de salário e empregos dos imigrantes sobre os nativos (CARD, 1990), efeitos dos benefícios de incapacidade temporal no tempo fora do trabalho depois de acidentar-se (Meyer, Viscusi, and Durbin, 1995), e o efeito das leis anti- aquisição sobre formas de empréstimos (GARVEY e HANKA(1999)).

O grande recurso da estimação de DD é dado pela simplicidade, como também pelo potencial para evitar muito dos problemas de endogeneidade que tipicamente surgem ao se fazer comparações entre indivíduos heterogêneos (MEYER, 1995). No entanto, a estimação DD tem suas limitações. Este método é apropriado quando a intervenção é tão boa quanto a aleatorização, condicionada ao tempo e ao efeito fixo dos grupos de

comparação, Assim, muitos dos debates sobre a validez da estimação DD tipicamente giram em torno da possível endogeneidade dos resultados da intervenção (BERTRAND, DUFLO e MULLAINATHAN, 2003).

Este método é útil avaliando mudanças de política em ambientes na quais as tendências de tempo subjacentes importantes estão presentes. Este método é mais popular na avaliação de mudanças de política de governo que acontecem em algumas unidades administrativas, como distritos escolares ou estados, mas não em unidades vizinhas. (ATHEY e IMBENS, 2002).

2. Comparações reflexivas:

É outro tipo de modelo não-experimental. Neste modelo, realiza-se uma pesquisa de referência junto aos participantes antes da intervenção do programa, com a qual é construído o contrafactual. Logo se realiza uma pesquisa de acompanhamento quando o programa está em andamento. Assim, são comparados os participantes de programa antes e depois da intervenção. O efeito de impacto é mensurado através da mudança nos indicadores de resultados antes e depois da intervenção. Este tipo de desenho é particularmente útil em avaliações de intervenções de cobertura total, tal como políticas de âmbito nacional e programas nos quais a população inteira participa e não há nenhum espaço para um grupo de controle. (BAKER, 2000)

A desvantagem principal no modelo de comparações reflexiva é que a situação dos participantes do programa antes e depois da intervenção pode mudar em grandes quantidades devido a razões independentes ao programa. Por exemplo, participantes em um programa de treinamento de trabalho podem ter melhorado a perspectiva do emprego depois do programa. Enquanto esta melhoria possa dever-se ao programa, também possa ser devido ao fato que a economia está recuperando-se de uma crise passada e o emprego está crescendo novamente.

A menos que eles não sejam cuidadosamente realizados, as comparações reflexivas podem não poder distinguir entre o programa e outros efeitos externos, e assim comprometer à confiabilidade dos resultados (BAKER, 2000).

3. Método das variáveis instrumentais.

No caso em que os indivíduos na amostra não são selecionados aleatoriamente da população para a qual se deseja avaliar um determinado programa, o suposto de independência condicionada não parece ser plausível. Neste caso, é necessário recorrer, para estimar o efeito causal, a outros procedimentos com base no suposto de identificadores diferentes e que precisam de informação adicional. Uma destas estimativas é dada pela aplicação do método das variáveis instrumentais. Este método adquiriu uma nova significância e interpretação pela aplicação nos trabalhos de Imbens e Angrist (1994), Heckman e Vytlacil (1999), e Angrist e Krueger (2001), que discutem as variáveis instrumentais, como instrumentos contínuos e discretos, e aplicação para a identificação de efeitos (DURÁN, 2004; RAVALLION, 2001).

O método das variáveis instrumentais utiliza uma ou mais variáveis que influem na participação do programa, mas não nos resultados dada a participação. Identifica a variação exógena nos resultados atribuíveis ao programa, reconhecendo que o estabelecimento não é aleatório, mas intencional. A variável instrumental (VI) é utilizada, primeiro, para predizer a participação no programa, e segundo, observa-se como varia a variável de resultado com os valores projetados.

ANEXO III: Tipos de pareamento (*matching*) baseados no escore de propensão

Matching Vizinho mais Próximo (_Nearest Neighbor Matching_ - NNM).

Existem dois tipos de *matching* de vizinho mais próximo, com e sem reposição, que determinam o número de unidades de controle que se devem parear a cada unidade de tratamento. O *matching* com reposição minimiza a distância no escore de propensão entre observações pareadas de controle e a unidade de tratamento, assim, cada unidade de tratamento pode ser pareada à unidade de controle mais próximo, ainda se a unidade de controle párea-se mais de uma vez. A vantagem desta técnica é que reduz o viés. Por outro lado, no *matching* sem reposição, pareia-se unidades de tratamento com as de controle que possivelmente são muito diferentes em termos do escore de propensão quando temos poucas unidades de controle similares às unidades de tratamento. Isto incrementa o viés, mas poderia melhorar a precisão das estimativas, além disso, para este caso os resultados são muitos sensíveis à ordem no quais as unidades de tratamento pareadas (ROSENBAUM, 1995).

Utilizando uma unidade de controle para cada unidade de tratamento, assegura-se a mínima distância no escore de propensão. Utilizando mais unidades de controle incrementa-se a precisão da estimação, ao custo de incrementar o viés. Uma vez realizado o *matching* de todas as unidades tratadas, a diferença entre o resultado destas e o resultado das unidades do grupo de controle que foram pareadas é calculada, e a média destas diferenças nos fornece a estimativa do ATT.

Para formalizar esta metodologia, define-se que A(i) representa as unidades dos grupos de controle que são pareados com as unidades tratadas i, com um valor estimado para o escore de propensão p(i) (BECKER e ICHINO, 2002). Depois o NMM que minimiza a diferença absoluta do escore de propensão entre as unidades i do grupo de tratamento e j do grupo controle é dado:

$$A_i(p(x)) = \left\{ p_j \mid \min \left\| p_j - p_i \right\| \right\} \qquad [1]$$

***Matching* Raio (*Radius Matching* (RM)) e *Matching* Calibrado (*Caliper Matching* (CM))**

Neste caso cada unidade tratada só será pareada com uma unidade do grupo de controle, quando este possuir um valor de escore de propensão que se encontra em uma distância pré-definida (o raio) do escore de propensão. O benefício desta técnica é que utiliza unicamente tantas unidades de controles como raio o permita, mas é possível que quanto menor seja o raio, algumas unidades tratadas não possam ser pareadas, por não encontrarem uma unidade no grupo de controle, a diferença do que ocorre no NNM. A formula nós diz que a unidade de tratamento i, se párea com a unidade de controle j, tal que:

$$A_i(p(x)) = \left\{ p_j \mid \min \left\| p_j - p_i \right\| < r \right\}. \qquad [2]$$

Onde, r>0 é um raio pré-especificado.

Para o método de *Matching* de Visinho mais próximo NMM e *Matching* de raio, RM, denota-se a comparação o número de controles pareados com as observações i que pertencem ao grupo de controle por N_C e w(i, j) denota o peso dados que o j-th se compara com o i-th caso do tratamento, $\sum_j w(i,j) = 1.$ $w_{ij} = 1/N_{ic}$ se j pertencem ao grupo controle, e w(i,j)=0 em outro caso. Então o estimador ATT para ambos *matching* é dado por:

$$\begin{aligned} \overline{\Delta} &= \frac{1}{N_T} \sum_{i \in \{D=1\}} \left[Y_{1i} - \sum_j w(i,j) Y_{0j} \right] \\ &= \frac{1}{N_T} \sum_{i \in \{D=1\}} Y_{1i} - \frac{1}{N_T} \sum_{j \in \{D=0\}} w_j Y_{oj}, \end{aligned} \qquad [3]$$

Onde 0< w(i,j) <1, o peso w_j são definidos por $w_j = \Sigma_i w_{ij}$, {D=1}é o conjunto dos individuos tratados, j é um elemento do conjunto de unidades de pares comparadas, e N_T denota o número de unidades no grupo de controle.

***Matching* de *Kernel* (*Kernel Matching* (KM))**

Para encontrar este estimador, se realiza uma média ponderada dos resultados das observações mais próximas a cada participante. Os pesos são alocados de forma

inversamente proporcional a distancia entre os escores de propensão dos grupos tratamento e controle. A média ponderada calcula-se com a seguinte formula:

$$w(i,j) = \frac{K\left(\frac{p_j - p_i}{h_h}\right)}{\sum_{j=1}^{N_{c,i}} K\left(\frac{p_j - p_i}{h_h}\right)}, \qquad [4]$$

Onde h_k é uma banda ou parâmetro de suavização, K, é a função de *Kernel* da diferença nos escores de propensão dos tratados e não tratados. Logo o estimador do *Matching* de *Kernel* será dado por:

$$\overline{\Delta}^{K} = \frac{1}{N_T} \sum_{i \in \{D=1\}} \left\{ Y_{1i} - \sum_{j \in \{D=0\}} \frac{K\left(p_j - p_i / h_h\right) Y_{0i}}{\sum_{j=1}^{N_{c,i}} K\left(p_j - p_i / h_h\right)} \right\} \qquad [5]$$

***Matching* Estratificado (*Stratification Matching* (SM))**

Este método baseia-se no mesmo procedimento de estratificação utilizado para estimar o escore de propensão. É importante destacar que para a construção, em cada bloco definido pelo este procedimento as covariâncias são balanceadas e a assinação ao tratamento pode ser considerada aleatória. Portanto, se q é o índice dos blocos definidos no intervalo do escore de propensão, dentro de cada bloco se calcula:

$$\overline{\Delta}_q^{S} = \frac{\sum_{i \in I(q)} Y_{1i}}{N_{T,q}} - \frac{\sum_{j \in I(q)} Y_{0j}}{N_{C,q}}, \qquad [6]$$

onde I(q) é o conjunto de unidades no bloco q enquanto, $N_{T,q}$ e $N_{C,q}$ representa o numero de unidades tratadas e de controle no bloco q. Logo o estimador ATT com base no método de estratificação é calculado com a seguinte formula:

$$\overline{\Delta}^{S} = \sum_{q=1}^{Q} \overline{\Delta}_q^{S} \frac{\sum_{i \in I(q)} D_i}{\sum_{\forall i} D_i} \qquad [7]$$

Onde o peso para cada bloco é dado pela correspondente fração das unidades tratadas e Q representa o número de blocos.

ANEXO IV: Programas sociais monitorada pelo Governo Federal.

i. Bolsa Escola.

Programa pioneiro no que diz respeito aos programas de transferência condicionada de renda, sendo um programa de garantia de renda mínima vinculada à educação. Este programa se transformou num dos mais amplos programas sociais do mundo quando foi criado pela Lei Nº 10.219, de 11 de Abril de 2001. Por meio desta lei, o governo federal criava um programa de transferência condicionada de renda onde as famílias recebem um benefício mensal, em dinheiro, desde que mantenham suas crianças matriculadas e freqüentando a escola. Quando este programa federal foi criado, os municípios que já tinham seus próprios programas de transferência de renda vinculados à educação puderam manter seus benefícios, aumentando o número de beneficiados ou o valor das bolsas, de acordo com suas necessidades, caso aderissem ao Bolsa Escola Federal (KASSOUF, 2004). Para ter direito ao benefício do Bolsa Escola, a família deve estar cadastrada no Cadastro Único de Programas Sociais do Governo Federal; além disso, comprovar residência no município, ter filhos ou dependentes, com idade entre seis e quinze anos, matriculados e freqüentando o ensino fundamental, e ter renda familiar mensal per capita de até R$ 90,00 (noventa reais) em 2002. O programa nacional previa um pagamento por criança e limita o número de crianças beneficiárias de uma mesma família de modo que não haja incentivo para aumento na taxa de fecundidade entre a população alvo. Segundo sua regra eram concedidas no máximo três bolsas mensais por família elegível, independentemente do número de crianças em idade escolar entre seus componentes. Atualmente O Bolsa Escola foi unificado ao PBF, assim as famílias do Bolsa Escola que cumpriam as exigências do PBF, passaram a receber o benefício do PBF, e os cadastros das famílias beneficiárias foram migrados para o Cadastro Único (BRASIL, 200-?c).

ii. Auxílio Gás

Programa criado em 2001 com o objetivo subsidiar o preço do gás liquefeito de petróleo para famílias de baixa renda. O subsídio é concedido a famílias que tenham um rendimento

per capita de até meio salário mínimo (R$90), e podem também receber benefícios de outros programas do governo federal como o Bolsa Escola e Bolsa Alimentação.

O valor do benefício em 2002 era de R$15,00 a cada dois meses e preferencialmente a mãe. O controle e fiscalização do programa ficaram sob responsabilidade do Ministério de Minas e Energia (CAIXA ECONÔMICA FEDERAL, 2002). Este programa atualmente forma parte do PBF, e os cadastros das famílias beneficiárias foram migrados para o Cadastro Único (BRASIL, 200-?c).

iii. Bolsa Alimentação

Foi criado pelo Ministério da Saúde em setembro de 2001. Programa de Renda Mínima vinculada à saúde, que consiste em melhorar as condições de saúde e nutrição de gestantes, mães que estão amamentando filhos menores de seis meses, e crianças de 6 meses a 6 anos e 11 meses. Em 2001, podiam ser atendidas pelo programa todas as famílias que possuam uma renda per capita de até R$90,00, no caso das crianças filhas de mães soropositivos para o HIV/aids poderiam receber o beneficio a partir do nascimento. Entre as condicionalidades do programas esta o compromisso das gestantes em realizar a consulta pré-natal e participar de atividades educativas que incluem orientação de alimentação durante a gestação e aleitamento materno. No caso das nutrizes, mães amamentando filhos de 0 a 6 meses, e mães com filhos de 6 meses a 6 anos e 11 meses, precisam registrar o nascimento da criança, manter a amamentação, e levá-la periodicamente para acompanhamento do crescimento e vacinação nas unidades de saúde do município. Em 2001 o programa compreenderia o pagamento do valor mensal de R$ 15,00 (quinze reais) por beneficiário, até o limite de R$ 45,00 (quarenta e cinco reais) por família beneficiada (BRASIL, 200-?a). Na atualidade, o programa Bolsa Alimentação foi também unificador ao PBF e as famílias beneficiárias deste programa, tiveram seus cadastros transferidos para o Cadastro Único.

iv. Cartão Alimentação

O Programa Nacional de Assistência Alimentar ou Cartão Alimentação - foi criado em 2003, com o objetivo de conceder um benefício às famílias em situação de insegurança alimentar. As famílias consideradas em condição de insegurança alimentar são aquelas que não têm acesso a alimentos de qualidade, em quantidade suficiente de modo permanente.

Este programa foi implantado prioritariamente em municípios da região do semi-árido brasileiro, bem como em áreas de grupos populacionais sujeitos à insegurança alimentar.

Os benefícios poderiam ser em dinheiro ou em alimentos em espécie, (os alimentos foram dados por questões culturais e hábitos alimentares, ocorrência de calamidades naturais e outras situações emergenciais, ou em caso de inexistência ou insuficiência de infra-estrutura varejista de distribuição de alimentos). Em caso do dinheiro, em 2003 o valor por mês era de R$50,00 (cinqüenta reais). Além disso, do benefício para cada pessoa ou família poderia ser até seis meses, prorrogáveis por, no máximo, mais dois períodos de seis meses, e somente concedido para pessoa ou família com renda familiar mensal per capita de até meio salário mínimo (R$ 100,00) em 2003 (KASSOUF, 2004). Os beneficiários podem participar em atividades comunitárias e educativas, inclusive aquelas de caráter temporário, e outras formas de contrapartidas sociais a serem definidas de acordo com as características do grupo familiar. O programa Cartão Alimentação hoje foi incorporado ao Bolsa Família e os beneficiários passaram a formar parte dos beneficiários do PBF (BRASIL, 200-?a).

v. Benefício de Prestação Continuada (BPC)

O programa Benefício da Prestação é um dos maiores programas de renda mínima da América Latina, Continuada (BPC) e que garante um salário mínimo mensal a idosos com 67 anos ou mais e a pessoas portadoras de deficiência incapacitadas para o trabalho e para a vida independente, seja por deficiência física, seja por deficiência mental. Em ambos os casos, a renda familiar per capita dos beneficiários deve ser inferior a 1/4 do salário mínimo.

O benefício visa proporcionar a essas pessoas uma vida independente. O programa está em vigor desde 1996. Para requerê-lo, o idoso ou a pessoa portadora de deficiência (PPD) deve se dirigir a uma agência do Instituto Nacional do Seguro Social (INSS), órgão responsável por sua operacionalização, sob coordenação e avaliação da Secretaria de Estado de Assistência Social. Como é um benefício assistencial, isto é, não exige qualquer contrapartida de quem o recebe, a própria lei que o regulamentou define a revisão das concessões a cada dois anos, garantindo o direto daqueles que realmente necessitam do benefício (BRASIL, 200-?b).

vi. Programa de Erradicação do Trabalho Infantil (PETI)

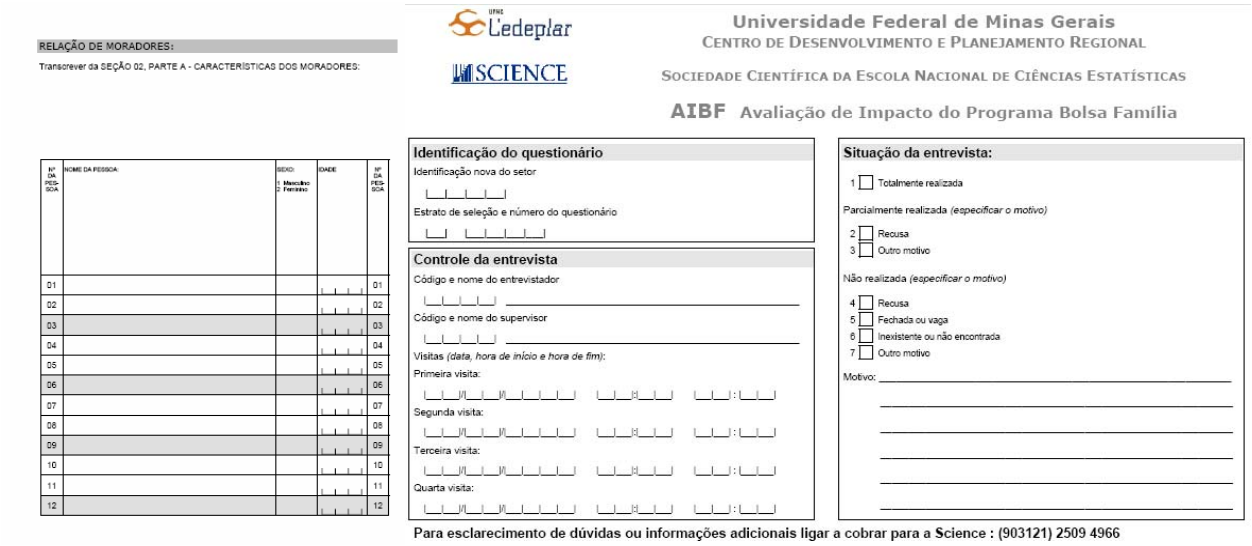
Este programa começou a ser implementado em 1999, e tem como objetivo eliminar, em parceria com os diversos setores dos governos estaduais e municipais e da sociedade civil, o trabalho infantil em atividades perigosas, insalubres e degradantes. Destina-se, prioritariamente, às famílias atingidas pela pobreza e pela exclusão social com filhos na faixa etária de 7 a 14 anos que trabalham em atividades dessa natureza. Em 2006, o valor do benefício era variável: as famílias, cujas crianças exercem atividades típicas da área urbana, tinham direito à bolsa mensal no valor de R$ 40 por criança. As que exercem atividades típicas da área rural recebiam R$ 25 ao mês, para cada criança cadastrada. Além disso, o programa destinava 20 nas áreas rurais e R$ 10 nas áreas urbanas (por criança ou adolescente) à denominada Jornada Escolar Ampliada, para o desenvolvimento, em período extracurricular, de atividades de reforço escolar, alimentação, ações esportivas, artísticas e culturais (BRASIL, 200-?d). As famílias contempladas a receber este benefício devem comprometer-se a que as crianças inscritas freqüentem no mínimo 85% das aulas no sistema formal de ensino, além de participar da Jornada Ampliada, e os pais comprometerem-se a não enviar seus filhos ao trabalho. Depois que os programas Bolsa Escola, Cartão Alimentação, Bolsa Alimentação, Auxílio-Gás, fossem unificados, a seguinte etapa para 2006 era a integração do PBF com o PETI, embora ainda esta unificação é processo que esta em andamento. (BRASIL, 200-?d).

vii. Projeto Agente Jovem de Desenvolvimento Social e Humano.

Programa criado em 2000 pelo Governo Federal com o objetivo de capacitar jovens de 15 a 17 anos para o trabalho, assim como para atuar em suas comunidades nas áreas de saúde, cultura, meio ambiente, cidadania, esporte e turismo. O público-alvo são jovens residentes em periferias urbanas, com prioridade para aqueles que estejam fora da escola, já que uma das exigências é a de que o adolescente retorne à algum tipo de atividade escolar. Também são priorizados os egressos de programas que atendem meninos e meninas em idade escolar tais como o de Erradicação do Trabalho Infantil, o Bolsa Escola e o Renda Mínima. O jovem atendido no projeto participa de curso de capacitação durante seis meses e depois começa a atuar em sua comunidade. Em 2006 durante todo o ano ele recebia uma bolsa mensal no valor de R$ 65,00. Recebem o beneficio os jovens regularmente cadastrados; e participante no mínimo, de 85% do total de aulas na escola e das atividades previstas no Programa (BRASIL, 200-?e).

ANEXO V: Questionário da coleta domiciliar da avaliação do Programa Bolsa Família (algumas seções)

4.1. Capa do questionário



RELAÇÃO DE MORADORES:

Transcrever da SEÇÃO 02, PARTE A - CARACTERÍSTICAS DOS MORADORES:

Nº DA PESSOA	NOME DA PESSOA	SEXO: 1 Masculino 2 Feminino	IDADE	Nº DA PESSOA
01				01
02				02
03				03
04				04
05				05
06				06
07				07
08				08
09				09
10				10
11				11
12				12

UFMG Cedeplar

SCIENCE

Universidade Federal de Minas Gerais
CENTRO DE DESENVOLVIMENTO E PLANEJAMENTO REGIONAL
SOCIEDADE CIENTÍFICA DA ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS
AIBF Avaliação de Impacto do Programa Bolsa Família

Identificação do questionário

Identificação nova do setor

|__|__|__|__|

Estrato de seleção e número do questionário

|__| |__|__|__|__|

Controle da entrevista

Código e nome do entrevistador

|__|__|__|__| _____

Código e nome do supervisor

|__|__|__|__| _____

Visitas *(data, hora de início e hora de fim)*:

Primeira visita:

|__|__|/|__|__|/|__|__|__|__| |__|__|h|__|__| |__|__|:|__|__|

Segunda visita:

|__|__|/|__|__|/|__|__|__|__| |__|__|h|__|__| |__|__|:|__|__|

Terceira visita:

|__|__|/|__|__|/|__|__|__|__| |__|__|h|__|__| |__|__|:|__|__|

Quarta visita:

|__|__|/|__|__|/|__|__|__|__| |__|__|h|__|__| |__|__|:|__|__|

Situação da entrevista:

1 ☐ Totalmente realizada

Parcialmente realizada *(especificar o motivo)*

2 ☐ Recusa
3 ☐ Outro motivo

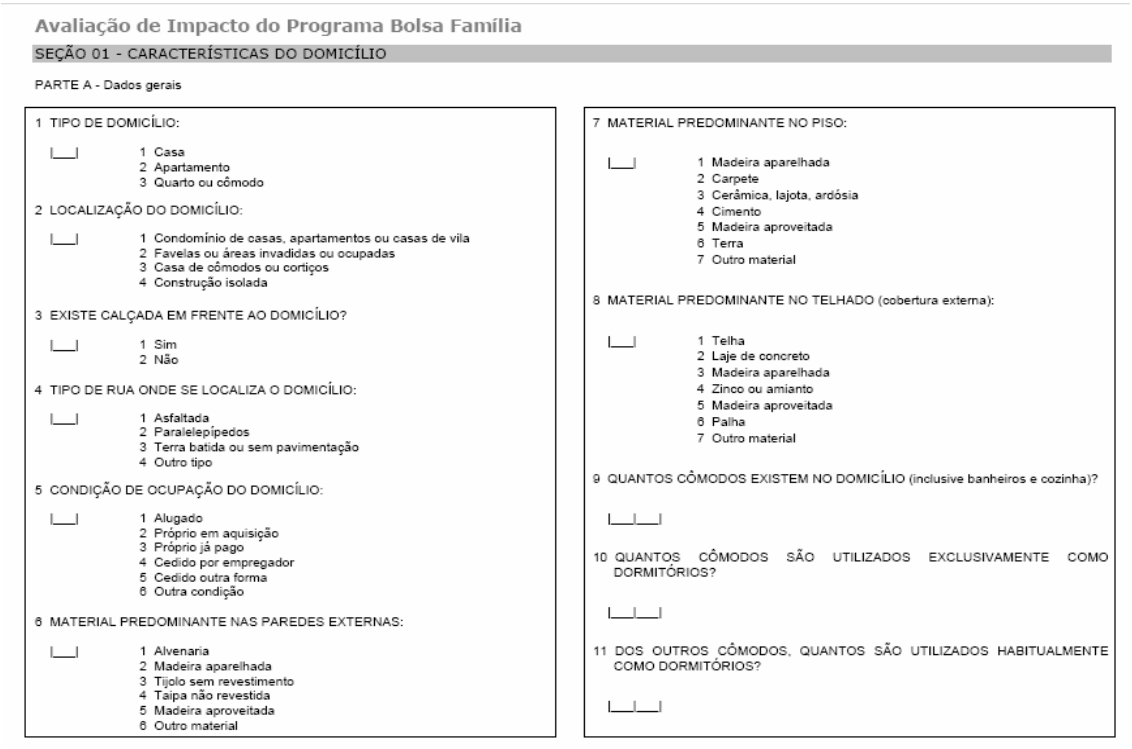
Não realizada *(especificar o motivo)*

4 ☐ Recusa
5 ☐ Fechada ou vaga
6 ☐ Inexistente ou não encontrada
7 ☐ Outro motivo

Motivo: _____

Para esclarecimento de dúvidas ou informações adicionais ligar a cobrar para a Science : (903121) 2509 4966

4.2. Seção 1: Características do domicilio.



Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 01 - CARACTERÍSTICAS DO DOMICÍLIO

PARTE A - Dados gerais

1 TIPO DE DOMICÍLIO:

|___| 1 Casa
2 Apartamento
3 Quarto ou cômodo

2 LOCALIZAÇÃO DO DOMICÍLIO:

|___| 1 Condomínio de casas, apartamentos ou casas de vila
2 Favelas ou áreas invadidas ou ocupadas
3 Casa de cômodos ou cortiços
4 Construção isolada

3 EXISTE CALÇADA EM FRENTE AO DOMICÍLIO?

|___| 1 Sim
2 Não

4 TIPO DE RUA ONDE SE LOCALIZA O DOMICÍLIO:

|___| 1 Asfaltada
2 Paralelepípedos
3 Terra batida ou sem pavimentação
4 Outro tipo

5 CONDIÇÃO DE OCUPAÇÃO DO DOMICÍLIO:

|___| 1 Alugado
2 Próprio em aquisição
3 Próprio já pago
4 Cedido por empregador
5 Cedido outra forma
6 Outra condição

6 MATERIAL PREDOMINANTE NAS PAREDES EXTERNAS:

|___| 1 Alvenaria
2 Madeira aparelhada
3 Tijolo sem revestimento
4 Taipa não revestida
5 Madeira aproveitada
6 Outro material

7 MATERIAL PREDOMINANTE NO PISO:

|___| 1 Madeira aparelhada
2 Carpete
3 Cerâmica, lajota, ardósia
4 Cimento
5 Madeira aproveitada
6 Terra
7 Outro material

8 MATERIAL PREDOMINANTE NO TELHADO (cobertura externa):

|___| 1 Telha
2 Laje de concreto
3 Madeira aparelhada
4 Zinco ou amianto
5 Madeira aproveitada
6 Palha
7 Outro material

9 QUANTOS CÔMODOS EXISTEM NO DOMICÍLIO (inclusive banheiros e cozinha)?

|___|___|

10 QUANTOS CÔMODOS SÃO UTILIZADOS EXCLUSIVAMENTE COMO DORMITÓRIOS?

|___|___|

11 DOS OUTROS CÔMODOS, QUANTOS SÃO UTILIZADOS HABITUALMENTE COMO DORMITÓRIOS?

|___|___|

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 01 - CARACTERÍSTICAS DO DOMICÍLIO

PARTE A - Dados gerais

12 QUANTOS BANHEIROS EXISTEM NESTE DOMICÍLIO (considere apenas os que têm chuveiro ou banheira e aparelho sanitário)?

|__| Se 0, siga para o quesito 13. Caso contrário, passe para o quesito 14.

13 NESTE DOMICÍLIO, TERRENO OU PROPRIEDADE EXISTE SANITÁRIO UTILIZADO PELOS MORADORES?

|__| 1 Sim
2 Não → Vá para o quesito 15

14 TIPO DE ESCOADOURO DO BANHEIRO OU SANITÁRIO:

|__| 1 Rede coletora de esgotos
2 Fossa séptica
3 Fossa rudimentar
4 Vala
5 Outro tipo
6 Não tem

15 EXISTE ÁGUA CANALIZADA DENTRO DO DOMICÍLIO?

|__| 1 Sim
2 Não

16 PRINCIPAL FONTE DE ABASTECIMENTO DE ÁGUA:

|__| 1 Rede geral
2 Poço ou nascente na propriedade — Vá para o quesito 18
3 Poço ou nascente fora da propriedade — Vá para o quesito 18
4 Bica pública — Vá para o quesito 18
5 Carro pipa — Vá para o quesito 18
6 Cisterna (água da chuva) — Vá para o quesito 18
7 Outra forma — Vá para o quesito 18

17 QUAL A FREQÜÊNCIA DE ABASTECIMENTO DE ÁGUA DA REDE GERAL?

|__| 1 Até 4 dias por semana
2 5 ou mais dias por semana

18 PRINCIPAL TIPO DE ÁGUA USADA PARA BEBER:

|__| 1 Filtrada
2 Fervida
3 Filtrada e fervida
4 Mineral
5 Natural
6 Direto da rede

19 PRINCIPAL TIPO DE ILUMINAÇÃO:

|__| 1 Elétrica (rede geral)
2 Gerador (domiciliar)
3 Lampião
4 Vela ou lamparina
5 Outro tipo

20 QUAL O PRINCIPAL TIPO DE COMBUSTÍVEL USADO PARA COZINHAR?

|__| 1 Eletricidade
2 Gás de botijão ou canalizado
3 Querosene
4 Carvão ou lenha
5 Outro tipo
6 Não utiliza

21 EXISTE TELEFONE FIXO NO DOMICÍLIO?

|__| 1 Sim
2 Não

22 PRINCIPAL DESTINO DO LIXO DOMICILIAR:

|__| 1 Coletado diretamente por serviço de limpeza
2 Coletado indiretamente
3 Queimado ou enterrado
4 Jogado em terreno baldio ou logradouro
5 Jogado em rio, lago, ou no mar
6 Outro destino

4.3. Seção 03: Educação (dados gerais).

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 03 - EDUCAÇÃO

Parte A – Dados gerais

1 NÚMERO DE ORDEM DA PESSOA QUE PRESTOU AS INFORMAÇÕES PARA ESTA SEÇÃO:

|__|__|

Nº DA PESSOA	2 [NOME] SABE LER E ESCREVER UM BILHETE SIMPLES NO IDIOMA QUE CONHECE? 1 Sim 2 Não	3 NO ANO PASSADO, FREQÜENTAVA ESCOLA OU CRECHE? 1 Sim, rede particular 2 Sim, rede pública municipal 3 Sim, rede pública estadual 4 Sim, rede pública federal 5 Não Vá para 6	4 QUAL É O CURSO QUE FREQÜENTAVA NO ANO PASSADO? 01 Creche 02 Pré-escolar 03 Classe de alfabetização 04 Alfabetização de adultos/AJA 05 Ensino fundamental ou 1º grau - regular seriado 06 Ensino fundamental ou 1º grau - regular não-seriado 07 Supletivo/EJA(ensino fundamental ou 1º grau) 08 Ensino médio ou 2º grau - regular seriado 09 Ensino médio ou 2º grau - regular não-seriado 10 Supletivo/EJA (ensino médio ou 2º grau) 11 Pré-vestibular 12 Superior - graduação 13 Mestrado ou doutorado AJA - Alfabetização de Jovens Adultos EJA - Educação de Jovens Adultos	5 QUAL É A SÉRIE QUE FREQÜENTAVA NO ANO PASSADO? 01 Primeira 02 Segunda 03 Terceira 04 Quarta 05 Quinta 06 Sexta 07 Sétima 08 Oitava 09 Curso não-seriado 10 Nenhuma	6 FREQÜENTA ESCOLA OU CRECHE ATUALMENTE? 1 Sim, rede particular Vá para 8 2 Sim, rede pública municipal Vá para 8 3 Sim, rede pública estadual Vá para 8 4 Sim, rede pública federal Vá para 8 5 Não, já freqüentou 6 Não, nunca freqüentou Vá para 29	7 POR QUE NÃO FREQÜENTA? 01 Precisa trabalhar 02 Precisa ajudar palmilhe no trabalho 03 Precisa tomar conta de irmãos 04 Não tem escola perto 05 É caro 06 Já formou 07 Engravidou/ casou 08 Doença 09 Não gosta de estudar 10 Outro motivo	8 QUANTOS ANOS TINHA QUANDO DEIXOU DE FREQÜENTAR ESCOLA? Se não sabe, preencher com: 00 Se responder a este quesito, vá para o quesito 26.	9 QUAL É O CURSO QUE FREQÜENTA ATUALMENTE? 01 Creche 02 Pré-escolar 03 Classe de alfabetização 04 Alfabetização de adultos/AJA 05 Ensino fundamental ou 1º grau - regular seriado 06 Ensino fundamental ou 1º grau - regular não-seriado 07 Supletivo/EJA(ensino fundamental ou 1º grau) 08 Ensino médio ou 2º grau - regular seriado 09 Ensino médio ou 2º grau - regular não-seriado 10 Supletivo/EJA (ensino médio ou 2º grau) 11 Pré-vestibular 12 Superior - graduação 13 Mestrado ou doutorado	10 QUAL É A SÉRIE QUE FREQÜENTA ATUALMENTE? 01 Primeira 02 Segunda 03 Terceira 04 Quarta 05 Quinta 06 Sexta 07 Sétima 08 Oitava 09 Curso não-seriado 10 Nenhuma	Nº DA PESSOA
01			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	01
02			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	02
03			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	03
04			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	04
05			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	05
06			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	06
07			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	07
08			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	08
09			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	09
10			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	10
11			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	11
12			|__|__|	|__|__|		|__|__|	|__|__|	|__|__|	|__|__|	12

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 03 - EDUCAÇÃO

Parte A – Dados gerais

Nº DA PESSOA	SOMENTE PARA QUEM FREQÜENTA ESCOLA (EXCETO ALUNOS DE AJA)			PARA QUEM FREQÜENTA ESCOLA							Nº DA PESSOA
	11 TURNO QUE FREQÜENTA: 1 Manhã 2 Tarde 3 Noite 4 Manhã e tarde 5 Manhã e noite 6 Tarde e noite	12 NOME DA ESCOLA OU CRECHE QUE FREQÜENTA:	13 CÓDIGO DA ESCOLA/CRECHE (Este código será preenchido após a coleta)	14 [NOME] FAZ ALGUMA REFEIÇÃO GRATUITA NA ESCOLA? 1 Sim 2 Não *Vá para 21*	15 COM QUE FREQÜÊNCIA [NOME] FAZ REFEIÇÕES GRATUITAS NA ESCOLA? 1 Um dia por semana 2 Dois ou três dias por semana 3 Quatro ou cinco dias por semana 4 Mais de cinco dias por semana	16 QUAL É O TIPO DA PRINCIPAL REFEIÇÃO GRATUITA QUE [NOME] FAZ NA ESCOLA? 1 Café da manhã/lanche 2 Almoço/ merenda 3 Jantar/ merenda	17 PARA ESTA REFEIÇÃO QUAL ERA A COMPOSIÇÃO PRINCIPAL? 1 Fruta / suco, 2 Pão com manteiga e café com leite 3 Arroz, feijão, carne, legumes 4 Sopa, mingau, munguzá, canjica, etc. 5 Outro	18 QUAL É O TIPO DA SEGUNDA PRINCIPAL REFEIÇÃO GRATUITA QUE [NOME] FAZ NA ESCOLA? 1 Café da manhã 2 Almoço 3 Lanche/merenda 4 Jantar 5 Só faz uma refeição **Vá para 21**	19 PARA ESTA REFEIÇÃO QUAL ERA A COMPOSIÇÃO PRINCIPAL? 1 Fruta / suco, 2 Pão com manteiga e café com leite 3 Arroz, feijão, carne, legumes 4 Sopa, mingau, munguzá, canjica, etc. 5 Outro	20 FAZ MAIS QUE DUAS REFEIÇÕES DIÁRIAS NA ESCOLA? 1 Sim 2 Não	
01			_____								01
02			_____								02
03			_____								03
04			_____								04
05			_____								05
06			_____								06
07			_____								07
08			_____								08
09			_____								09
10			_____								10
11			_____								11
12			_____								12

4.4. Seção 12: Benefícios.

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 12 - BENEFÍCIOS

PARTE A - Bolsa Família

1 VOCÊ OU ALGUM OUTRO MORADOR DESTE DOMICÍLIO JÁ SE INSCREVEU OU FOI CADASTRADO PARA RECEBER BENEFÍCIO DE ALGUM PROGRAMA DO GOVERNO FEDERAL?

|___| 1 Sim 2 Não *Vá para a Parte B*

2 EM QUE MÊS E ANO VOCÊ OU SUA FAMÍLIA SE INSCREVEU OU FOI CADASTRADO PARA RECEBER BENEFÍCIO DE ALGUM PROGRAMA DO GOVERNO FEDERAL?

|__|__|/|__|__|__|__| PREENCHER 88/8888 PARA NÃO SABE

3 ESTA INSCRIÇÃO/CADASTRAMENTO FOI PARA RECEBER O BENEFÍCIO DO PROGRAMA BOLSA FAMÍLIA?

|___| 1 Sim 2 Não *Vá para a Parte B*

4 COMO, OU POR MEIO DE QUEM, O SR/SRA TEVE CONHECIMENTO DO PROGRAMA DO BOLSA FAMÍLIA?

|___| 1 Prefeitura 2 Parentes 3 Vizinhos 4 Amigos 5 Televisão/Rádio/Jornal 6 Escola/Creche 7 Posto de saúde 8 Outro

5 ONDE FOI PREENCHIDO E QUEM PREENCHEU O FORMULÁRIO DE CADASTRAMENTO?

|___| 1 Funcionário/técnico da prefeitura em algum órgão municipal, escola, posto de saúde
2 Funcionário/técnico da prefeitura no domicílio da família
3 Alguém da própria família em algum órgão municipal, escola, posto de saúde
4 Alguém da própria família no domicílio da família
5 Outra pessoa que não era da família nem funcionário/técnico da prefeitura (parente em outro domicílio, amigo, vizinho, líder comunitário)
6 Outra situação

6 TEVE QUE MOSTRAR ALGUM DOCUMENTO PARA SE CADASTRAR?
(Admite múltiplas respostas)

1 ☐ Não
2 ☐ Título de eleitor
3 ☐ CPF
4 ☐ Carteira de identidade
5 ☐ Carteira de trabalho
6 ☐ Comprovante de residência
7 ☐ Outro documento

7 VOCÊ, OU ALGUM OUTRO MORADOR DESTE DOMICÍLIO, RECEBE OU JÁ RECEBEU O BENEFÍCIO BOLSA FAMÍLIA?

|___| 1 Sim 2 Não *Vá para a Parte B*

8 QUANTO TEMPO LEVOU ENTRE O PREENCHIMENTO DO FORMULÁRIO E A ENTREGA DO CARTÃO PARA SAQUE/RECEBIMENTO DO BENEFÍCIO DO BOLSA FAMÍLIA?

|__|__| meses

9 ONDE PEGOU OU QUEM LHE ENTREGOU O CARTÃO PARA SAQUE/ RECEBIMENTO DO BENEFÍCIO DO BOLSA FAMÍLIA?

|___| 1 Prefeitura
2 Agência da Caixa Econômica Federal (CEF)
3 Casa lotérica/Correspondente bancário
4 Parente, amigo, vizinho
5 Outro. Qual? _____

10 ONDE RECEBE OU RECEBIA O BENEFÍCIO DO BOLSA FAMÍLIA?

|___| 1 Agência da Caixa Econômica Federal (CEF)
2 Casa lotérica/Correspondente bancário
3 Através de parente, amigo, vizinho
4 Outro. Qual? _____

11 O SR/SRA JÁ TEVE ALGUM DIFICULDADE/PROBLEMA PARA RECEBER O BENEFÍCIO DO BOLSA FAMÍLIA?

|___| 1 Sim 2 Não

12 A QUEM O SR / A SRA RECORREU OU RECORRERIA EM CASO DE DIFICULDADE/ PROBLEMA PARA RECEBER O BENEFÍCIO DO BOLSA FAMÍLIA?

|___| 1 Prefeitura 2 Atendimento 0800 3 Agência da Caixa (CEF) 4 Ministério Público 5 Líder comunitário 6 Político 7 Parente, amigo, vizinho 8 Outro. Qual? _____

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 12 - BENEFÍCIOS

PARTE B - Para cada morador do domicílio - Informar os benefícios que recebe ou já recebeu

1 ALGUM MORADOR DESTE DOMICÍLIO RECEBE OU RECEBEU ALGUM DOS SEGUINTES BENEFÍCIOS [*LISTA DE BENEFÍCIOS*]?

|__| 1 Sim
2 Não Encerre a entrevista

BF - Bolsa família
BA - Bolsa alimentação
CA - Cartão alimentação
BE - Bolsa escola
VG - Vale gás
BPC idoso - Benefício de prestação continuada para idoso
BPC PPD - Benefício de prestação continuada para pessoa portadora de deficiência (física ou mental)
RMV - Renda mensal vitalícia
PETI - Programa de erradicação do trabalho infantil
Agente jovem
Benefício recebido de igreja
Benefício recebido de ONG (Organização Não Governamental)
Benefício recebido de sindicato
Bolsa escola municipal
Outro benefício

MARCAR COM **X** O BENEFÍCIO QUE CADA MORADOR RECEBE

	BF	BA	CA	BE	VG	BPC idoso	BPC PPD	RMV	PETI	Agente jovem	Igreja	ONG	Sindi-cato	Bolsa Escola Muni-cipal	OUTRO (Especificar)	
01	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	01
02	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	02
03	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	03
04	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	04
05	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	05
06	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	06
07	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	07
08	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	08
09	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	09
10	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	10
11	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	11
12	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	0 ☐	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	12

Avaliação de Impacto do Programa Bolsa Família

SEÇÃO 12 - BENEFÍCIOS - QUESTIONÁRIO POR BENEFÍCIO E MORADOR

PARTE C - Informação de cada morador do domicílio e cada benefício recebido

|__|__|__|__| SETOR

|__| |__|__|__|__| ESTRATO DE SELEÇÃO E NÚMERO DO QUESTIONÁRIO

1 NÚMERO DE ORDEM DA PESSOA:

|__|__|

2 ESTA PESSOA TEM NIS?

|__| 1 Sim
2 Não Vá para o quesito 4

3 NIS DA PESSOA:

|__|__| |__|__|__|__| |__|__|__|__| |__|__|__|__| - |__|__|

4 BENEFÍCIO:

|__|__| 01 BF - Bolsa família
02 BA - Bolsa alimentação
03 CA - Cartão alimentação
04 BE - Bolsa escola
05 VG - Vale gás
06 BPC idoso - Benefício de prestação continuada para idosos
07 BPC PPD - Benefício de prestação continuada para pessoa portadora de deficiência (física ou mental)
08 RMV - Renda mensal vitalícia
09 PETI - Programa de erradicação do trabalho infantil
10 Agente jovem
11 Benefício recebido de igreja
12 Benefício recebido de ONG (Organização Não Governamental)
13 Benefício recebido de sindicato
14 Bolsa Escola Municipal
15 Outro benefício _____

5 EM QUE MÊS E ANO COMEÇOU A RECEBER ESTE BENEFÍCIO?

|__|__| / |__|__|__|__|

6 AINDA RECEBE ESTE BENEFÍCIO?

|__| 1 Sim Vá para o quesito 10
2 Não

7 EM QUE MÊS E ANO PAROU DE RECEBER ESTE BENEFÍCIO?

|__|__| / |__|__|__|__|

8 QUAL O VALOR DO ÚLTIMO BENEFÍCIO RECEBIDO?

|__|__|__|,|__|__|

9 POR QUE PAROU DE RECEBER ESTE BENEFÍCIO?

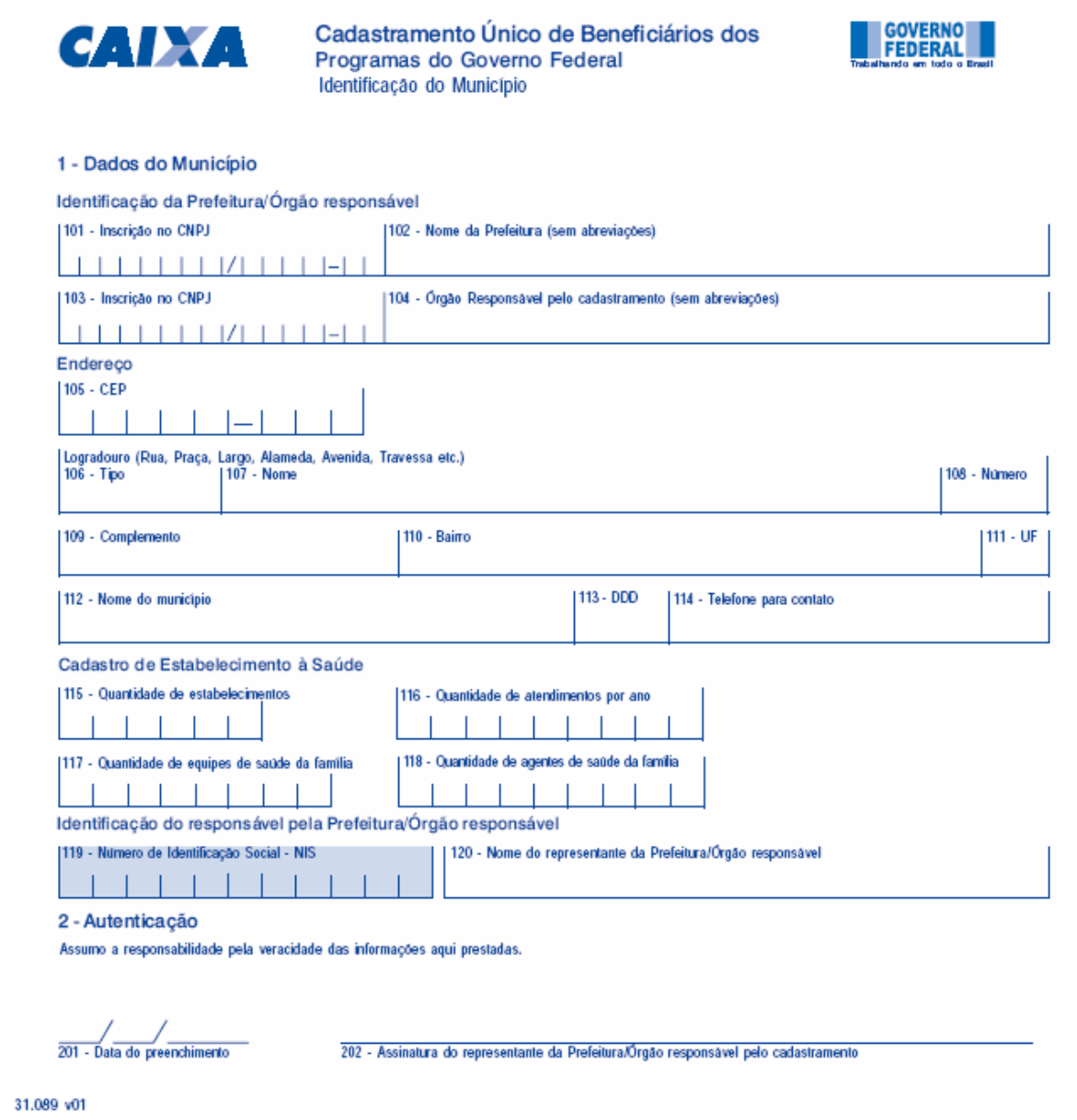
|__| 1 Não cumpriu condicionalidades (agenda de saúde e freqüência escolar)
2 Aumento da renda familiar (per capita)
3 Crianças/adolescentes completaram a idade limite
4 Prefeitura cancelou o benefício
5 Recebia mais de um benefício
6 Não precisava mais do benefício
7 Não sabe o motivo

ENCERRE A PARTE

10 QUAL O VALOR DOS 12 ÚLTIMOS RECEBIMENTOS?

|__|__|__|,|__|__| outubro 2005
|__|__|__|,|__|__| setembro 2005
|__|__|__|,|__|__| agosto 2005
|__|__|__|,|__|__| julho 2005
|__|__|__|,|__|__| junho 2005
|__|__|__|,|__|__| maio 2005
|__|__|__|,|__|__| abril 2005
|__|__|__|,|__|__| março 2005
|__|__|__|,|__|__| fevereiro 2005
|__|__|__|,|__|__| janeiro 2005
|__|__|__|,|__|__| dezembro 2004
|__|__|__|,|__|__| novembro 2004

ANEXO VI: Questionário do cadastro único domicílios e pessoas



CAIXA

Cadastramento Único de Beneficiários dos Programas do Governo Federal
Identificação do Município

GOVERNO FEDERAL
Trabalhando em todo o Brasil

1 - Dados do Município

Identificação da Prefeitura/Órgão responsável

101 - Inscrição no CNPJ

102 - Nome da Prefeitura (sem abreviações)

103 - Inscrição no CNPJ

104 - Órgão Responsável pelo cadastramento (sem abreviações)

Endereço

105 - CEP

Logradouro (Rua, Praça, Largo, Alameda, Avenida, Travessa etc.)
106 - Tipo
107 - Nome
108 - Número

109 - Complemento
110 - Bairro
111 - UF

112 - Nome do município
113 - DDD
114 - Telefone para contato

Cadastro de Estabelecimento à Saúde

115 - Quantidade de estabelecimentos
116 - Quantidade de atendimentos por ano
117 - Quantidade de equipes de saúde da família
118 - Quantidade de agentes de saúde da família

Identificação do responsável pela Prefeitura/Órgão responsável

119 - Número de Identificação Social - NIS
120 - Nome do representante da Prefeitura/Órgão responsável

2 - Autenticação

Assumo a responsabilidade pela veracidade das informações aqui prestadas.

201 - Data do preenchimento
202 - Assinatura do representante da Prefeitura/Órgão responsável pelo cadastramento

31.089 v01

CAIXA

Cadastramento Único para Programas Sociais do Governo Federal

Identificação do Domicílio e da Família

GOVERNO FEDERAL

1 - Dados de controle

101 - Tipo (de): 0 1 | 102 - Nº de ordem: 0 0 | 103 - Data da pesquisa | 104 - Número de identificação Social - NIS do entrevistador | Código Domiciliar: 018155693

105 - Nome do entrevistador | 106 - CNPJ da Prefeitura/Órgão/Empresa | 107 - Modalidade: 1 - Inclusão 2 - Alteração

Atenção básica à saúde

108 - Nome do estabelecimento de assistência à saúde | 109 - Código EAS/MS

2 - Identificação do domicílio e da família

Endereço

201 - CEP

Logradouro (Rua, Praça, Largo, Alameda, Avenida, Travessa etc.)
202 - Tipo | 203 - Nome | 204 - Número

205 - Complemento | 206 - Bairro | 207 - UF

208 - Nome do município | 209 - DDD | 210 - Telefone para contato

Características do domicílio

211 - Tipo de localidade: 1 - Urbana 2 - Rural

212 - Domicílio coberto por: 1 - PACS - Programa de Agentes Comunitários de Saúde 2 - PSF - Programa de Saúde da Família 3 - Similares ao PSF 4 - Outro

213 - Situação: 1 - Próprio 2 - Alugado 3 - Amansado 4 - Cedido 5 - Invasão 6 - Financiado 7 - Outra

214 - Tipo: 1 - Casa 2 - Apartamento 3 - Cômodos 4 - Outro

215 - Número de cômodos

216 - Tipo de construção: 1 - Tijolo/Alvenaria 2 - Adobe 3 - Taipa revestida 4 - Taipa não revestida 5 - Madeira 6 - Material aproveitado 7 - Outro

217 - Tipo de abastecimento de água: 1 - Rede pública 2 - Poço/Nascente 3 - Carro pipa 4 - Outro

218 - Tratamento de água: 1 - Filtração 2 - Fervura 3 - Cloração 4 - Sem tratamento 5 - Outro

219 - Tipo de iluminação: 1 - Relógio próprio 2 - Sem relógio 3 - Relógio comunitário 4 - Lampião 5 - Vela 6 - Outro

220 - Escoamento sanitário: 1 - Rede pública 2 - Fossa rudimentar 3 - Fossa séptica 4 - Vala 5 - Céu aberto 6 - Outro

1 - Coletado 2 - Queimado 3 - Enterrado 4 - Céu aberto 5 - Outro

222 - Quantidade de pessoas | 223 - Quantidade de mulheres grávidas | 224 - Quantidade de mães amamentando | 225 - Quantidade de deficientes

Lista de pessoas residentes no domicílio

Nº de ordem	Nome	Nº de ordem	Nome
01		07	
02		08	
03		09	
04		10	
05		11	
06		12	

3 - Autenticação

Assumo a responsabilidade pela veracidade das informações aqui prestadas.

301 - Assinatura do entrevistado

302 - Assinatura do entrevistador

303 - Assinatura do representante da Prefeitura/Órgão responsável pelo cadastramento

31.090 v02

31.089 v01

CAIXA

Cadastramento Único para Programas Sociais do Governo Federal
Identificação da Pessoa

GOVERNO FEDERAL
Trabalhando em todo o Brasil

1 - Dados de controle

101 - Tipo Doc. 0 2

102 - Número de ordem da pessoa

103 - Modalidade
1 - Inclusão
2 - Alteração

Código domiciliar

2 - Identificação da pessoa

201 - Nome completo da pessoa sem abreviações. Caso necessário abreviar, vide instruções.

202 - Data de nascimento

203 - Sexo
1 - Masculino
2 - Feminino

204 - Nacionalidade
1 - Brasileira
2 - Brasileiro naturalizado
3- Estrangeira

205 - País de origem (se estrangeiro)

206 - Data de chegada ao Brasil

207 - Cód. IBGE munic.nasc.

208 - UF munic. nasc.

209 - Nome do município de nascimento

210 - Nome completo do pai (sem abreviações)

211 - Nome completo da mãe (sem abreviações)

212 - Estado civil
1 - Solteiro(a)
2 - Casado(a)
3 - Divorciado(a)
4 - Separado(a)
5 - Viúvo(a)

213 - Se o(a) esposo(a) ou o(a) companheiro(a) reside no domicílio, informar o nº de ordem correspondente, se não reside, informar 99

214 - Tipo de deficiência (assinalar com "X")
☐ Cegueira ☐ Mudez ☐ Surdez ☐ Mental ☐ Física ☐ Nenhuma ☐ Outro

215 - Raça/Cor
1 - Branca
2 - Negra
3 - Parda
4 - Amarela
5 - Indígena

Documentos

216 - Número de Identificação Social - NIS

217 - Certidão civil tipo
91 - Nascimento
92 - Casamento

218 - Número termo

219 - Livro

220 - Folha

221 - Data de emissão

222 - UF

223 - Nome do cartório (órgão emissor)

Documento de identidade
224 - Número

225 - Complemento

226 - Data de emissão

227 - UF

228 - Sigla do órgão emissor

Carteira de Trabalho e Previdência Social
229 - Número

230 - Série

231 - Data de emissão

232 - UF

233 - CPF

Título de eleitor
234 - Número

235 - Zona

236 - Seção

31.088 v03

Qualificação escolar

237 - Frequenta escola
1 - Pública municipal
2 - Pública estadual
3 - Pública federal
4 - Particular
5 - Outra
6 - Não freqüenta

238 - Grau de instrução
1 - Analfabeto
2 - Até 4ª série incompleta do ensino fundamental
3 - Com 4ª série completa do ensino fundamental
4 - De 5ª a 8ª série incompleta do ensino fundamental
5 - Ensino fundamental completo
6 - Ensino médio incompleto
7 - Ensino médio completo
8 - Superior incompleto
9 - Superior completo
10 - Especialização
11 - Mestrado
12 - Doutorado

239 - Série escolar
1 - Maternal I
2 - Maternal II
3 - Maternal III
4 - Jardim I
5 - Jardim II
6 - Jardim III
7 - CA (alfabetização)
8 - 1ª série do ensino fundamental
9 - 2ª série do ensino fundamental
10 - 3ª série do ensino fundamental
11 - 4ª série do ensino fundamental
12 - 5ª série do ensino fundamental
13 - 6ª série do ensino fundamental
14 - 7ª série do ensino fundamental
15 - 8ª série do ensino fundamental
16 - 1ª série do ensino médio
17 - 2ª série do ensino médio
18 - 3ª série do ensino médio

240 - Nome da Escola

241 - Código censo INEP

Qualificação profissional

242 - Situação no mercado de trabalho
1 - Empregador
2 - Assalariado com carteira de trabalho
3 - Assalariado sem carteira de trabalho
4 - Autônomo com previdência social
5 - Autônomo sem previdência social
6 - Aposentado/Pensionista
7 - Trabalhador rural
8 - Empregador rural
9 - Não trabalha
10 - Outra

243 - Nome da empresa em que trabalha, se desempregado, último emprego

244 - CNPJ/CEI da empresa

245 - Data de admissão

246 - Ocupação

247 - Remuneração deste emprego

248 - Renda de aposentadoria/pensão
R$

249 - Renda de Seguro-Desemprego
R$

250 - Renda de pensão alimentícia
R$

251 - Outras rendas
R$

Características da família

252 - Tempo de moradia
(Ano(s)) (Mês(es))

Despesas mensais da família (preencher somente para a mãe/responsável legal da família)

253 - Aluguel
R$

254 - Prestação habitacional
R$

255 - Alimentação
R$

256 - Água
R$

257 - Luz
R$

258 - Transporte
R$

259 - Medicamentos
R$

260 - Gás
R$

261 - Outras despesas
R$

262 - Número de pessoas que vivem da renda desta família

Relação familiar

263 - Nº de ordem da mãe/responsável legal da família.

264 - Parentesco em relação a mãe/responsável legal da família, se o próprio, informar 01

265 - Se reside com o pai informar o número de ordem do pai, se não, informar 99.

Parentescos
01 - Mãe/responsável legal
02 - Esposo(a)
03 - Companheiro(a)
04 - Filho(a)
05 - Pai
06 - Avô/Avó
07 - Irmão/Irmã
08 - Cunhado(a)
09 - Genro/Nora
10 - Sobrinho(a)
11 - Primo(a)
12 - Sogro(a)
13 - Neto(a)
14 - Tio(a)
15 - Adotivo(a)
16 - Padrasto/Madrasta
17 - Enteado(a)
18 - Bisneto(a)
19 - Sem parentesco
20 - Outro

266 - Se reside com a mãe informar o número de ordem da mãe, se não, informar 99.

267 - Se criança de 0 a 6 anos, com quem fica?
1 - Pai/Mãe
2 - Irmão/Irmã
3 - Avó/Avó
4 - Sozinho
5 - Creche
6 - Outro

268 - Se grávida, informar mês de gestação

269 - Amamentando
1 - Sim
2 - Não

270 - Participa de algum programa do Governo Federal ou recebe algum benefício social? (assinalar com "X")

☐ Bolsa Criança Cidadã - PETI
Data de inclusão ____/____/_____

☐ Agente Jovem
Data de inclusão ____/____/_____

☐ Bolsa Escola

☐ Bolsa Alimentação

☐ Nenhum

Tipo de benefício 1 - Rural 2 - Urbano

☐ LOAS/BPC

☐ Previdência Rural

☐ PRONAF

☐ PROGER

Valor do benefício - R$ _____

☐ Outro _____

Data de início de participação ____/____/_____

271 - Beneficiário prioritário para o Programa Bolsa Alimentação
1 - Sim
2 - Não

CAIXA

Cadastramento Único de Beneficiários dos Programas do Governo Federal

Identificação do Agricultor Familiar

GOVERNO FEDERAL
Trabalhando em todo o Brasil

1 - Dados de controle

101 - Tipo Doc: 0 3

102 - Número de ordem da pessoa Agricultora:

103 - Modalidade: 1 - Inclusão; 2 - Alteração

104 - Número de Identificação Social - NIS:

Código domiciliar:

2 - Identificação do beneficiário(a)

201 - Nome completo do(a) agricultor(a) (sem abreviações)

202 - Apelido do(a) agricultor(a) (sem abreviações)

203 - Número de ordem da mãeresponsável legal da família

Organização Social

204 - Organização social a que pertence

☐ Sindicato ☐ Cooperativa ☐ Associação ☐ Quilombos ☐ Nenhuma ☐ Outra _____

3 - Estrutura da atividade agropecuária

301 - Localização do domicílio
Reside em: 1 - Estabelecimento rural; 2 - Aglomerado rural próximo; 3 - Aglomerado urbano próximo

302 - Condição de posse e uso da terra

☐ Proprietário(a) ☐ Parceiro(a)/Meeiro(a) ☐ Assentado(a) pelo INCRA ☐ Posseiro(a)
☐ Arrendatário(a) ☐ Comodatário(a) ☐ Beneficiário(a) do Banco da Terra ☐ Não se aplica

303 - Caracterização da atividade

☐ Agricultor(a) ☐ Pescador(a) artesanal ☐ Aqüicultor(a) ☐ Extrativista vegetal ☐ Silvicultor(a) ☐ Outra _____

304 - Área do estabelecimento (em hectares)

4 - Força de trabalho além da familiar

401 - Contrata empregados(as) eventuais: 1 - Sim; 2 - Não

402 - Número de empregados(as) permanentes contratados(as)

403 - Administração do estabelecimento: 1 - Pela família; 2 - Por administrador(a) remunerado(a); 3 - Por outro(a)

5 - Composição da renda bruta familiar anual

Ano agrícola

501 - Ano agrícola (mês/ano)

de ____/____ até ____/____

Perdas na produção

502 - Teve perdas na produção agropecuária neste ano agrícola: 1 - Sim; 2 - Não

Quanto (em percentual) _____ (%)

503 - Renda bruta das atividades agropecuárias

Renda bruta proveniente de avicultura, bovinocultura de leite, caprinocultura, ovinocultura, suinocultura, sericicultura, fruticultura e/ou olericultura

R$

504 - Renda bruta de outras atividades agropecuárias

Renda bruta proveniente de outras atividades agropecuárias

R$

505 - Renda bruta de atividade não agropecuária

Renda bruta de atividade não agropecuária, excluídos os proventos de benefícios previdenciários

R$

6 - Declaração do(a) beneficiário(a)

Declaro, sob as penas da lei (art. 299 do Código Penal), que as informações acima correspondem à verdade.

_____, ____/____/_____
Local e data

_____
Assinatura do(a) beneficiário(a)

31.097 v01

APÊNDICE I.

TABELA A1. 1 – Número de registros iniciais para o relacionamento probabilístico e os pares formados. Brasil. 2006. Etapa 2.

Região	Pesquisa AIBF (A)	CadÚnico (B)	AxB	Pares formados segundo a Blocagem
Norte	5.148	1.378.693	7.097.511.564	87.882
Nordeste	7.974	4.214.553	33.606.845.622	1.068.793
Centro este	2.948	1.000.960	2.950.830.080	61.638
Sudeste	9.301	3.129.043	29.103.228.943	7.876.103
Sul	1.409	596.486	840.438.911	98.096
Total	26.780	10.319.735	73.598.855.120	9.192.512

Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

TABELA A1. 2 – Número de registros iniciais para o relacionamento probabilístico e os pares formados. Brasil. 2006. Etapa 3

Região	Pesquisa AIBF (A)	CadÚnico (B)	Pares formados segundo a Blocagem	Pares formados segundo a Blocagem
Norte	1.054	1.378.722	1.453.172.988	438.962
Nordeste	3.034	4.215.192	12.788.892.528	12.980.698
Centro este	1.688	1.001.491	1.690.516.808	457.532
Sudeste	2.504	3.129.651	7.836.646.104	1.307.487
Sul	574	596.692	342.501.208	21.135
Total	8.854	10.321.748	24.111.729.636	15.205.814

Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

TABELA A1. 3 – Número de registros iniciais para o relacionamento probabilístico e os pares formados. Brasil. 2006. Etapa 4

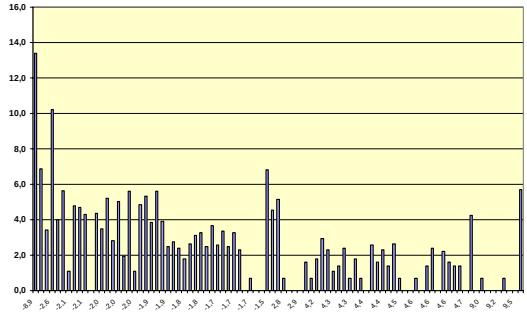
Região	Pesquisa AIBF (A)	CadÚnico (B)	AxB	Pares formados segundo a Blocagem
Norte	4.649	1.378.185	6.407.182.065	1.899.389
Nordeste	7.033	4.213.585	29.634.143.305	30.580.091
Centro este	2.685	1.000.697	2.686.871.445	681.066
Sudeste	8.112	3129043	25.382.796.816	5.905.615
Sul	1.112	596.479	663.292.432	53.406
Total	23.591	10.317.989	64.774.286.063	39.119.567

Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

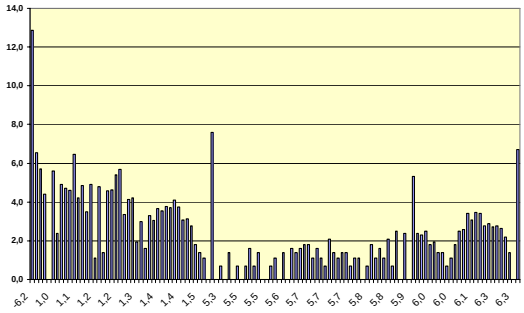
APÊNDICE II:

GRAFICO A2. 1 – Distribuição de freqüência dos pesos totais do relacionamento. Probabilístico. Regiões. Brasil 2006. Etapa 1

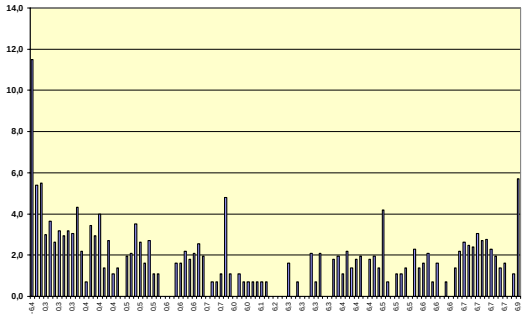
NORTE



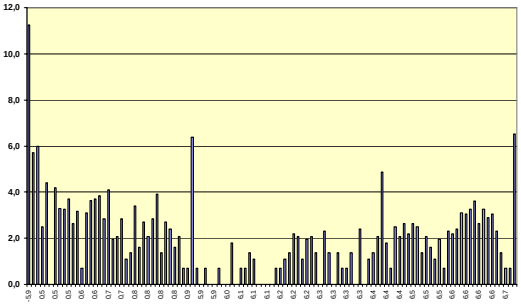
NORDESTE



CENTRO-OESTE



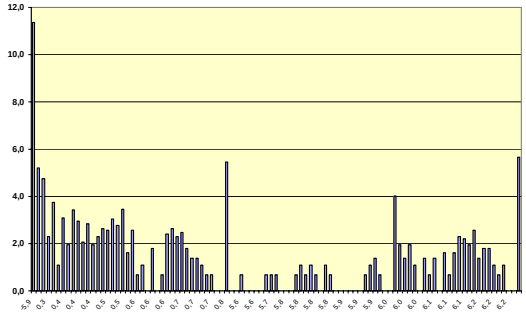
SUDESTE



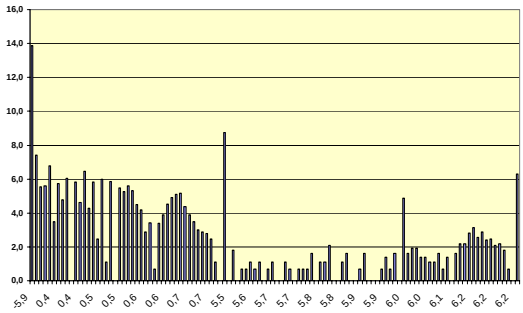
Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

GRAFICO A2. 2 – Distribuição de freqüência dos pesos totais do relacionamento. Probabilístico. Regiões. Brasil 2006. Etapa 2

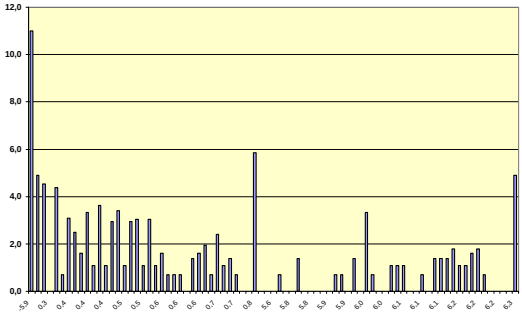
NORTE



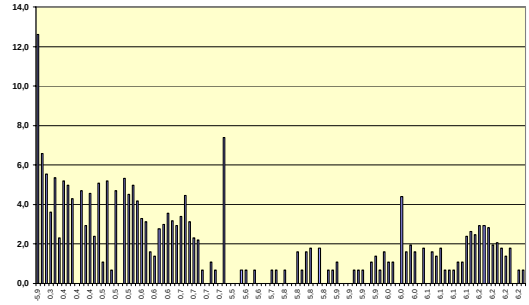
NORDESTE



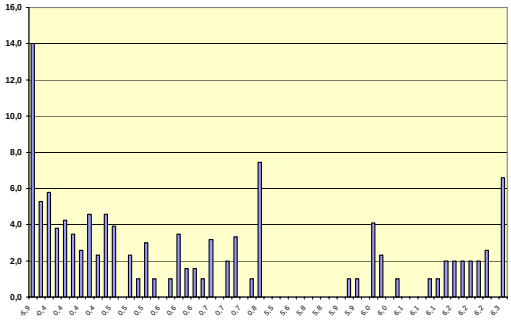
CENTRO-OESTE



SUDESTE



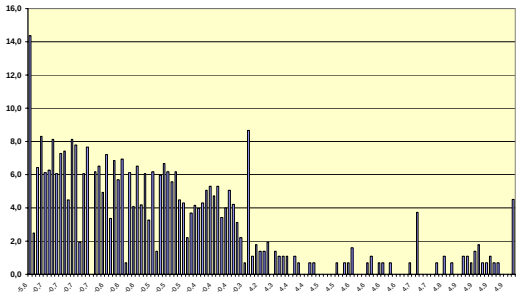
SUL



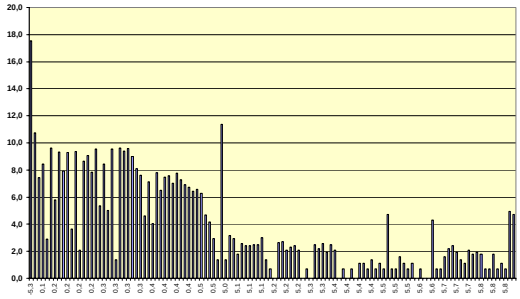
Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

GRAFICO A2. 3 – Distribuição de freqüência dos pesos totais do relacionamento. Probabilístico. Regiões. Brasil 2006. Etapa 3

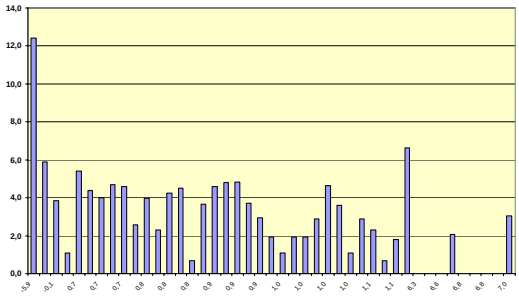
NORTE



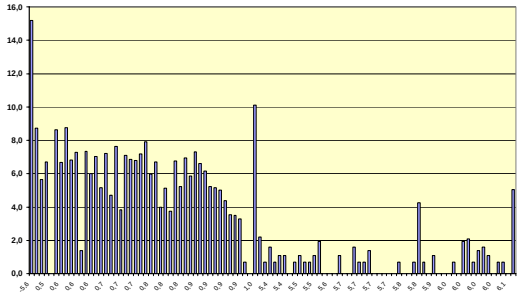
NORDESTE



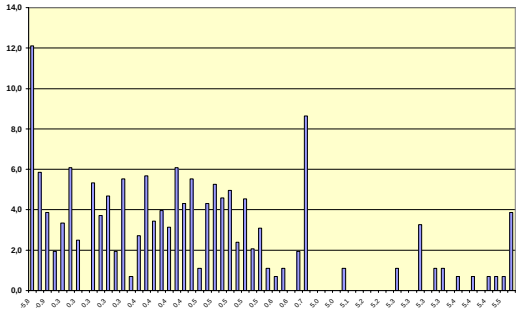
CENTRO-OESTE



SUDESTE



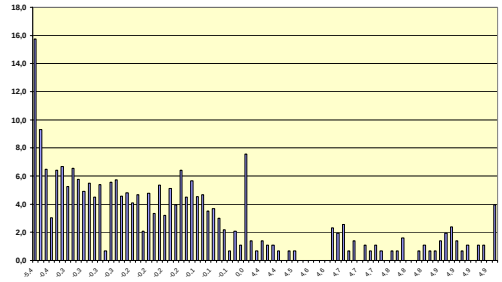
SUL



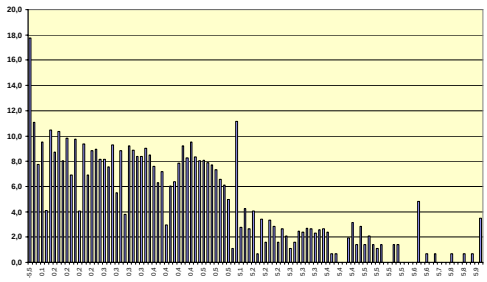
Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

GRAFICO A2. 4 – Distribuição de freqüência dos pesos totais do relacionamento. Probabilístico. Regiões. Brasil 2006. Etapa 4

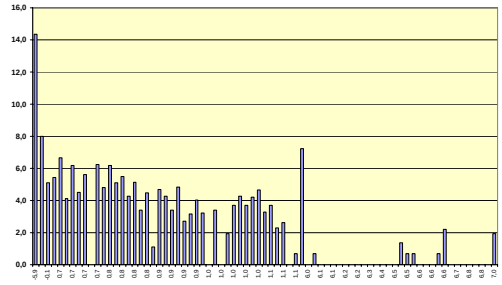
NORTE



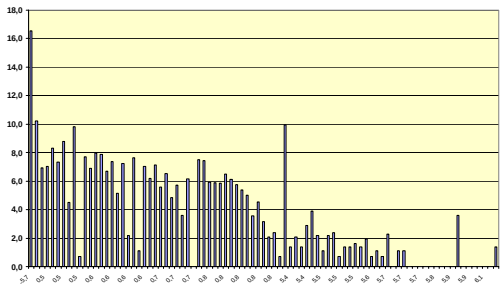
NORDESTE



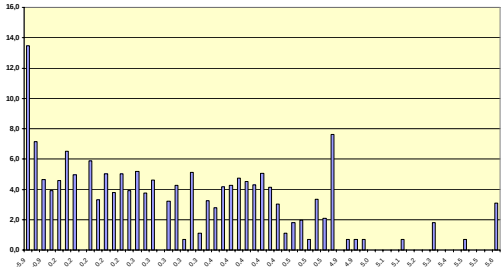
CENTRO-OESTE



SUDESTE



SUL



Fonte: Tabela elaborada com os dados da base da pesquisa de campo AIBF e dos registros administrativo do CadÚnico. 2006

APÊNDICE III

TABELA A3. 1 – Variáveis utilizadas na especificação dos modelos equilibrados do escore de propensão, segundo os cortes de renda e regiões, considerando os grupos de comparação segundo AIBF.

Nro	Nome de Variável	Variável	Tratamento x Comparação 2											
			Até 200				Até 100				Até 50			
			3	3	2	2	3	1	2	2	3	1	2	2
			Br	Nd	Nc	Ss	Br	Nd	Nc	Ss	Br	Nd	Nc	Ss
1	Dummy chefe não-branco	cor_h	x	x		x	x	x	x	x	x	x	x	x
2	Dummy chefe mulher	sexo_h	x	x	x	x	x	x	x	x	x	x	x	x
3	Dummy domicílio de qualidade inferior	quali3	x		x	x		x	x	x	x	x	x	x
4	Dummy domicílio de qualidade média	quali2	x	x	x	x	x	x	x	x	x	x	x	x
5	Dummy presença de pessoa de 60 anos ou mais	adul60a	x	x	x	x	x	x	x	x	x	x	x	x
6	dummy mãe de chefe alfabetizada	mae_alf_h	x	x	x	x	x	x	x	x	x	x	x	x
7	dummy mulher responsável presente	mulher	x	x	x	x	x	x	x	x	x	x	x	x
8	altura em metros da mulher responsável*	altura_mul		x	x	x	x	x	x	x	x	x	x	x
9	dummy homem responsável presente	homem	x	x	x	x	x	x	x	x	x	x	x	x
10	altura em metros do homem responsável*	altura_hom	x	x	x	x	x	x	x	x	x	x	x	x
11	número de membros do domicílio	tamdom		x	x	x		x	x	x	x		x	x
12	proporção de crianças entre 0 e 13 anos de idade	prc0a13a	x		x	x		x	x	x	x	x	x	x
13	Dummy de presença de crianças de 0 a 13 anos de idade	cri0a13a	x	x	x	x	x	x	x	x	x	x	x	x
14	proporção de crianças entre 0 e 6 anos de idade	prc0a6a		x	x	x		x		x	x	x	x	x
15	proporção de crianças mulheres 7a14/ crianca 0a14	razmul7a14	x	x	x	x	x	x	x	x	x	x	x	x
16	dummy casal com filhos até 14 anos	casalcfp	x	x	x	x	x	x	x	x	x	x	x	x
17	dummy chefe com até 3 anos de estudos	esc_h3		x		x	x	x	x	x	x	x	x	x
18	dummy chefe com até 4 anos de estudos	esc_h4	x	x	x	x	x	x	x	x	x	x	x	x
19	dummy chefe com até 7 anos de estudos	esc_h7	x	x	x	x	x	x	x	x	x	x	x	x
20	dummy chefe com menos de 50 anos	idad50_h	x	x	x	x	x	x	x	x	x	x	x	x
21	dummy domicílio em área urbana	urbano		x	x	x	x	x	x	x	x	x	x	x
22	dummy chefe menos de 10 anos no município	mig10a_h		x	x	x		x	x	x	x	x	x	x
23	dummy chefe menos de 5 anos no município	mig5a_h		x	x	x		x	x	x	x		x	x
24	dummy chefe viveu até os 14 anos em área rural	inf_rur_h	x	x	x	x	x	x	x	x	x	x	x	x
25	dummy região Nordeste	NE	x				x				x			
26	dummy região Norte ou Centro-Oeste	N_CO					x				x			

Br= Brasil; Nd = Nordeste; Nc = Norte e Centro Oeste; Ss = Sudeste e Sul.

Fonte: AIBF, 2005 e CadÚnico 2005.

TABELA A3. 2 – Variáveis utilizadas na especificação dos modelos equilibrados do escore de propensão, segundo os cortes de renda e regiões, considerando os grupos de comparação segundo CadÚnico.

Nro	Nome de Variável	Variável	Tratamento x Comparação 2											
			Até 200				Até 100				Até 50			
			4 Br	3 Nd	2 Nc	2 Ss	3 Br	2 Nd	2 Nc	3 Ss	2 Br	3 Nd	3 Nc	3 Ss
1	dummy chefe não-branco	cor_h	x	x	x	x	x	x	x	x	x	x	x	x
2	Dummy chefe mulher	sexo_h	x	x	x	x	x	x	x	x	x	x	x	x
3	dummy domicílio de qualidade inferior	quali3	x	x	x	x	x	x	x	x	x	x	x	x
4	dummy domicílio de qualidade média	quali2	x	x	x	x	x	x	x	x	x	x	x	x
5	dummy presença de pessoa de 60 anos ou mais	adul60a	x	x	x	x	x	x	x	x	x	x	x	x
6	dummy mãe de chefe alfabetizada	mae_alf_h	x	x	x	x	x	x	x	x	x	x	x	x
7	dummy mulher responsável presente	mulher	x	x	x	x	x	x	x	x		x	x	
8	altura em metros da mulher responsável*	altura_mul	x	x	x	x	x	x	x	x	x	x	x	x
9	dummy homem responsável presente	homem	x	x	x	x	x	x	x	x	x	x	x	x
10	altura em metros do homem responsável*	altura_hom	x	x	x	x	x	x	x	x	x	x	x	x
11	número de membros do domicílio	tamdom		x		x	x	x	x	x	x	x	x	x
12	proporção de crianças entre 0 e 13 anos de idade	prc0a13a	x		x	x	x	x	x	x	x	x	x	x
13	Dummy de presença de crianças de 0 a 13 anos de idade	cri0a13a	x	x	x	x	x	x	x	x	x	x	x	x
14	proporção de crianças entre 0 e 6 anos de idade	prc0a6a	x	x	x	x	x	x	x	x	x	x	x	x
15	proporção de crianças mulheres 7a14/ crianca 0a14	razmul7a14	x	x	x	x	x	x	x	x	x	x	x	x
16	dummy casal com filhos até 14 anos	casalcfp	x	x	x	x	x	x	x	x	x	x	x	x
17	dummy chefe com até 3 anos de estudos	esc_h3		x			x	x	x	x	x	x	x	x
18	dummy chefe com até 4 anos de estudos	esc_h4	x	x	x	x	x	x	x	x	x	x	x	x
19	dummy chefe com até 7 anos de estudos	esc_h7	x	x	x	x	x	x	x	x	x	x	x	x
20	dummy chefe com menos de 50 anos	idad50_h	x	x	x	x	x	x	x	x	x	x	x	x
21	dummy domicílio em área urbana	urbano	x	x	x	x	x	x	x	x	x	x	x	x
22	dummy chefe menos de 10 anos no município	mig10a_h		x					x			x	x	
23	dummy chefe menos de 5 anos no município	mig5a_h		x	x			x	x			x	x	
24	dummy chefe viveu até os 14 anos em área rural	inf_rur_h	x	x	x	x	x	x	x	x	x	x	x	x
25	dummy região Nordeste	NE	x				x				x			
26	dummy região Norte ou Centro-Oeste	N_CO	x				x				x			

Br= Brasil; Nd = Nordeste; Nc = Norte e Centro Oeste; Ss = Sudeste e Sul.

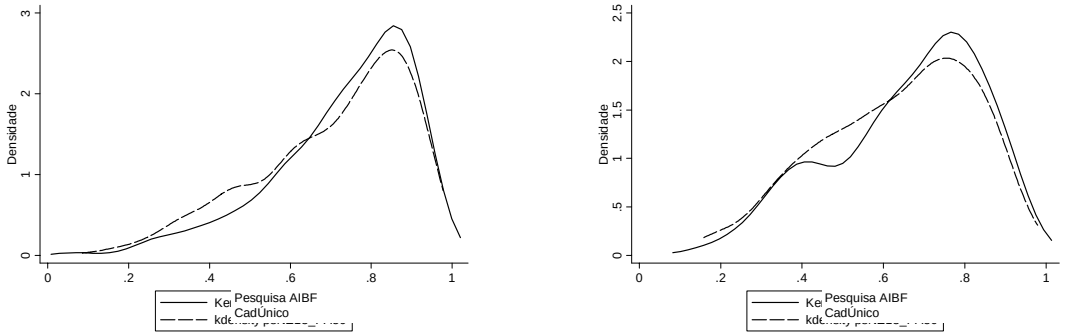
Fonte: AIBF, 2005 e CadÚnico 2005.

APÊNDICE IV

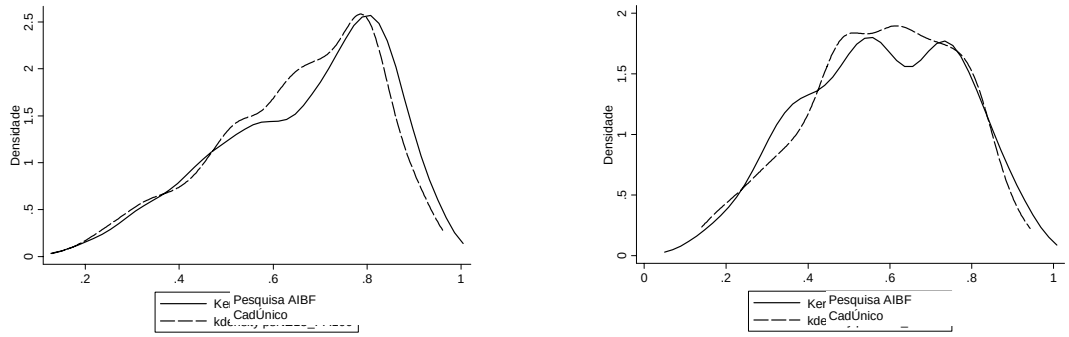
TABELA A4. 1 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Nordeste. 2005

Beneficiários do PBF | Não beneficiários de nenhum Programa

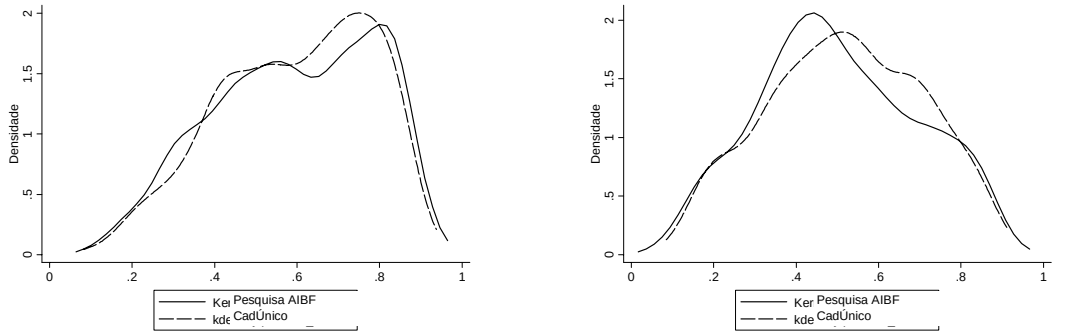
Corte de renda até R$ 50,00



Corte de renda até R$ 100,00



Corte de renda até R$ 200,00

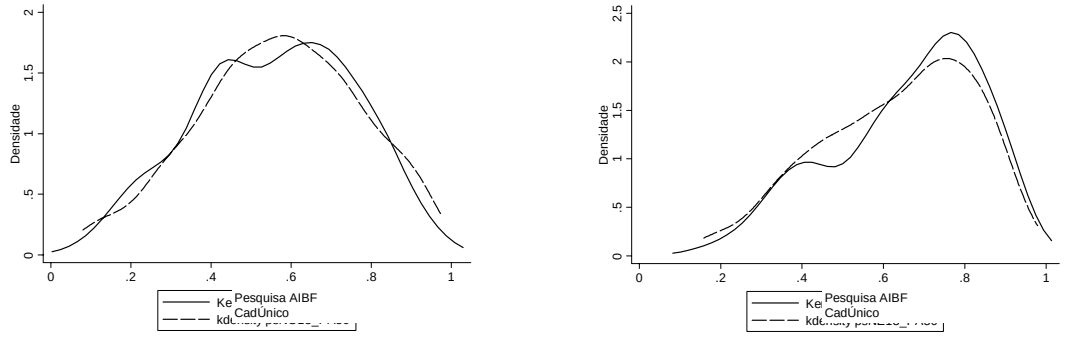


Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.

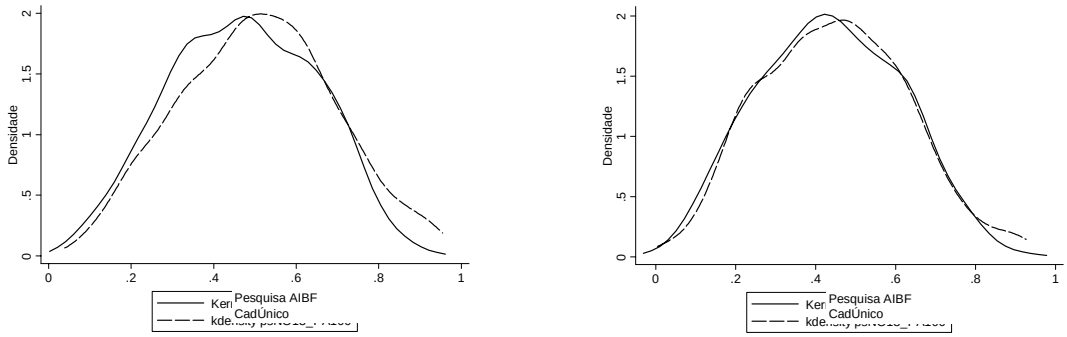
TABELA A4. 2 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Norte-Centro-Oeste. 2005

Beneficiários do PBF | Não beneficiários de nenhum Programa

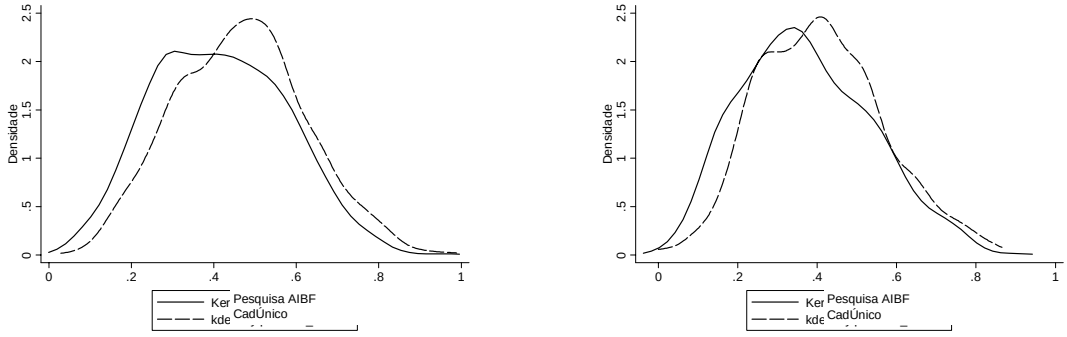
Corte de renda até R$ 50,00



Corte de renda até R$ 100,00

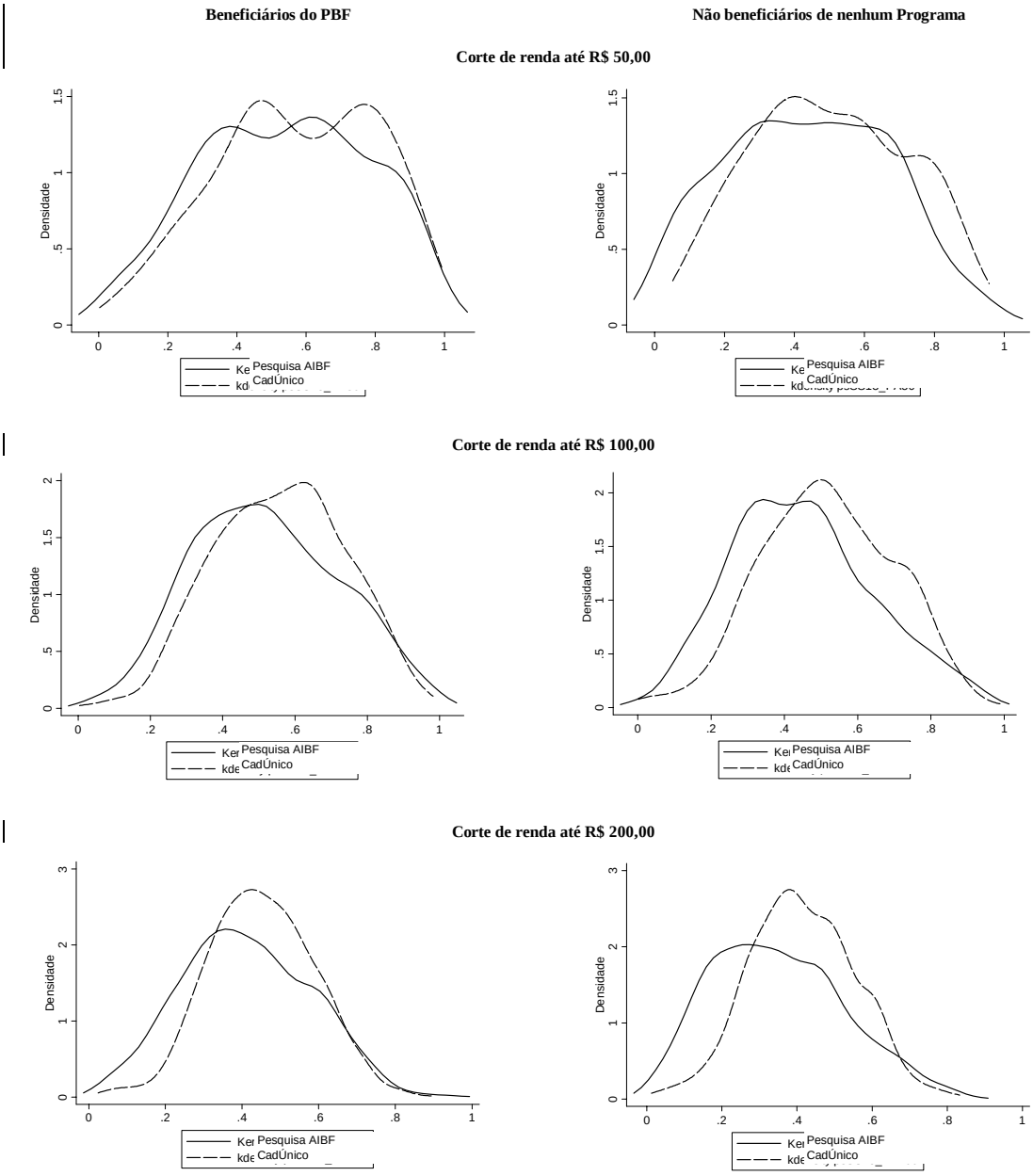


Corte de renda até R$ 200,00



Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.

TABELA A4. 3 – Distribuição de densidade da estimação do escore de propensão do balanceamento realizado entre os domicílios elegíveis, segundo tipo de alocação utilizada. Sudeste e Sul. 2005



Fonte: elaboração a partir dos dados da pesquisa de campo AIBF e registros administrativos CadÚnico.