1
Introdução

1.1.
Considerações gerais

Atualmente, em prazos cada vez mais limitados, as tomadas de decisões exigem a maior quantidade de informação possível. Usadas com este objetivo, as pesquisas amostrais, apesar da sua importância, podem tornar-se de alto custo em tempo e dinheiro. Além disso, um formulário longo, necessário para numerosos itens, pode prejudicar a qualidade dos dados, caso apresente altas taxas de não-resposta em razão da sua carga de questões (Rodrigues, 2003). Uma solução prática é explorar o máximo possível da informação já disponível nas diferentes amostras realizadas, isto é, desenvolver uma integração estatística das informações já coletadas. De um modo geral, numerosas aplicações de metodologias que integram dados, de diferentes fontes, têm sido conduzidas. Três dessas são: *merging, record linkage* e *statistical matching*. Enquanto as duas primeiras metodologias objetivam unir as mesmas unidades de dois ou mais arquivos diferentes, a terceira trata do problema de fusão de dados, que concatena unidades similares, segundo critério estabelecido, quando as unidades não têm identificadores que possam ser usados para a concatenação devido a leis que protegem o respeito à privacidade da informação, ou não possuem as mesmas unidades por serem oriundas de pesquisas diferentes.

O emparelhamento estatístico (*statistical matching*) não é uma área de pesquisa recente, mas devido ao intenso crescimento de fluxo de dados

disponibilizados e dos avanços computacionais, vem ganhando uma importância cada vez maior. O seu objetivo prático é extrair alguma informação adicional a partir das pesquisas amostrais já disponíveis. Usada desde os anos 70, (Okner, 1972), é uma metodologia de apoio à decisão que, ao unir microdados oriundos de duas ou mais pesquisas amostrais, cria um único arquivo-síntese, composto de variáveis combinadas das diversas pesquisas, mesmo que essas variáveis não tenham sido coletadas conjuntamente ou não se refiram às mesmas unidades amostrais. Tal processo objetiva, no aspecto prático, obter o máximo de informação possível a um menor custo, maior velocidade e enfoque mais amplo. No aspecto teórico, é necessário avaliar se o processo é justificável do ponto de vista estatístico.

Para evitar algum tipo de confusão conceitual, antes de qualquer procedimento de concatenação de dados, deve-se investigar o que representa cada terminologia usada para os diferentes procedimentos de integração de dados, de diferentes bases (D'Orazio et al., 2006).

Indubitavelmente, a integração de dados de duas ou mais bases diferentes de dados estabelece a possibilidade de obter uma informação conjunta, mais rica em informação. O emparelhamento estatístico (*statistical matching*) visa construir um arquivo-sintético, onde análises multivariadas sejam possíveis, no conjunto de variáveis combinadas dessas diversas pesquisas, mesmo que essas variáveis não tenham sido coletadas simultaneamente.

Antes de analisar os procedimentos de emparelhar estatisticamente, nas próximas seções se estabelecem o objetivo e a estrutura estatístico-matemática do problema do emparelhamento estatístico, formalizando a notação usada.

A seção 1.2 descreve a estrutura geral do emparelhamento estatístico.

1.2.
Estrutura do emparelhamento estatístico

Admita-se duas amostras que consistem de registros independentes gerados a partir de modelos apropriados, de acordo com o plano amostral utilizado. Para analisar o problema de emparelhar estatisticamente essas duas amostras independentes A e B , a estrutura teórica é a que segue:

Considere (**X**, **Y**, **Z**) uma variável aleatória com função de densidade ou de probabilidade *f* (**x, y, z**), $\boldsymbol{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$ *e* $\mathbf{z} \in \mathcal{Z}$ da família de distribuições $\Im = \{f\}$, e assume-se que a matriz de vetores (***X***, ***Y***, ***Z***) tem uma distribuição conjunta multivariada. Pode-se indicar, sem perda de generalidade, que **X** = $(X_1, \ldots\ldots\ldots\ldots\ldots, X_P)$**'**, **Y**=$(Y_1, \ldots\ldots\ldots\ldots\ldots, Y_Q)$**'** e **Z**= $(Z_1, \ldots\ldots\ldots\ldots\ldots, Z_R)$**'** como vetores de variáveis aleatórias de dimensão *P, Q* e *R* respectivamente.

Admita-se que A e B são duas amostras aleatórias de tamanho n_A e n_B, com observações geradas por *f* (**x, y, z**), independentes e identicamente distribuídas (*i.i.d.*).

Além disso, as unidades da amostra A têm o vetor ***Z*** omisso e nas unidades de B o faltante é o vetor ***Y.***

Considere

$$(x_a^A, y_a^A) = (x_{a1}^A \ldots\ldots\ldots x_{aP}^A, y_{a1}^A \ldots\ldots\ldots y_{aQ}^A),$$

$a = 1, \ldots\ldots\ldots\ldots\ldots, n_A$, os valores observados das unidades da amostra A , e,

$$(x_b^B, z_b^B) = (x_{b1}^B \ldots\ldots\ldots x_{bP}^B, z_{b1}^B \ldots\ldots\ldots z_{bR}^B),$$

$b = 1, \ldots\ldots\ldots\ldots\ldots, n_B$, os valores observados das unidades da amostra B .

A não ser que sejam necessários para a compreensão, para simplificar a notação os subscritos A e B serão omitidos no que se segue.

Ao se obter informação síntese sobre a distribuição conjunta de (***X***, ***Y***, ***Z***) usando as amostras observadas A e B, se está lidando com o emparelhamento estatístico.

Essas amostras, representadas na Figura 1, podem ser consideradas como uma única estrutura $A \cup B$ contendo n_A + n_B observações *(i.i.d.)* geradas pela distribuição f(***x***, ***y***, ***z***), onde a parte em branco indica as variáveis omissas, sendo caracterizada por:

i. dados faltantes e o mecanismo de geração dos mesmos;

ii. a ausência da informação conjunta ou da suposição da distribuição conjunta em X , Y e Z .

Sobre o primeiro ponto, existe uma vasta literatura estatística das possíveis caracterizações de mecanismos de geração de dados faltantes e como lidar com esta situação (vide Little e Rubin, 2002).

A segunda questão é a essência do problema do emparelhamento estatístico.

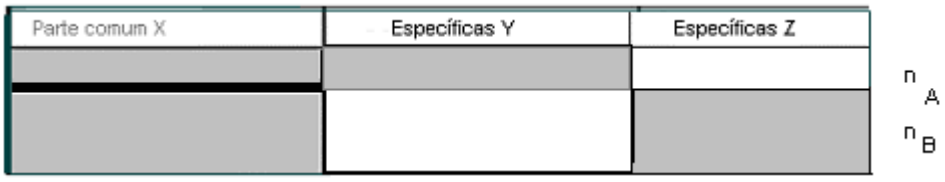

Figura 1 - Problema geral: duas amostras independentes com n_A e n_B unidades, com suas variáveis específicas Y e Z , com variáveis comuns X .

O problema do emparelhamento estatístico e suas possíveis soluções consideram a estrutura estatística dos dados e o mecanismo de dados faltantes. O mecanismo de dados faltantes no problema de emparelhamento estatístico, formalizado em Rubin (1976) apud D'Orazio et al., 2006, p. 6, define 3 diferentes

modelos de dados faltantes, que podem ser, geralmente, assumidos pelo analista como: MCAR[1], MAR[2] e MNAR[3].

O problema de emparelhamento estatístico tem uma propriedade particular: a omissão dos dados é induzida pelo desenho amostral e, portanto definida *a priori* e independentemente dos dados observados ou não.

Quando A e B são considerados conjuntamente como um único arquivo de $n_A + n_B$ unidades independentes, geradas a partir da mesma distribuição $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, com vetor Z omisso em A e o vetor Y faltante em B fica caracterizado um mecanismo de MCAR de dados faltantes para o problema de emparelhamento estatístico.

Caso se pressuponha um modelo específico, a referência é um conjunto de técnicas com uma única estrutura $A \cup B$ contendo $n_A + n_B$ observações geradas pela distribuição $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Um modelo natural identificável é aquele que assume a independência condicional de Y e Z dado X. Essa suposição é chamada de independência condicional e denominada como CIA (*conditional independence assumption*). Na literatura, são apresentadas diversas descrições e análises de diferentes abordagens de emparelhamento estatístico com o pressuposto da CIA. Entretanto, o conjunto de modelos identificáveis com a estrutura $A \cup B$, onde a CIA é válida é pequeno. Para solucionar esse problema, é necessário definir procedimentos de emparelhamentos estatísticos apropriados, de acordo, com os diferentes papéis das duas amostras A e B .

Caso a CIA não seja válida, a opção alternativa considerada é usar alguma informação auxiliar disponível, tema que será tratado na seção 1.3.3.

Recentes desenvolvimentos, no procedimento de emparelhamento estatístico têm comprovado que esse dispõe de uma vastíssima gama de

1 MCAR é a situação na qual a estrutura de dados faltantes é independente tanto dos dados observados quanto dos não observados.

2 MAR é a situação na qual a estrutura de dados faltantes depende só da parte observada.

aplicações em pesquisas e avaliações. Muitas aplicações teóricas desse método têm sido direcionadas a modelos de micro-simulação de Políticas Públicas, pesquisa de Marketing e de Estatísticas Oficiais. Dentre essas aplicações, Bourgeois (1996) escreveu um artigo sobre pobreza na Indonésia para o Banco Mundial, na Austrália (Abello et al., 2004) gera uma base de dados síntese, na área farmacêutica que serve de entrada para simular o equilíbrio entre os benefícios e taxas pagas pelos beneficiários. Agências estatísticas governamentais americanas e canadenses, assim como companhias de pesquisa em Marketing, especialmente na Europa usaram e continuam usando o emparelhamento estatístico (*statistical matching*), que na Europa é denominado fusão de dados (*data fusion*). No Canadá, entre vários trabalhos cito Alter (1974) que cria um arquivo sintético ao concatenar os registros da pesquisa *Canadian Survey of Consumer Finances* com a pesquisa *Family Expenditure Survey* de 1970. Nos Estados Unidos, uma lista de aplicações em emparelhamento estatístico pode ser consultada em Cohen, 1991b. Na Itália, Coli et al., (2005) produziu a SAM (matriz de contabilidade social) emparelhando duas pesquisas italianas. No Japão, Okauchi (2002) relata sua experiência ao fundir duas amostras japonesas.

Sendo um campo simultaneamente para muitas pesquisas teóricas, merecem registro: Kadane (1978), Rodgers (1984), Rubin (1974, 1976, 1986, 1987), Cohen (1991b), Moriarity e Scheuren (2001, 2003, 2004), Kamakura e Wendel, (1997, 2003), Rodrigues (2003), Rässler (2002, 2003, 2004) e D'Orazio et al. (2002, 2005a, 2005b, 2006) entre outros. Os referidos trabalhos representam uma formalização de alguns procedimentos de emparelhamento estatístico aplicados até agora.

No Brasil, Elbers et al. (2004) escreveram um artigo sobre pobreza e desigualdade, onde novas estimativas de medidas de pobreza e desigualdade foram realizadas a partir da combinação das pesquisas PPV (Pesquisa sobre Padrões de Vida) e PNAD (Pesquisa Nacional por Amostra de Domicílios), da Fundação IBGE.

3 MNAR é a situação na qual a estrutura de dados faltantes depende tanto dos dados observados quanto dos não observados.

1.3. Objetivo, abordagem e representatividade do emparelhamento estatístico

A escolha da metodologia do emparelhamento estatístico depende da combinação entre o objetivo do processo de integração e da informação disponível, onde o mais importante é o seu resultado, ou melhor, a representatividade da base de dados concatenada *vis-à-vis* a distribuição original $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, que será formalizada na seção 1.3.2. A informação auxiliar pode orientar a abordagem utilizada.

1.3.1. Objetivo e abordagem do emparelhamento estatístico

Existem dois possíveis objetivos para se realizar o emparelhamento estatístico: micro ou macro.

- Objetivo micro:

  Consiste na transformação apropriada de registros, que podem ser domicílios, indivíduos, empresas e etc., de arquivos distintos, em uma base de dados integrada, cujos registros contêm todas as variáveis de interesse extraídas dessas fontes distintas, a partir de unidades similares, segundo um critério estabelecido. Essa base integrada, chamada arquivo síntese ou sintético é útil para solucionar o problema de confidencialidade nos micro arquivos de uso público.

- Objetivo macro:

  As fontes de dados são usadas para estimar diretamente a função de distribuição conjunta, ou alguma característica chave, por exemplo, a correlação ou uma tabela, de variáveis de interesse, que não tenham sido coletadas conjuntamente.

Define-se a abordagem num emparelhamento estatístico como paramétrica, não-paramétrica ou uma mistura dessas estruturas – método misto (*mixed method*).

Numa abordagem paramétrica pressupõe-se uma família paramétrica conhecida de distribuição para $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Algumas vezes, a informação existente não é suficiente para se decidir por uma ou outra família, e a abordagem utilizada não pode se apoiar em tal premissa. Nestes casos a abordagem será não-paramétrica. Nos casos intermediários utiliza-se uma abordagem mista que consiste em dois passos:

Passo 1:

Paramétrico, onde um modelo paramétrico é suposto, seus parâmetros são estimados e valores são preditos para serem doados.

Passo 2:

Nesse passo, o arquivo síntese é gerado por um procedimento micro não-paramétrico, dentre outros, Moriarity e Scheuren, 2001), é um exemplo.

Subentenda-se por método misto, um procedimento com esses dois passos. No primeiro passo se aplica uma regressão, seguido da utilização de um procedimento não-paramétrico da família *hot deck*, que finaliza a fusão, obtendo o arquivo síntese ampliado (Singh, Mantel, Kinack e Rowe 1993). No caso da variável ser categórica, um modelo log-linear é a referência.

Os procedimentos não-paramétricos, quando não se supõe uma distribuição paramétrica particular para as observações de uma amostra, são mais robustos, segundo D'orazio et al. (2006).

Os procedimentos não-paramétricos mais freqüentemente usados são os de imputação da família *hot deck* (*random, rank* e *distance*), descritos na seção 3.2.

Em resumo, os procedimentos de emparelhamento estatístico em relação ao objetivo e à abordagem são indicados no Quadro 1, a seguir:

	Abordagens do Emparelhamento Estatístico		
Objetivos	Paramétrico	Não-paramétrico	Misto
Macro	♦	♦	
Micro	♦	♦	♦

Fonte: D'orazio et al. 2006b

Quadro 1 - Abordagens e objetivos do emparelhamento estatístico.

A maioria das aplicações iniciais de emparelhamento estatístico teve o objetivo micro e foram utilizados métodos não-paramétricos. Atualmente, a maioria das abordagens de emparelhamento estatístico ainda se mantém a nível micro. Entre as razões para esse fato, algumas vezes, nas aplicações dos modelos de micro-simulação, o arquivo sintético é usado como entrada de dados de algum outro procedimento. Em outros casos, prefere-se analisar o arquivo síntese ampliado, ao invés de dois ou mais arquivos contendo somente as variáveis originais. Finalmente, a informação conjunta de variáveis, que não tenham sido observadas simultaneamente, reunidas em um único arquivo pode ser de interesse para pesquisa.

Em resumo:

i. Assume-se que duas amostras A e B consistem de registros gerados independentemente a partir de duas amostras diferentes;

ii. Supõe-se que valores para um vetor de variáveis (X, Y) foram coletados em uma pesquisa A. Enquanto, uma pesquisa B coletou valores para um vetor de variáveis (X, Z);

iii. O vetor X, contém as variáveis comuns a ambos os arquivos, que servem como ponte de ligação para realizar o emparelhamento estatístico, criando o arquivo síntese (X, Y, Z);

iv. Geralmente *a priori*, é comum existir pouca ou nenhuma informação auxiliar sobre a distribuição de (Y, Z) ou sobre a distribuição conjunta de (X, Y, Z);

v. Além disso, em geral, para todas as finalidades práticas, supõe-se que existem poucas, ou nenhuma, sobreposição das unidades observadas entre os dois arquivos;

vi. O objetivo do emparelhamento estatístico é combinar esses dois arquivos, usando uma função do vetor X, para obter um arquivo ampliado que contenha (X, Y, Z);

vii. Finalmente, admita-se que se está interessado na estimação de alguma função da distribuição conjunta de (X, Y, Z), para avaliar a aderência do arquivo síntese *vis-à-vis* o arquivo teórico original (não observado na prática). Na prática, no caso brasileiro, pode-se considerar uma combinação das informações das pesquisas PNAD e POF, que poderia fornecer informações sobre parâmetros, mais abrangentes, em relação a rendimentos e despesas do que as disponíveis atualmente, como por exemplo, as variâncias e covariâncias do arquivo síntese (X, Y, Z).

1.3.2.
Representatividade do arquivo síntese predito

No início da seção 1.3 a importância da representatividade da base de dados concatenada, resultante do emparelhamento estatístico foi citada.

De forma generalizada, quatro categorias avaliam a representatividade de um procedimento de emparelhamento estatístico, vide Rässler (2002), a saber:

i. registros no arquivo síntese coincidentes com a informação verdadeira (usualmente não observável);

ii. a distribuição conjunta de todas as variáveis coincidentes com a informação verdadeira (usualmente não observável);

iii. a estrutura de correlação das variáveis preservada;

iv. as distribuições marginais e conjuntas das variáveis dos arquivos originais preservadas.

Usa-se a preservação da estrutura de correlação ou covariância das variáveis (vide D'Orazio et al., 2006), como teste para a representatividade já que o escopo deste texto não é identificar distribuições subjacentes nos arquivos a serem concatenados.

Em termos de representatividade dos arquivos preditos, quando se lida com uma abordagem paramétrica com objetivo micro, é necessário entender se o arquivo síntese criado pode satisfazer os usuários, isto é, fazer inferências a partir do arquivo síntese sobre a distribuição conjunta de (X, Y, Z) que é $f(x,y,z,\theta)$[4]. Como conseqüência, o arquivo sintético deveria ser representativo da distribuição de $f(x,y,z,\theta)$ ou, em outras palavras, o arquivo síntese poderia ser considerado como tendo sido gerado a partir da distribuição $f(x,y,z,\theta)$. Nesta abordagem, consideram-se os estimadores de máxima verossimilhança (EMV). Uma das propriedades do estimador de máxima verossimilhança, sob

[4] Sempre que necessário será explicitado que os dados foram gerados por uma distribuição paramétrica, e incluir θ na representação.

condições razoavelmente gerais, é ser consistente. Conseqüentemente, pelo menos para grandes bases de dados, $\hat{\theta}$ pode ser considerado aproximadamente igual ao parâmetro real e desconhecido θ (θ é o limite em probabilidade do estimador $\hat{\theta}$). Como resultado, o arquivo gerado ou as estimativas dos seus parâmetros, podem ser considerados como aproximadamente representativo de $f(x,y,z,\theta)$.

1.3.3.
O caso particular da CIA

Vários procedimentos são usados, dentre estes, o de imputação, não paramétrica, *hot deck* baseados nos valores das variáveis comuns X, que é a variável de concatenação, caso suponha implicitamente a independência de Y e Z dado X, que usualmente é referenciada como suposição de independência condicional ou CIA, tem como conseqüência, que a densidade conjunta f(**x**, **y**, **z**) pode ser fatorada como:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{Z|X}(\mathbf{z}|\mathbf{x})\, f_{Y|X}(\mathbf{y}|\mathbf{x})\, f_X(\mathbf{x}),$$

$$\forall\ \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y} e \mathbf{z} \in \mathcal{Z},$$

(1.3.3.1)

onde $f_{Y|X}$ é a densidade condicional de Y dado X, e $f_{Z|X}$ é a densidade condicional de Z dado X.

Pressupor a CIA, vide capítulo 3, quando se considera um modelo paramétrico, torna esse modelo identificável[5] para $A \cup B$, vide Figura 1. Então, seus parâmetros podem ser diretamente estimados. Para a equação (1.3.3.1) ser estimada é suficiente concatenar os registros com valores similares de X, o vetor de variáveis comuns, das distribuições marginais de X e Y e de X e

[5] Um modelo é identificável se todos os seus parâmetros são estimáveis a partir dos dados disponíveis.

Z . Realmente, essa informação está disponível nas amostras distintas A e B , e pode ser usada para essa estimação.

Um exemplo, em uma abordagem paramétrica de emparelhamento estatístico, supõe-se que o vetor (X, Y, Z) tem uma distribuição, por exemplo, normal multivariada, com um vetor de médias e cuja matriz de covariância (ou de correlação) é não singular:

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}$$

(1.3.3.2)

Note que todos os elementos de Σ podem ser estimados usando os arquivos A e B , exceto Σ_{YZ} ou $cor(Y,Z)$,e as suas transpostas.

Pressupondo a independência condicional e a normalidade multivariada, a matriz de covariância ou de correlação de Y e Z, Σ_{YZ}, poderia ser estimada sem usar emparelhamento estatístico, bastaria usar a propriedade da normal multivariada (Anderson, 1958, p. 28-29):

$$\Sigma_{YZ.X} = \Sigma_{YZ} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ}$$

(1.3.3.3)

Usando a suposição de independência condicional $\Sigma_{YZ.X} = 0$ e resolvendo para Σ_{YZ} ,

$$\Sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ}$$

(1.3.3.4)

Esse modelo tem tido um papel extremamente importante no emparelhamento estatístico. E nos primórdios das aplicações de emparelhamento estatístico foi suposto de maneira explícita ou implícita, mesmo

que não fosse válido. A razão é simples: esse modelo é identificável para $A \cup B$ e diretamente estimável.

Na prática, freqüentemente, e implicitamente essas amostras independentes têm sido tratadas como se fossem uma única estrutura $A \cup B$ contendo n_A + n_B observações. Isso equivale a considerar Σ_{YZ} igual a uma matriz de zeros, dado X, para as amostras a serem emparelhadas, ou seja, a suposição de independência condicional (CIA). Ou melhor, o pressuposto de que $A \cup B$ é um único arquivo gerado, das observações *(i.i.d.)*, pelo mecanismo de geração de dados faltantes da f (***x***, ***y***, ***z***).

Cumpre destacar que a CIA é uma suposição que não pode ser testada a partir dos arquivos A e B. Desta forma, não é aconselhável fazer tal suposição implicitamente. Pois caso essa suposição esteja errada, assumir a CIA conduziria a resultados viesados no relacionamento conjunto no arquivo síntese dessas variáveis, levando a decisões erradas e a avaliações inadequadas. Um exemplo, nas aplicações reais, de emparelhamento estatístico, onde a CIA pode não ser válida, é usar duas amostras realizadas em períodos diferentes.

Outra solução é utilizar a informação auxiliar externa sobre o relacionamento conjunto de (X, Y, Z) ou entre (Y, Z) deve ser seguida, sempre que for possível obter resultados confiáveis nos procedimentos de emparelhamento estatístico, evitando o pressuposto de CIA.

A informação auxiliar pode ser usualmente na forma paramétrica, isto é, se os valores de alguns parâmetros da distribuição de (X, Y, Z) são conhecidos, ou é uma amostra adicional C, onde Y e Z são observados simultaneamente. Nesta tese considera-se outro tipo de informação auxiliar, a saber uma relação existente entre as variáveis Y e Z.

Uma variante importante do emparelhamento estatístico é apresentada na seção 1.4.

1.4.
Variante importante do emparelhamento estatístico

Algumas vezes, o emparelhamento estatístico é usado como parte do planejamento de uma pesquisa, como em aplicações de amostragem matricial (*matrix sampling*). Esse método é usado para reduzir o número de itens, nos questionários administrados para as unidades de pesquisa. Essa é uma metodologia usada quando uma pesquisa é dimensionada como grande (número de variáveis do questionário e população são de grande tamanho), onde diferentes e menores grupos de questões selecionadas ao acaso, a partir do conjunto integral dos itens da pesquisa, são analisados em diferentes grupos, de mais informantes aleatoriamente selecionados.

Muitos estudiosos de amostragem matricial, dentre os quais Adams e Darwin (1982), Dillman, Sinclair e Clark (1993), Roszkowski e Bean (1990), mostram evidências empíricas de que pesquisas com questionários longos tendem a apresentar altas taxas de não-resposta. Herzog e Bachman (1980) sugeriram que a qualidade das respostas também pode ser afetada com o aumento do tamanho do questionário, onde a não-resposta é só um caso extremo. Aplicações de amostragem matricial, apresentam uma vastíssima gama de pesquisas e avaliações, produzem amostras independentes, com um número menor de variáveis, para que diferentes questionários sejam aplicados a um número maior de entrevistados, os quais respondem somente a um desses questionários. Isso vem a gerar, posteriormente uma amostra sintética, concatenada, com o emparelhamento estatístico dos diferentes questionários respondidos por unidades similares.

O emparelhamento é usado na amostragem matricial, objetivando reduzir os custos e a carga de resposta do entrevistado, bem como, para aumentar as taxas de respostas da pesquisa, tão importante para a qualidade da pesquisa.

Pesquisas que usam amostragem matricial (*matrix sampling*), também denotada como SQS (*Split Questionnaire Surveys*), têm a importante característica de serem planejadas com definições e métodos harmonizados, com papéis definidos para as variáveis: X, a variável comum que concatena os diferentes questionários e também para Y e Z, onde os diferentes questionários são respondidos por unidades similares. Um exemplo de amostragem matricial, Renssen (1998), está descrita na seção 2.3. No caso de emparelhamento de dois ou mais arquivos independentes, freqüentemente, uma custosa harmonização tem que ser realizada antes de emparelhá-los, pois as inconsistências existentes precisam ser eliminadas, vide seção 2.3.

Entre os artigos recentes que avaliam diferentes estimadores para um desenho de amostragem matricial, cito o artigo de Chipperfield e Steel (2009).

Rodrigues (2003) fez uma simulação de "*matrix sampling*" a partir dos dados do Censo Demográfico de 1991 e avaliou as distribuições marginais dos questionários originais e dos concatenados, depois da emulação da amostragem matricial *(matrix sampling)*.

Algumas aplicações de amostragem matricial são apresentadas na seção 1.4.1.

1.4.1. Aplicações de Amostragem Matricial

Nos Estados Unidos, o arquivo NHIS (*National Health Interview Survey*) aplica nos domicílios diferentes sub-questionários sobre incidência de doenças.

A técnica da *Amostragem Matricial* não é nova, segundo se depreende da documentação do Censo Americano de 1950 e 1960[6] (veja *Bureau of the*

[6] A Census that Mirrors America: Interim Report (1993) p. 36.

Census, 1965 e Goldfield, 1992). Em testes de variadas alternativas, de número de sub-questionários, para o Censo 2000 foi recomendado que fosse avaliado o mérito da *Amostragem Matricial* que usaria vários questionários intermediários no lugar de um único longo.

Observe-se que tal processo é utilizado desde 1990 pelo NAEP[7], nos Estados Unidos para avaliar alunos em geral, a partir da triagem de escolas. Os resultados publicitados de escrita do NAEP 2002 apresentam um acompanhamento de alguns níveis de proficiência desde 1998. Esse instituto avalia os estudantes americanos que respondem a uma bateria de perguntas, de várias especificações, cujas estruturas são complexas, além de possibilitarem avaliações de Economia, Língua Estrangeira, Geografia, Matemática, Ciência, História dos Estados Unidos, História do Mundo, Inglês (gramática, escrita e leitura).

A *Amostragem Matricial* é usada para construir testes quando os objetivos que estão sendo avaliados requerem mais tempo do que o disponível, circunstância na qual esse levantamento escolar se enquadra. O NAEP usa a *Amostragem Matricial* quando combina as necessidades de determinadas perguntas de avaliação. Com a *Amostragem Matricial*, os subgrupos separados de perguntas são colocados em livretos também separados. Entretanto, os efeitos da ordem das perguntas devem ser considerados em cada livreto, que contém um jogo de perguntas específicas, pois nesse contexto os efeitos práticos mostram que as perguntas que aparecem na extremidade do livreto podem ser subestimadas, por causa da fadiga ou do excesso de tempo requerido para as respostas, que pode levar o estudante a não respondê-las.

Administrar essa coleção inteira de perguntas, fortemente cognitivas, com cada estudante respondendo a todos os itens, seria tarefa demasiadamente árdua, pois consumiria um espaço de tempo bem superior ao que seria prático obter. A *Amostragem Matricial* divide em partes diferentes o conjunto inteiro de perguntas. Assim em vez de imprimir um único caderno com todas as questões a serem aplicadas a cada estudante em uma única amostra selecionada de estudantes, opta-se por imprimir, separadamente, cadernos menores (BIB), que são administrados em amostras diferentes, com um maior número de

[7] NAEP - *National Center for Education Statistics - htpp://* nces.ed.gov

estudantes. A adoção da Amostragem Matricial de conteúdos está conjugada a uma metodologia de construção de provas denominada Blocos Incompletos Balanceados (BIB) com distribuição em espiral (*Balanced Incomplete Block* (BIB) *spiraling design*). Essa metodologia permite a aplicação de 169 itens de forma a cobrir a Matriz de Referência em cada série e disciplina. Em seguida, divide-se esse conjunto em 13 blocos com 13 itens cada, agrupando-os de três em três, em 26 cadernos diferentes de prova. Dessa forma, apesar de estar avaliando um amplo escopo de conteúdos, cada aluno responde apenas a 39 itens. Este plano amostral reduz o tempo de avaliação requerido pelo estudante, além de propiciar a cobertura completa do assunto que está sendo avaliado naquela parte. Uma das vantagens do NAEP é que os livretos de teste são impressos para um assunto particular, e cada livreto pode conter uma seleção diferente de blocos cognitivos. Isto permite que o NAEP avalie a área escolhida (grupo de escolas a nível nacional ou menor nível) como um todo, dentro de uma quantidade razoável de tempo para o teste. As coleções de subgrupos são distribuídas e depois concatenadas levando em conta os estudantes similares, (é estabelecido um critério mínimo de distância entre alunos). Existe uma variação de planejamento de livretos de teste, que permite ao NAEP gerar estimativas precisas do desempenho do estudante e maximizar os recursos escassos de dados disponíveis de informação, tais como o tempo dos estudantes e também dos professores, que é medido no teste. Portanto, essas diversas perguntas necessitam ser testadas com confiabilidade.

A parte do método de *Amostragem Matricial* focalizada pelo NAEP requer que cada estudante responda a perguntas de somente uma área selecionada. O livreto (BIB) dessa matéria possibilita que os estudantes recebam seções diferentes dos formulários de avaliação, permitindo ao NAEP verificar que qualquer grupo de estudantes é avaliado, por meio de números aproximadamente iguais de questões nos diferentes livretos.

O NAEP pode provar que um número suficiente de estudantes é capaz de obter resultados precisos para cada pergunta, ao consumir, em geral, uma média de aproximadamente metade do tempo que cada estudante consumiria, se recebesse todos os itens do teste. Nesse projeto de *Amostragem Matricial*, as perguntas comuns apareceriam sempre no mesmo lugar em cada livreto de avaliação. Assim, a perícia do estudante para as perguntas que aparecem na extremidade do livreto poderia ser mais bem avaliada. Lembre-se que o cansaço do aluno pode levá-lo a subestimar determinadas perguntas, o que explica as suas não-respostas no caderno completo da pesquisa. A *Amostragem Matricial*,

por meio do contraste, permite um método mais sofisticado na criação de livretos (BIB), produzindo dados que são relativamente livres desses efeitos da ordem de colocação. Cumpre realçar que, nos BIBs do NAEP, os blocos cognitivos são equilibrados. Cada livreto cognitivo é emparelhado com pelo menos outro livreto cognitivo de um aluno similar. Desta forma, o projeto do BIB do NAEP varia de acordo com a área selecionada.

A seguir, a Figura 2 apresenta um exemplo simplificado do BIB, que é baseado no projeto-piloto de 1990 da área de Matemática do NAEP.

Uma amostra de estudantes é dividida em sete grupos equivalentes, e a cada grupo de estudantes é distribuído um dos sete livretos do teste.

Figura 2

A Model of BIB Matrix Sampling

Booklet version	Position 1 cognitive block	Position 2 cognitive block	Position 3 cognitive block
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

Fonte: NAEP, 1990

Para assegurar a distribuição apropriada no tempo da avaliação, os livretos são embalados de 1 a 7, novamente de 1 a 7, e assim por diante. O coordenador do teste distribui aleatoriamente esses livretos aos estudantes em cada sessão de administração do teste. A distribuição dos livretos garante tamanhos de amostra comparáveis para cada versão do livreto, o que permite que tais amostras sejam aleatórias e independentes. Assim, reduz-se a probabilidade de que os estudantes que estejam próximos recebam um livreto idêntico.

Uma metodologia especifica as operações do levantamento de dados, aplicando o plano de amostragem nacional e estadual do NAEP. Uma equipe de funcionários de campo, que recebe o treinamento extensivo, administra a avaliação nacional. Embora cada estado participante seja responsável pelo

levantamento de dados para o NAEP, essa equipe assegura a uniformidade dos procedimentos por meio dos manuais processuais detalhados, bem como de treinamento, supervisão e monitoração do controle de qualidade. Como o processo é complexo, a coleta de dados do NAEP é monitorada de perto. O controle rigoroso deste processo contribui para a qualidade, permitindo a comparação das avaliações dos resultados do todo e dos estados. Em resumo, o NAEP tem conduzido as avaliações escolares reduzindo a carga de questões que os estudantes devem responder, sem que sérias rupturas nos procedimentos ou nos problemas principais possam ameaçar a validade dessas avaliações.

Entre as prioridades estabelecidas pelo Ministério da Educação (MEC), destaca-se a melhoria permanente da Educação Básica no Brasil. Contribuindo para a realização de tal objetivo, o Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) implantou, em 1990, um dos mais amplos e completos esforços na coleta e sistematização de dados e análise de informações sobre o ensino fundamental e médio em nosso país, o SAEB, sendo este um importante subsídio para a compreensão dos fatores associados ao processo de ensino e aprendizagem, em diversas séries e disciplinas. Com base nas informações coletadas por ele, o MEC e as secretarias estaduais e municipais de Educação definem ações voltadas para a correção das distorções e debilidades identificadas, dirigindo seu apoio técnico e financeiro para o crescimento das oportunidades educacionais e da qualidade do sistema educacional brasileiro, em seus diferentes níveis. No Brasil, o SAEB[8], também utiliza, na prova de proficiência uma amostra de quesitos de um banco de questões para avaliar a proficiência média de uma região/estado. Em 2001, o SAEB cumpriu, de 22 a 26 de outubro, seu sexto ciclo de aplicação, avaliando o desempenho em Língua Portuguesa e Matemática dos alunos brasileiros da 4ª e da 8ª série do ensino fundamental e da 3ª série do ensino médio. Para tanto, o SAEB 2001 utilizou dois instrumentos: provas, pelas quais foram medidos os desempenhos dos alunos em Língua Portuguesa e Matemática; e questionários contextuais, pelos quais foram coletadas informações sobre alunos, turmas, professores, diretores e escolas. Para a realização do SAEB 2001, foram envolvidos 287.719 alunos, 11.737 turmas, 6.935 escolas, 21.754 funções docentes (professores) e 6.820

[8] SAEB - Sistema Nacional de Avaliação da Educação Básica - http://www.inep.gov.br/basica/saeb

diretores de escolas das redes estadual, municipal e particular em todos os Estados brasileiros e no Distrito Federal.

A partir de 1995 uma renovação do SAEB melhorou e refinou a coleta de dados para produzir informações sobre o desempenho do aluno e os fatores a ele associados, bem como a respeito das condições em que ocorre o processo ensino e aprendizagem, o SAEB utiliza procedimentos metodológicos de pesquisa formais e científicos, que garantem sua confiabilidade. Para tanto, são utilizadas provas elaboradas com um grande número de itens. Estes são distribuídos em vários cadernos de provas com 1700 itens para a montagem de cadernos de prova em blocos de itens balanceados (BIB) , o que permite uma ampla cobertura dos conteúdos e das habilidades (com seus diferentes graus de complexidade), em todas as séries avaliadas. Os itens das provas são elaborados com base nas Matrizes de Referência do SAEB, instrumentos que são produto de uma ampla consulta nacional sobre os conteúdos praticados nas escolas brasileiras do ensino fundamental e médio. Essas matrizes incorporam a reflexão de professores, pesquisadores e especialistas sobre cada área que é objeto da avaliação. A cada levantamento, além das provas, são também utilizados questionários contextuais que permitem conhecer as características da escola, do diretor, do professor, da turma e dos alunos que participam da avaliação. As escolas e turmas que participam do SAEB são escolhidas aleatoriamente, por meio de rigorosos métodos estatísticos. Como a pesquisa é amostral, cada aluno, professor ou diretor que participa do SAEB representa outros colegas, devido a expansão amostral. Deve ser destacado, ainda, que as informações coletadas pelo SAEB são sigilosas. Quando ocorre a divulgação dos resultados da avaliação, alunos, professores, diretores e escolas que integram a amostra não são identificados.

O resultado de uma prova de aplicação ampla, como a utilizada pelo SAEB, está diretamente relacionado à qualidade dos itens que a compõem. É imprescindível contar com itens elaborados com o máximo rigor metodológico, para se obter uma prova de alta qualidade técnica e fazer inferências válidas sobre o desempenho dos alunos. Em cada aplicação do SAEB, são utilizados diversos cadernos de provas(BIBs) para avaliar os conhecimentos e habilidades dos alunos em diferentes séries e disciplinas. Tais cadernos são montados por meio da Amostragem Matricial de conteúdos. Essa técnica propicia a cobertura de um amplo espectro curricular em cada levantamento, permitindo inferências sobre o sistema educacional brasileiro e não sobre os conhecimentos individuais de cada aluno.

Para garantir a comparabilidade das séries históricas, mantêm-se alguns blocos comuns e/ou itens já aplicados em anos anteriores. Por sua vez, para garantir a comparabilidade do desempenho dos alunos entre as três séries avaliadas, aplicam-se blocos da 4ª série do ensino fundamental na 8ª série do ensino fundamental, bem como blocos da 8ª série do ensino fundamental na 3ª série do ensino médio. Os resultados são analisados utilizando-se a Teoria da Resposta ao Item _TRI, que permite a comparação e a colocação dos mesmos em uma escala única de desempenho. Com isso é possível avaliar o nível médio de desempenho dos alunos nas áreas selecionadas, ainda que estes tenham respondido a diferentes conjuntos de itens.

O primeiro grande critério de estratificação refere-se à série em que o aluno está matriculado, a saber: 4ª e 8ª séries do ensino fundamental e 3ª série do ensino médio. Algumas vezes são utilizadas indiferentemente as palavras estrato ou subpopulação. Resumidamente este critério será indicado por série.

O segundo critério importante é o da unidade da Federação. A pesquisa produz estimativas para cada um dos 26 estados e para o Distrito Federal, totalizando 27 estratos nesta categoria.

Os critérios que virão a seguir são todos aplicados dentro de cada série e unidade da Federação. O terceiro critério pré-estabelecido é a dependência administrativa da escola (rede): estadual, municipal e particular. Em algumas situações, as duas primeiras são reunidas em uma única categoria intitulada como pública. O quarto critério é o da localização da escola, referindo-se ao fato dela estar sediada na capital ou no interior. Para as localizadas no interior, no estrato da 4ª série do ensino fundamental, as escolas foram ainda subdivididas entre aquelas situadas na área rural ou urbana. A partir de 1997, foram excluídas as escolas federais, as escolas rurais da Região Norte e as turmas multisseriadas. Estudos realizados em 1999 justificaram a eliminação adicional das escolas rurais de todos os estados, excetuando-se as escolas rurais com alunos na 4ª série do ensino fundamental, nos estados da Região Nordeste, de Minas Gerais e do Mato Grosso do Sul. Para a definição da Amostra do SAEB 2001, decidiu-se manter as mesmas exclusões realizadas nos dois levantamentos anteriores. O quinto critério é dividir as escolas de cada estrato de interesse em dois grupos: escolas com uma ou duas turmas regulares da série e escolas com três ou mais turmas regulares da série.

Na primeira etapa do plano amostral foram selecionadas as escolas.

Na segunda etapa, foram selecionadas as turmas dentro dessas escolas, em cada uma das séries. Se uma turma for selecionada para participar da

avaliação, todos os seus alunos serão testados (ao menos os presentes no dia da avaliação), embora em diferentes disciplinas.

Nesse sentido, podemos considerar que a amostra de alunos de cada disciplina é obtida mediante três etapas: a seleção de escolas, a seleção de turmas, e a seleção, dentro da turma, de um grupo de alunos para participarem da avaliação de cada disciplina.

Por último, vale observar que uma mesma escola pode participar de mais de um universo e, portanto, das amostras de mais de uma série, desde que tenha turmas e alunos de mais de uma das séries consideradas.

Outras aplicações citam um desenho de pesquisa proposto, visto como uma extensão do tipo Amostragem Matricial Múltipla (*Multiple Matrix Sampling Design* - MMS), descrito não só por Shoemaker (1973), mas também por Munger e Lloyd (1988). A Amostragem Matricial Múltipla obtém as respostas em amostras aleatórias de itens, a partir de todas as unidades selecionadas. Esse tipo de desenho é útil para estimativas populacionais, embora a seleção aleatória item a item, não seja operacionalizável numa pesquisa do tamanho do Censo Brasileiro, principalmente pelo encadeamento de certos conjuntos de quesitos e a quantidade razoavelmente numerosa dos itens.

Por outro lado, no livro *A Census that Mirrors America* [9], publicado em 1993, o desenho da *Amostragem Matricial* recebe a denominação de Amostra por Conteúdo, impondo certas regras na investigação das propriedades de inferências das quantidades populacionais em função do número de formulários, do número de itens de dados e da fração amostral de cada formulário, avaliando e fazendo recomendações ao Censo Americano, que após as simulações realizadas aplique-se um teste piloto usando esse plano.

Navarro e Griffin (1993) discutiram como determinar a melhor forma de dividir o conteúdo do questionário da amostra (longo) do Censo Americano de 2000 em diversos formulários que definem vários planos de Amostragem Matricial, também em função do número de formulários e da especificação de dados, bem como da fração amostral correspondente a cada formulário.

Raghunathan e Grizzle (1995) introduziram um desenho de pesquisa com o questionário dividido, onde cada componente continha um número de questões

[9] National Research Council (1993) p.24 - 40.

aproximadamente igual. O método de divisão, baseado no desenho da Amostragem Matricial, tem sido usado desde então nos testes para avaliação de resultados educacionais nos Estados Unidos.

Rässler et al. (2001) [10], baseados em dados reais de uma pesquisa sobre consumidores alemães, dividiram seu questionário para avaliar um desenho de Amostragem Matricial, a partir do trabalho de Raghunathan e Grizzle. Os resultados foram positivos para essa aplicação, ou seja esses foram superiores aos obtidos usando o plano amostral anterior a Amostragem Matricial.

10 www.icis.dk/ICIS_papers/D1_2_3.pdf

1.5.
O “problema de identificação” do método

A principal característica do problema de emparelhamento estatístico é a ausência de uma informação simultaneamente coletada sobre as variáveis de interesse. A suposição de associação entre variáveis nunca observadas simultaneamente é não-identificável, além de não poder ser estimada. Dá-se a isso o nome de “problema de identificação do método”, vide Manski (1995) ou “incerteza no método”, vide D’Orazio et al. (2006), sendo inerente ao emparelhamento estatístico. O “problema de identificação do método”, em termos de incerteza nos parâmetros do modelo é uma forma particular da incerteza para o problema de emparelhar estatisticamente, vide D’Orazio et al. 2006, página 99.

Com a finalidade de resolver esse problema, duas soluções foram consideradas na seção 1.3.3. A primeira solução foi uma suposição particular, a CIA, que conduz a um modelo identificável para $A \cup B$.

A segunda solução foi usar a informação auxiliar. Existem outros modelos além daqueles definidos pela CIA que conduzem a um modelo identificável para $A \cup B$, vide D’Orazio et al. (2006).

Entretanto, é possível que nenhuma dessas soluções propostas seja apropriada: a CIA pode ser uma suposição inválida, e a informação auxiliar pode não estar disponível. Em geral, é inviável construir, de forma precisa, a distribuição de (X,Y,Z) a partir da distribuição de (X,Y), oriunda de uma pesquisa A, e de (X,Z) proveniente de uma pesquisa B, caso não exista uma informação auxiliar sobre a distribuição de (X,Y,Z) ou (Y,Z).

Essa situação produz um tipo de incerteza no modelo de (X,Y,Z) : enquanto, dada uma amostra, problemas estatísticos padrão são caracterizados por uma única estimativa do modelo de parâmetros, isto é, o parâmetro estimado $\hat{\theta}$ é o que maximiza a função de verossimilhança, no problema do emparelhamento estatístico a informação amostral é incapaz de distinguir num conjunto, e algumas vezes em um conjunto muito grande, os possíveis parâmetros. Como resultado, as técnicas de emparelhamento estatístico devem visar:

- um conjunto de estimativas paramétricas igualmente plausíveis , quando o objetivo for macro;
- uma coleção de arquivos sínteses, sob as diferentes estimativas igualmente plausíveis, quando o objetivo for micro (ver Moriarity e Scheuren, 2001).

Em relação ao parâmetro estimado $\hat{\theta}$ que maximiza a função de verossimilhança, alerta-se que, tipicamente, após a criação de arquivos emparelhados estatisticamente, os mesmos são processados por *softwares* padrões que geram estimativas de variância e covariância. Entretanto, ignorar que um arquivo foi gerado a partir de um método de criação usando emparelhamento estatístico e processá-lo usando métodos padrões, pode conduzir a realização de inferências incorretas. Um arquivo criado pelo emparelhamento estatístico não pode ser tratado como se tivesse a mesma informação estatística de um arquivo contendo (X,Y,Z) que foram observados simultaneamente para todos os registros. A capacidade de obter estimativas confiáveis para um conjunto de dados integrado depende da representatividade desse arquivo síntese, conforme seção 1.3.2. Sims (1972a) alertou que tratar dados emparelhados como se tivessem sido observados simultaneamente

equivale a implicitamente supor uma independência condicional de Y e Z, dado X.

Relembra-se que mesmo sem possuir uma informação auxiliar sobre a distribuição de (X,Y,Z) ou (Y,Z), a maioria dos métodos de emparelhamento estatístico tem sido baseados no pressuposto de independência condicional (CIA). Mas ignorar que a CIA pode não ser válida tem como conseqüência o risco de permitir a produção de estimativas seriamente viesadas para as distribuições resultantes.

Na seção 1.6, discutem-se algumas alternativas de procedimentos para implementar soluções que tem sido propostas.

1.6. Soluções propostas

Em busca de solução para esses dilemas, no caso de distribuição normal trivariada, dentre outros estudos de simulação desenvolvidos vide D'Orazio et al., 2006b, para:

(i) comparar procedimentos mistos *versus* procedimentos paramétricos de emparelhamento estatístico, sob a CIA.

(ii) avaliar alguns procedimentos paramétricos de emparelhamento estatístico que fazem uso de estimativas externas de $cor(Y,Z)$ ou $cor(Y,Z \mid X)$.

(iii) comparar alguns procedimentos mistos de emparelhamento estatístico que fazem uso de estimativas externas de $cor(Y,Z)$ ou $cor(Y,Z \mid X)$, como exemplo consulte o procedimento misto proposto por Moriarity e Scheuren (2003), cujo procedimento similar é baseado na estimação do método de máximo verossimilhança (MV) para os parâmetros.

Uma abordagem mais geral para emparelhar estatisticamente consiste na avaliação da incerteza associada às estimativas geradas pelos métodos macro de emparelhamento estatístico quando a CIA não é válida (vide D'Orazio, et al. 2006).

A incerteza no modelo refere-se ao fato, que devido à estrutura clássica do emparelhamento estatístico, existe mais do que uma estimativa para o parâmetro de interesse dado as várias suposições sobre a distribuição de (Y,Z) que podem ser estabelecidas. Para cada uma das suposições é criado um arquivo síntese e então avalia-se a variação das estimativas feitas a partir

desses grupos de arquivos sintéticos gerados por esse procedimento e exibe-se a quantidade de incerteza devido aos emparelhamentos.

Kadane (1978) e Rubin (1986) descreveram a incerteza nos modelos paramétricos do emparelhamento estatístico. Suas idéias foram analisadas e aprofundadas por Moriarity e Scheuren (2001) que, usando o método misto, (seção 1.3.1), após o passo de regressão e antes de emparelhar somaram resíduos aleatórios às estimativas do primeiro passo; formalizam-se essas idéias na seção 2.6. A incerteza é avaliada ao se checar, para todos os parâmetros compatíveis com a informação conhecida, as estimativas geradas por esses parâmetros, quando as variáveis podem ser consideradas sob o modelo de distribuição normal.

Ainda sob a condição de normalidade, Rubin (1987) usa idéias de imputação múltipla[11] com a finalidade de obter uma coleção de arquivos sintéticos, e ao mesmo tempo, um conjunto de estimativas plausíveis. Rässler (2002) define uma técnica, totalmente bayesiana que conduz a uma abordagem apropriada de imputação múltipla.

Para as variáveis categóricas, Kamakura e Wedel (1997) usam um procedimento de imputação múltipla. D'Orazio et al. (2004) fazem uma abordagem usando máxima verossimilhança.

No caso de variáveis contínuas:

Moriarity e Scheuren (2001), no caso de uma distribuição normal trivariada (X,Y,Z) não singular, afirmam que a incerteza pode ser medida para intervalo dado por 1.6.1; com a exigência de que a matriz de covariância de (X,Y,Z), Σ,

[11] A idéia de imputação múltipla estabelece que cada dado ausente seja imputado **m** vezes, gerando **m** banco de dados completos. Esses bancos são analisados estatísticamante e combinados para uma análise final.

da equação (1.3.3.2) deve ser positiva definida, para que a $Cor(Y,Z)$ esteja contida no intervalo:

$$(Cor(X,Y) * Cor(X,Z)) \pm \sqrt{(1-(Cor(X,Y))^2) * (1-(Cor(X,Z))^2)} \qquad (1.6.1)$$

Note que $Cor(Y,Z)$ é igual a $(Cor(X,Y) * Cor(X,Z))$ no caso especial de independência condicional de (Y,Z) dado X.

Para ilustrar se ambos $Cor(X,Y)$ e $Cor(X,Z)$, forem iguais a 0,75 (ou ambos iguais a -0,75), o intervalo de valores de $Cor(Y,Z)$ é (0,125 ; 1). Pode-se atribuir valores desse intervalo para $Cor(Y,Z)$, e para cada uma dessas correlações atribuídas gerar um arquivo correspondente.

Se $Cor(X,Y)$ e $Cor(X,Z)$ forem iguais a 0,4, o intervalo de valores de $Cor(Y,Z)$ admissíveis é (-0,68 ; 1), revelando-se na prática uma restrição de pouca utilidade, que devido a amplitude do intervalo desaconselharia o uso do emparelhamento estatístico. Essa é uma das razões comumente alegada pelos pesquisadores que invocam, implicitamente, a suposição de independência condicional, seja essa suposição justificável ou não. Ou seja, se a correlação entre (X,Y) e (X,Z) for pequena, em valor, absoluto, a amplitude do intervalo estimado que contém $Cor(Y,Z)$ é grande e a incerteza é alta (os dados são compatíveis com um número grande de classes de modelo).

Moriarity e Scheuren (2001), ao revisar o trabalho de Kadane (1978), realizam uma avaliação da incerteza, para a distribuição normal trivariada usando a equação 1.6.1. Essa revisão usa simulação com computação intensiva, quantificando a incerteza nas estimativas do método de Kadane (1978) e realiza algumas correções no procedimento que reduz essa incerteza, obtendo estimativas mais confiáveis, que preservam o valor de $Cor(Y,Z)$. A correção

fundamental, dentre outras, foi a adição de resíduos aleatórios (seção 2.6), aos valores preditos no passo de regressão, antes do passo de emparelhamento dos registros. Lembre-se que o método misto, descrito na seção 1.3.1, utiliza-se de dois passos.

Com essa abordagem, pelo menos, quantifica-se a incerteza nas estimativas devida ao procedimento de emparelhamento estatístico, vide Kadane (1978), Rubin (1986), Rässler (2002) e D'Orazio et al. (2004, 2005).

Na literatura têm sido propostas diferentes formas para estimação do parâmetro no passo de regressão. A extensão para maiores dimensões, em um caso mais geral, pode ser encontrada na literatura recente em D'Orazio et al. (2006), dentre outros.

No caso de variáveis categóricas os procedimentos são:

i. explorar a incerteza na estimação dos parâmetros de um modelo multinomial usando o algoritmo EM.

ii. mostrar que para reduzir a incerteza dos parâmetros usam-se restrições paramétricas nas células, que evitam os zeros estruturais, por exemplo, não se contabiliza a fecundidade masculina ou uma pessoa de pouca idade com curso universitário, entre outras situações.

D'Orazio, Di Zio, e Scanu (2004) tratam do caso de incerteza e restrições lógicas, no emparelhamento estatístico de variáveis categóricas.

1.7.
Contribuição dessa tese

Neste texto, conceitua-se e investiga-se o emparelhamento estatístico, em casos onde seja consistente propor a afiliação estocástica (ver seção 2.4 para a caracterização do conceito), evitando a necessidade de usar fontes externas de informação auxiliar. A adição de resíduos, aqui supostos com uma distribuição conhecida, aos valores preditos, antes do emparelhamento, conduz a resultados bem sucedidos na preservação dos valores originais das matrizes de covariância e de correlação Σ_{XYZ}.

Em um estudo de simulação, baseados na hipótese de afiliação estocástica, dois métodos de concatenação estatística são desenvolvidos e comparados com o método clássico, que se baseia na suposição de independência condicional – CIA e com um método que equivale a uma regressão sem adição dos resíduos aleatórios.

Os métodos propostos usam um objetivo micro, para um emparelhamento estatístico, irrestrito, não-paramétrico, usando um procedimento *distance hot deck*, realizado em classes.

Os objetivos específicos desta tese ficam assim definidos:

1. Introduzir uma estrutura de informação auxiliar consistente, a partir dos arquivos A e B, usando a teoria econômica, na qual se baseia a hipótese do relacionamento teórico de afiliação estocástica entre duas variáveis contínuas, no nosso caso, Y (renda) e Z (aluguel), no passo de emparelhamento estatístico.

2. Construir arquivos síntese com registros completos (X, Y, Z) e investigar se estes preservam tanto a covariância quanto a correlação entre Y (renda) e Z (aluguel) do arquivo original. Comparar características do arquivo original, por meio da figura de mérito, o EQM (erro quadrático médio), com as características correspondentes dos quatro métodos computacionalmente intensivos, que aqui simulam a reconstrução da base original. Desses métodos, os dois aqui propostos, usam a informação auxiliar e soma os resíduos aleatórios, para garantir a estrutura de covariâncias e correlações. O terceiro usa o pressuposto de CIA, sem supor um modelo paramétrico. O último considera o equivalente a uma regressão (relação definida pela afiliação estocástica), sem adicionar os resíduos aleatórios.

3. Avaliar as propostas usando simulação de Monte Carlo, gerando 500 amostras-síntese para os quatro métodos em estudo, calcular as correlações Σ_{YZ} e verificar as médias das mesmas para cada método, com o objetivo de verificar se os resultados são equivalentes, ou algum desses é inferior, a partir de considerações estatísticas.

Neste estudo, admite-se que o plano amostral da PNAD (base de dados utilizada na simulação) é equivalente ao de uma amostra simples, e o fator de correção de população finita é aproximadamente 1.

A investigação do emparelhamento estatístico apropriando-se do conceito de afiliação estocástica da teoria econômica é uma novidade apresentada nessa tese.

1.8. Estrutura da tese

Esse primeiro capítulo apresenta o assunto emparelhamento estatístico e ressalta a questão relativa à ausência de informação auxiliar sobre a distribuição de (Y,Z) quando se conduz o emparelhamento estatístico.

A organização dessa tese é:

O Capitulo 2 realiza uma revisão da literatura do emparelhamento estatístico, descrevendo os procedimentos alternativos para emparelhar estatisticamente duas amostras aleatórias independentes *A* e *B* (equivalente a uma situação MCAR – ver primeira nota de rodapé da seção 1.2) , restringindo ou não, que cada registro seja doado uma única vez, bem como os cuidados prévios ao emparelhamento, vide seção 2.3. Nas seções 2.4 e 2.5 discutem-se afiliação e dependência positiva, seguida de transformação percentil monotônica.

No Capitulo 3 a suposição da CIA é uma questão crucial. Então descreve-se e analisa-se emparelhamentos estatísticos diferentes sob o pressuposto da CIA.

O Capitulo 4 descreve o uso da informação auxiliar no emparelhamento estatístico.

O Capitulo 5 formaliza os métodos aqui propostos para concatenar os arquivos A e B , classificados em classes pelo número de cômodos, gerando uma base de dados emparelhados.

O Capitulo 6 aplica os quatro métodos. Realiza e avalia um estudo computacional intensivo, onde se simulam as duas alternativas de métodos propostos, o método usando afiliação estocástica sem somar resíduos e o sob pressuposto da CIA. Avaliam-se os arquivos sintéticos resultantes dos quatro

tipos de emparelhamentos via erro quadrático médio – EQM - e o viés da matriz de variância e da matriz de covariância Σ_{XYZ}.

O Capitulo 7 dedica-se às considerações finais e futuros trabalhos.