5
Método proposto e simulação realizada

O objetivo do processo de integração e a combinação da informação disponível são determinantes para a escolha da metodologia do emparelhamento estatístico. Conforme mencionado na seção 1.3 existem diversos critérios de classificação para os procedimentos de emparelhamento estatístico. Dentre estes, a natureza tanto da variável X a ser usada como referência para o emparelhamento quanto das variáveis específicas Y e Z são determinantes. Quando a variável é quantitativa, grande parte da literatura faz referência ao método de regressão ou ao método misto. Usando variáveis contínuas e pressupondo um modelo paramétrico, de distribuição normal trivariada, em suas simulações, alguns exemplos desses métodos que podem ser consultados são Moriarity e Scheuren (2001) e Rässler (2002) dentre outros.

Neste texto a variável da parte comum X é categórica e as demais são variáveis contínuas. Além disso, não se pressupõe nem um modelo paramétrico, nem a CIA.

Em um novo procedimento de emparelhamento estatístico, os métodos propostos usam informação auxiliar, num objetivo micro, irrestrito e com uma abordagem não-paramétrica, usando *distance hot deck*, realizado em classes.

O primeiro procedimento avaliado usa informação auxiliar sobre a relação dos percentis das variáveis Y e Z. Os dois outros procedimentos aqui propostos usam além da informação auxiliar, a adição de resíduos.

A informação auxiliar, fundamentada na teoria econômica, elabora uma transformação relacional, baseada na suposição de afiliação estocástica entre as variáveis contínuas Y (renda) e Z (aluguel). Grosso modo, a suposição é de que rendas e aluguéis são localmente correlacionadas, e positivamente.

Num primeiro passo calculam-se os percentis de Y e de Z. Depois, para cada registro em A (receptor), busca-se o registro em B (doador) com a menor distância entre os percentis de Y e de Z. Este último passo é equivalente a uma regressão não-paramétrica e utiliza o *distance hot deck*. O resultado dos passos anteriores é considerado com e sem interpolação. No procedimento sem interpolação, verifica-se o valor mais próximo do resultado estimado no arquivo doador. Os procedimentos são detalhados na seção seguinte, 5.1.

O estudo investiga os resultados dos arquivos sintéticos, comparando-os com o resultado do arquivo original.

A avaliação da eficiência dos procedimentos propostos de emparelhamento dos arquivos investiga a preservação de $\sum_{YZ}$, matrizes de covariância e de correlação via erro quadrático médio. O procedimento que se baseia no equivalente à regressão sem adição de resíduos e o que supõe a CIA, também são comparados com os métodos aqui desenvolvidos.

5.1.
Introdução ao Método Proposto

No caso trivariado o emparelhamento estatístico pode ser apresentado de forma resumida como uma situação típica na qual existem duas amostras, uma de tamanho n_A contendo (X_j, Y_j), $j = 1,.....n_A$, do arquivo A e outra de tamanho n_B contendo (X_i, Z_i), $i = 1,.....n_B$, do arquivo B, oriundas de duas amostras diferentes mas com um conjunto de quesitos em comum, no caso a variável X.

O objetivo do emparelhamento estatístico consiste na integração desses arquivos para obter uma estimativa da informação conjunta de (X, Y, Z) ou (Y, Z) resultando no arquivo $\text{Sintese}\hat{}\,final$, denotado como $\hat{S}$, que representa a base de dados emparelhada ou arquivo síntese. Essas distribuições resultantes de emparelhamento não podem ser estimadas com base somente na informação disponível nesses arquivos A e B, a menos que seja suposto que as variáveis Y e Z são condicionalmente independentes dado X, o que relacionaria as variáveis Y e Z linearmente e equivaleria à hipótese da CIA.

O método de emparelhamento estatístico, aqui proposto é irrestrito, usa classes e se baseia na existência de informação auxiliar. Usa-se como informação auxiliar a hipótese sobre Y e Z, dado X, ou seja, a relação de afiliação estocástica entre as variáveis Y e Z, dado X, que é uma restrição mais geral do que a CIA. Na verdade equivale a uma quasi-linearidade local. Aqui a renda e o aluguel são teoricamente relacionados como afiliadamente estocásticas, a partir da teoria Econômica, sendo esta a informação usada a

partir de uma transformação relacional dessas variáveis. Na seção 2.4 discute-se afiliação e dependência positiva. Existem outros diferentes tipos de informação auxiliar, conforme descrito no capítulo 4.

Observe que como a simulação é feita para reproduzir a base original de dados conhecida, os arquivos síntese são gerados como se estivessem usando amostragem matricial, vide seção 1.4.

As três alternativas são aplicadas, comparadas entre si, e comparadas considerando a CIA implicitamente.

Os procedimentos são aqui denominados de:

I - Transformação não-paramétrica relacional (sem adição de resíduos) – equivale à regressão;

II - Transformação não-paramétrica relacional intervalar (com adição de resíduos) e busca – **TNRIr**;

III - Transformação não-paramétrica relacional com interpolação (com adição de resíduos) e busca - **TNRIo**

e são aplicados em cada uma das classes do valor de X.

Em cada um dos procedimentos, para criar os seus referidos arquivos sínteses, utilizam-se B e A como doadores, com a imputação de Z no arquivo A e Y no arquivo B. Finalmente, concatenam-se os respectivos arquivos sintéticos resultantes da doação dos arquivos B e A, em cada classe X, que resulta nos arquivos sintéticos relativos a cada um dos procedimentos.

Um estudo de simulação é conduzido usando uma base real de dados (X,Y,Z), cujas matrizes de covariância e de correlação são conhecidas, e que pode servir para validar cada procedimento proposto.

Para cada procedimento de emparelhamento dos arquivos A e B, e uma vez criado o seu respectivo arquivo Síntese, verifica-se se a estrutura de covariância e de correlação das variáveis é preservada. O enfoque principal é investigar a preservação de $\sum_{YZ}$ do arquivo original, nos processos, que é uma das indicações da representatividade da base de dados emparelhada dos mesmos. Verifica-se se existem grandes diferenças entre os 4 procedimentos, através do erro quadrático médio e compararam-se os resultados de viés, dos dois tipos de emparelhamentos, a partir das suas estimativas da covariância e correlação das variáveis Y e Z.

Quando o arquivo A for usado como doador, um procedimento simétrico aplicará os mesmos processos formalizados quando o arquivo B foi considerado doador (este último será descrito).

5.2.
Metodologia da simulação

5.2.1.
Base reais de dados

Usa-se uma amostra de valores de vetores (X,Y,Z) retirada da PNAD, 2005, referente aos domicílios em imóveis alugados na região metropolitana do Rio de Janeiro.

Supõe-se que os registros estejam classificados em L classes de emparelhamento, preliminarmente realizado pela variável X: número de cômodos do domicílio.

Por conveniência, supõe-se que X, Y e Z são valores de vetores univariados. Y e Z representam as transformações logarítmicas da renda e do aluguel, respectivamente.

A variáveis Y e Z são submetidas à transformação logaritmo. Realiza-se a transformação monotônica $g(.)$, percentil das variáveis Y (ou Z). Associa-se o ordenamento da variável Y, ou seja a ordem da variável p_Y (ou p_Z), sendo equivalente a trabalhar diretamente com a renda e o aluguel, mas a transformação logarítmica deixa a distribuição mais simétrica, e obviamente, não afeta o cálculo do percentil. Teoricamente a partir da distribuição de Y, $f_Y(y)$ pode-se calcular o percentil correspondente a partir do inverso da função cumulativa $F_Y(y)$, ver detalhes em 2.6, o mesmo se aplica a Z.

Esse arquivo passa a ser composto pela quíntupla ordenada (X,Y,Z,p_Y,p_Z). A vantagem do uso da transformação monotônica percentil é a

facilidade de identificação da distribuição do par (p_Y, p_Z), que deveriam ser do mesmo tipo, uniformemente distribuída em $[0,1]^2$, onde se busca a relação quase linear entre p_Y e p_Z, que é a localmente linear. Apesar da existência de uma relação direta entre Y e Z, a mesma é de difícil identificação, dado o relacionamento não-linear entre a renda e o aluguel.

Antes da divisão da quíntupla ordenada (X, Y, Z, p_Y, p_Z), uma vez em ordem crescente as variáveis Y, Z são denotadas por:

$$Y_{(1)} \leq Y_{(2)} \leq \ldots\ldots\ldots\ldots \mathrm{Y}_{(n-1)} \leq Y_{(n_A)}$$

e

$$Z_{(1)} \leq Z_{(2)} \leq \ldots\ldots\ldots\ldots \mathrm{Z}_{(n-1)} \leq Z_{(n_B)}$$

Associa-se a cada $Y_{(k)}$ e $Z_{(k)}$ uma transformação relacional $g(.)$ e $f(.)$, na verdade, os percentis (monotônicos) p_{Yj} e p_{Zi}. Como se está lidando com amostras, poderiam existir poucos pontos (registros), mas isso não ocorre nas nossas amostras.

Na base (X, Y, Z, p_Y, p_Z), podem-se calcular os resíduos, como a diferença entre os percentis p_{Yk} e p_{Zk}:

$$residuo_k = p_{Yk} - p_{Zk} \qquad k = 1,\ldots\ldots\ldots\ldots, n_A + n_B \qquad (5.2.1)$$

Dada essa relação de afiliação estocástica entre os pares de variáveis (Y, Z), a suposição confiável de ordenamento crescente de Y e Z justifica o uso da transformação das variáveis aleatórias Y e Z, associando a cada Y e a

cada Z uma transformação relacional, na verdade, utiliza-se o percentil p_Y e p_Z, respectivamente, a ser realizada gerando essa informação auxiliar, representada pela quíntupla (X,Y,Z,p_Y,p_Z), selecionada e transformada conjuntamente, onde os resíduos são construídos usando a equação (5.2.1). Estes resíduos na prática seriam oriundos de outra pesquisa e sua distribuição seria também informação auxiliar.

Com respeito ao resíduo, note que por causa da simetria, o resíduo para p_{Zk} é igual a menos o resíduo em relação a p_{Yk}. Esses resíduos são guardados em um banco de dados de resíduos.

A base de quíntuplas ordenadas (X,Y,Z,p_Y,p_Z) com $n_A + n_B$ registros é dividida por amostragem sistemática simples, e consideradas como amostras aleatórias independentes da mesma população, em 2 arquivos, A com (X,Y,p_Y) e B com (X,Z,p_Z), não existindo sobreposição de registros.

Os arquivos divididos, A e B, são condicionados ao valor de X, onde X indica o número de cômodos, cujas distribuições são denotadas por $f_{Y|X}(y \mid x)$ e $f_{Z|X}(z \mid x)$.

As variáveis renda e aluguel são variáveis aleatórias com uma dada distribuição conjunta. Essas variáveis apresentam um modelo comportamental, que no processo de decisão sobre a escolha de uma alternativa de aluguel, é influenciado por fatores racionais e subjetivos. Os fatores racionais são aqueles explicados a partir de características sócio-econômicas dos indivíduos residentes nos domicílios. Os fatores subjetivos são aqueles que não são expressos

diretamente a partir de conceitos econômicos, advindos de fatores aleatórios, da decisão subjetiva associada a cada domicílio i ou j.

$$Y_j \qquad j = 1,.....,n_A$$

$$Z_i \qquad i = 1,.....,n_B$$

Conforme definido em (2.4.2) no conceito de afiliação estocástica, para dois domicílios quaisquer i , j:

$$\forall \varepsilon > 0 \qquad \exists \quad \delta > 0 \quad \ni$$

Se $Y_i > Y_j + \delta \Rightarrow P(Z_i > Z_j) < \varepsilon$

Diz-se então que Y e Z são estocasticamente afiliados.

A suposição de rendas e aluguéis serem localmente correlacionadas, e positivamente conduz ao conceito de dependência positiva. Considere-se o caso bivariado e suponha que as duas variáveis aleatórias reais (p_Y, p_Z) tem distribuição conjunta *F* e uma função de densidade estritamente crescente *f*. A partir dos conceitos formalizados em (2.4), a propriedade (VII) é satisfeita e podemos dizer que renda e aluguel são estocasticamente afiliados.

O espaço da família de distribuições $\Im = \{f\}$ conjuntas possíveis das triplas (X,Y,Z) após a transformação, torna-se o espaço das triplas (X, p_Y, p_Z).

Por exemplo, com as 8 classes de número de cômodos, X, encontradas na PNAD (a última classe inclui 8 cômodos ou mais), se Y_j = ln ($renda_j$) para $j = 1,.....,n_A$ e Z_i = ln ($aluguel_i$) para $i = 1,.....,n_B$, para a base real de

dados usada, a figura 3 mostra a diferença de amplitude entre os *box-plots* de Y_j e Z_i versus p_Y e p_Z.

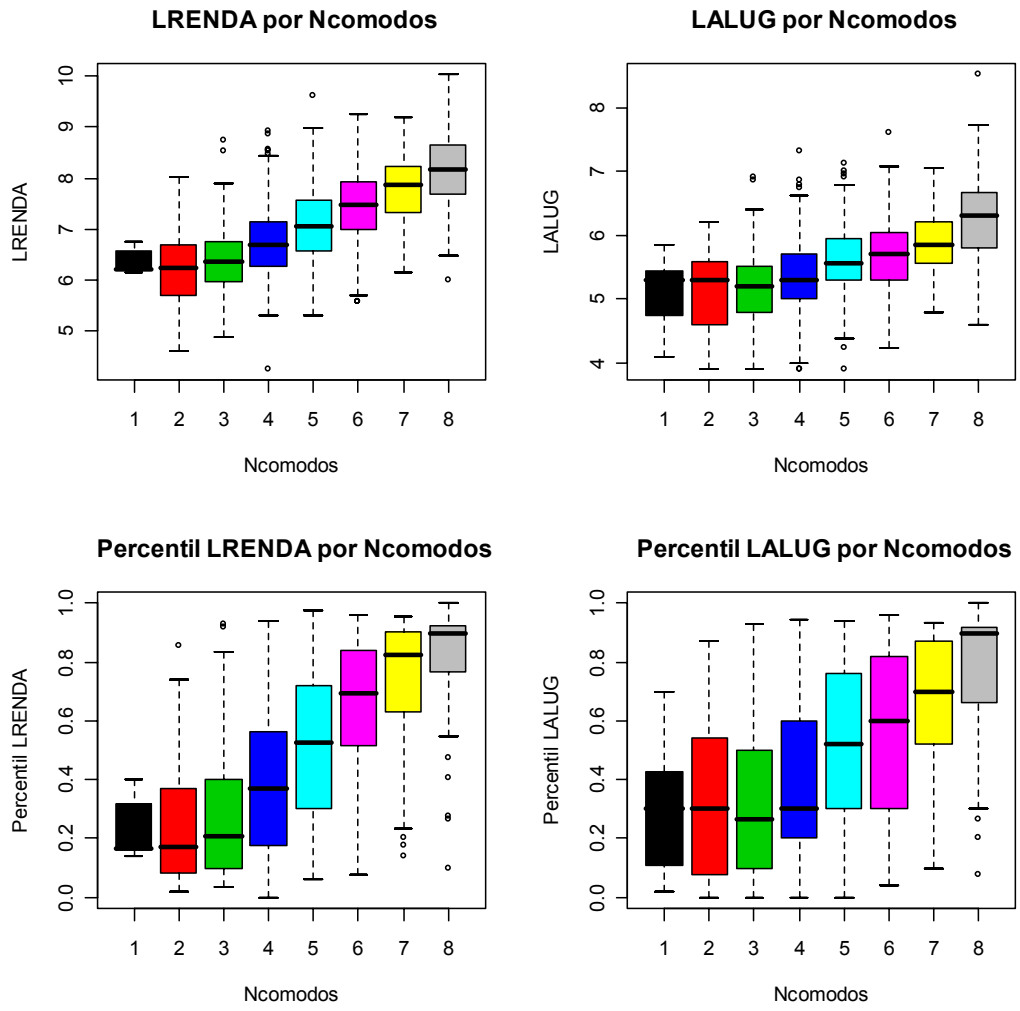

Figura 3 - *Box-plot* Lrenda, Laluguel , p_Y e p_Z por número de cômodos

5.2.2.
Propostas de procedimentos não-paramétricos de emparelhamento estatístico

Nesta seção descrevem-se os procedimentos propostos, que consideram uma abordagem não-paramétrica de emparelhamento estatístico irrestrito nas classes de X . Utilizam informação auxiliar, gerada por uma transformação relacional envolvendo as variáveis contínuas, estocasticamente afiliadas, Lrenda e Laluguel.

Avalia-se o uso da transformação não-paramétrica relacional no emparelhamento e numa das alternativas cria-se um ponto fictício a partir da interpolação linear e convexa de dados existentes, por isso as denominações de não-paramétrica relacional e transformação não-paramétrica relacional com interpolação, para os procedimentos que a partir de agora serão denotadas respectivamente pelas siglas TNRIr e TNRIo.

A base de dados (X, Y, Z) tem dimensão (n_A + n_B) = n . Ambos os arquivos A receptor e B doador têm dimensão $n/2$.

Para cada valor $j = 1,.....,n_A$, que são os valores observados das unidades da amostra A , existe um percentil p_{Y_j}, que é considerado determinístico. No arquivo doador B , seus percentis p_{Z_i} estão ordenados.

Em cada classe, executam-se os seguintes passos:

Primeiro passo: Transformação não-paramétrica

Devido ao fato dos indivíduos residentes nos domicílios exibirem características não previsíveis em sua conduta, inclui-se um componente

aleatório associado à condição de afiliação estocástica, para superar estas limitações. O componente aleatório chamado neste texto de resíduo correspondente - res_j é somado ao p_{Yj}. A suposição de afiliação estocástica permite que esta "regressão" seja denominada $\ddot{p}_{Zj}$, e equivale a uma transformação relacional não paramétrica. O resíduo aleatório res_j, escolhido aleatoriamente do banco de dados de Resíduos, é um arquivo separado de dimensão n, e foi inicialmente calculado no arquivo original. Como já comentado, na prática, seria uma informação "auxiliar" oriunda de outras fontes.

Uma igualdade, em cada classe, com um resíduo aleatório correspondente res_j é adicionado a p_{Yj}, obtendo $\hat{\ddot{p}}_{Z_j}$:

$$\hat{\ddot{p}}_{Z_j} = p_{Y_j} + res_j \qquad (5.2.2)$$

com inversa da sua função de distribuição acumulada empírica é :

$$\hat{\ddot{Z}}_j = \hat{F}^{-1}(\ddot{P}_{Z_j}) = \inf\{z : \hat{F}(z) \geq \ddot{P}_{Z_j}\} \quad \text{onde} \quad \hat{\ddot{p}}_{Z_j} \in [0,1]. \qquad (5.2.3)$$

Para garantir os valores dos percentis de $\hat{\ddot{p}}_{Z_j}$ entre zero e um, os resíduos gerados usados têm que estar dentro do limite das retas (5.2.2.a), (5.2.2.b) e (5.2.2.c), preservando $\hat{\ddot{p}}_{Z_j}$. A figura 4 ilustra esses limites para os resíduos.

$$p_{Yj} \in [0;0,2] \Rightarrow res_j \in \left[-\frac{3}{4}p_{Yj}, 3p_{Yj}\right] \qquad (5.2.2.a)$$

$$p_{Y\,j} \in [0{,}2;0{,}8] \Rightarrow res_j \in \left[-\frac{3}{4}p_{Y\,j}, \frac{3}{4}-\frac{3}{4}p_{Y\,j}\right] \qquad (5.2.2.b)$$

$$p_{Y\,j} \in [0{,}8;1] \Rightarrow res_j \in \left[-3+3p_{Y\,j}, \frac{3}{4}-\frac{3}{4}p_{Y\,j}\right] \qquad (5.2.2.c)$$

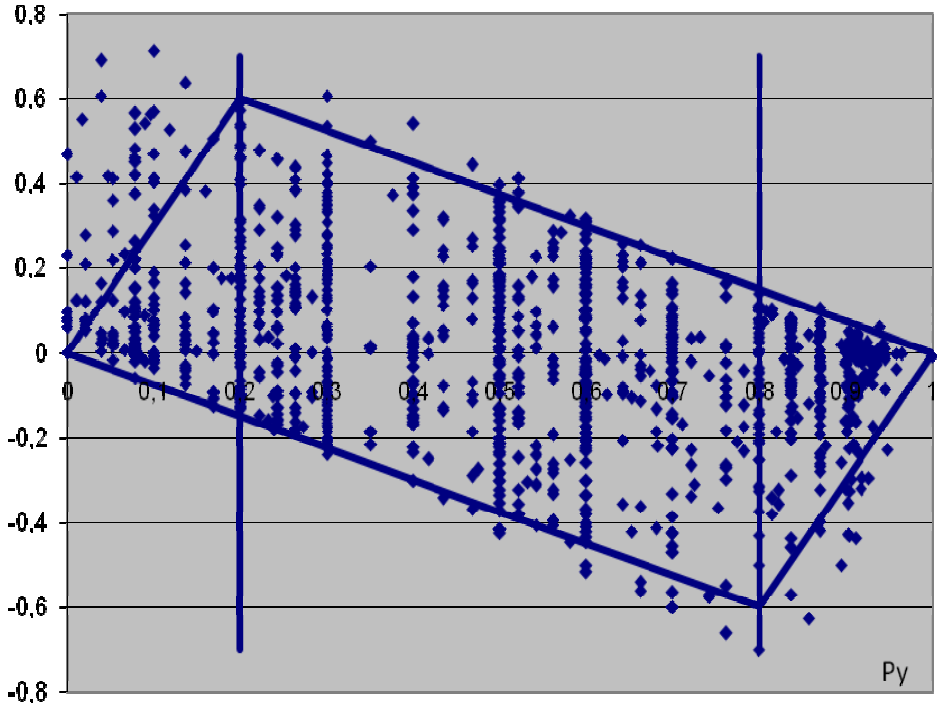

Figura 4 - Distribuição uniforme bidimensional dos resíduos, limitados pelas retas que garantem os percentis preditos entre 0 e 1.

Um gráfico similar pode ser feito para garantir os valores dos percentis de $\hat{p}_{Y_i}$ entre zero e um, os resíduos gerados usados têm que estar dentro do limite das suas respectivas retas, para preservar $\hat{p}_{Y_i}$ e limitar os resíduos.

Segundo passo: Predição do valor a ser imputado

Na lista dos p_{Z_i} ordenados do arquivo doador B, de dimensão $n/2$, existe um intervalo, com um limite inferior e superior, que contém a estimativa do valor predito $\hat{\ddot{p}}_{Z_j}$.

Para um determinado $\hat{\ddot{p}}_{Z_j}$ existe um intervalo, tal que, $\hat{\ddot{p}}_{Z_j} \in [p_{Z(k-1)}, p_{Z(k)}]$ e esses possuem valores observados relacionados $[Z_{(k-1)}, Z_{(k)}]$ no arquivo B, conforme mostrado no Quadro 2 abaixo:

p_Z	Z
…	…
$p_{Z(k-1)}$	$Z_{(k-1)}$
$\hat{\ddot{p}}_{Z_j}$	$?=\hat{Z}_j$
$p_{Z(k)}$	$Z_{(k)}$
…	…

Quadro 2 - Valores do intervalo que contém $\hat{\ddot{p}}_{Y_i}$

Esses percentis do intervalo, os valores observados, $p_{Z(k-1)}$ e $p_{Z(k)}$, são usados para determinar o valor de $\hat{Z}_j$, a ser imputado no arquivo receptor A. Com esses dois valores $p_{Z(k-1)}$ e $p_{Z(k)}$, se propõem duas alternativas para estimar o valor de $\hat{Z}_j$, **TNRIr** e **TNRIo**, descritas a seguir.

5.2.2.1. TNRIr

Na primeira alternativa é a transformação não-paramétrica relacional intervalar - **TNRIr**, escolhe-se o valor mais próximo de $\hat{\ddot{p}}_{Z_j}$ entre os valores observados $p_{Z(k-1)}$ e $p_{Z(k)}$, do intervalo, que será a referência para a doação ao arquivo A . Para cada intervalo que contém esse percentil estimado $\hat{\ddot{p}}_{Z_j}$, seu limite inferior e superior dos percentis observados, tem valores observados correspondentes $Z_{(K-1)}$ e $Z_{(K)}$, respectivos, no arquivo B . Dentre esses dois valores de p_Z , escolhe-se o valor mais próximo de $\hat{\ddot{p}}_{Z_j}$ para ser o $\hat{Z}_j$ doador de A .

Ou seja, para cada $a = 1,.....,n_A$, um valor observado de Z_i do arquivo B é imputado no $a-ésimo$ registro em A , usando o Z_i cujo p_{Z_j} seja o mais próximo de $\hat{\ddot{p}}_{Z_j}$, que é o predito, a partir de um procedimento *distance hot deck*, usando a distância absoluta, descrita na seção 3.2.3, em cada classe. Essa distância absoluta é avaliada como:

$$d_j = \min | \hat{\ddot{p}}_Z - p_i | \qquad i = k-1, k \qquad (5.2.2.1)$$

Em resumo, para um determinado $\hat{\ddot{p}}_{Z_j}$ existe um intervalo tal que $\hat{\ddot{p}}_{Z_j} \in [p_{Z(k-1)}, p_{Z(k)}]$ e esses possuem valores observados relacionados $[Z_{(k-1)}, Z_{(k)}]$ no arquivo B .

O valor observado de Z_i , $i = k-1$ ou k, do arquivo B imputado é aquele com a menor distância d_j . Identifica-se essa variável como $\tilde{\tilde{Z}}_j$.

Então, o arquivo A síntese aumentado, com reposição, após a imputação com procedimento **TNRIr** será escrito como $(X_j, Y_j, \tilde{\tilde{Z}}_j)$, e notado por $\mathbf{Sintese}_{\mathbf{A}}^{\bullet\bullet}$.

Na prática, se pode utilizar *distance hot deck* numa sub-área com dados semelhantes.

5.2.2.2. TNRIo

A segunda alternativa é a transformação não-paramétrica relacional com interpolação - **TNRIo**. A partir de uma interpolação da estimativa $\ddot{Z}_j$ obtida, calcula-se o valor de $\hat{Z}_j$ que é o valor obtido dessa interpolação, a ser imputado no registro do arquivo receptor A , da forma a seguir.

O limite inferior e superior dos percentis observados no intervalo, que contém o percentil estimado $\hat{\ddot{p}}_{Z_j}$, tem os valores observados correspondentes $Z_{(K-1)}$ e $Z_{(K)}$ respectivos, no arquivo B . Os valores $p_{Z(k-1)}$, $p_{Z(k)}$, $Z_{(K-1)}$ e $Z_{(K)}$ são usados na interpolação linear para estimar $\hat{Z}_j$:

$$\hat{Z}_j = (1-r_j)Z_{(K-1)} + r_j Z_{(K)} \qquad (5.2.2.2.1)$$

onde,

$$r_j = \frac{\ddot{p}_{Z_j} - p_{Z(k-1)}}{p_{Z(k)} - p_{Z(k+1)}} \qquad (5.2.2.2.2)$$

Observe que apesar da baixa probabilidade, pode acontecer de um valor observado de p_Z ser exatamente igual a $p_{Y(j)}$ + res_j. Também pode ocorrer de existirem vários p_Z iguais, neste caso utiliza-se o último dos mesmos para usar na interpolação.

Após a imputação com procedimento de Interpolação, o arquivo A síntese aumentado com o vetor $\hat{Z}$ estimado, será o arquivo sintético $(X_j, Y_j, \hat{Z}_j)$, chamado $\text{Sintese}_{\text{A}}^{\wedge}$.

Os resultados da simulação analisam se os procedimentos propostos são uma alternativa que pode produzir resultados consistentes na prática ao emparelhar estatisticamente, e serão investigados no próximo capítulo.