Pontifícia Universidade Católica
do Rio de Janeiro

Nádia Maria Coelho Rodrigues

Concatenação Estatística de Dados e Afiliação Estocástica

Tese de Doutorado

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da PUC_Rio como requisito parcial para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Reinaldo C. Souza
Co-orientador: Kaizô I. Beltrão

Rio de Janeiro, 13 de outubro de 2009.

PUC-Rio - Certificação Digital Nº 0410318/CA

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Nádia Maria Coelho Rodrigues

Concatenação Estatística de Dados e Afiliação Estocástica

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Reinaldo C. Souza
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Kaizô I. Beltrão
Co-Orientador
ENCE/IBGE

Profa. Monica Barros
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Victor Hugo de Carvalho Gouvêa
UFF

Prof. Sérgio da Costa Cortes
Departamento de Informática – PUC-Rio

Profa. Ana Carolina Letichevsky
Fundação Cesgranrio

Prof.José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de janeiro, 13 de outubro de 2009.

PUC-Rio - Certificação Digital Nº 0410318/CA

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, da autora e do orientador.

Rodrigues, Nádia Maria Coelho

Na Escola Nacional de Ciências Estatísticas graduou-se em Estatística em 1976 e no curso de mestrado em Estudos Populacionais e Pesquisas Sociais em 2003.

PUC-Rio - Certificação Digital Nº 0410318/CA

Ficha Catalográfica

Rodrigues, Nádia Maria Coelho

Concatenação estatística de dados e afiliação estocástica / Nádia Maria Coelho Rodrigues ; orientador: Reinaldo C. Souza ; co-orientador: Kaizô I. Beltrão. – 2009.

159 f. ; 30 cm

Tese (Doutorado em Engenharia Elétrica)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Emparelhamento estatístico. 3. Suposição de independência condicional (CIA). 4. Informação auxiliar. 5. Procedimento hot deck. I. Souza, Reinaldo C. II. Beltrão, Kaizô I. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

PUC-Rio - Certificação Digital Nº 0410318/CA

A meus pais e aos meus filhos por seu amor, carinho e seu contínuo apoio e estímulo.

Agradecimentos

Aos meus orientadores Professores Reinaldo C. Souza e Kaizô I. Beltrão, pelo apoio, confiança e pela sua dedicada orientação.

Ao professores examinadores Ana Carolina Letichevsky, Monica Barros, Sérgio da Costa Cortes e Victor Hugo de Carvalho Gouvêa.

Ao IBGE pelo apoio financeiro. Aos funcionários e colegas do IBGE, principalmente da Diretoria de Informática.

À Pontifícia Universidade Católica de Rio de Janeiro, a todos os professores e funcionários do Departamento.

A meus amigos Mário Gama e Emilia Matos por contínuo seu apoio e colaboração.

A todos os amigos e familiares que de uma forma ou de outra me estimularam ou ajudaram.

PUC-Rio - Certificação Digital Nº 0410318/CA

Resumo

Rodrigues, Nádia Maria Coelho; Souza, Reinaldo C.. **Concatenação Estatística de Dados e Afiliação Estocástica.** Rio de Janeiro, 2009. 159p. Tese de Doutorado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O emparelhamento estatístico é a técnica de combinar informações de duas ou mais fontes de dados, possivelmente de pesquisas independentes, que possuam um subconjunto comum de variáveis, para produzir uma informação mais abrangente e coerente, em um arquivo de dados síntese, onde as variáveis observadas nas diferentes amostras são gravadas conjuntamente. Métodos computacionalmente intensivos viabilizam novas formas de emparelhamento estatístico. Em um estudo de simulação, um único arquivo de dados da PNAD é dividido em duas bases de dados para emular um caso de amostragem matricial. Esses dois arquivos são emparelhados estatisticamente utilizando quatro metodologias, e os resultados das mesmas são comparados com as do arquivo único original. A CIA (*conditional independence assumption*) não parece ser válida. Para evitar a suposição de independência condicional (CIA*)*, os três métodos de emparelhamento estatístico desenvolvidos são baseados na hipótese do relacionamento teórico de afiliação estocástica entre as variáveis contínuas renda e aluguel, e dois deles usam também a informação de resíduos. Os métodos são comparados entre si e com o método clássico, que se baseia na suposição de independência condicional – CIA. Em uma abordagem não-paramétrica, com um objetivo micro, os métodos de emparelhamento estatístico propostos são irrestritos, e realizam-se em classes, definidas pela variável de número de cômodos. Usam um procedimento *distance hot deck*, além de adicionar os resíduos supostos conhecidos. Esse estudo investiga os resultados de viés e do EQM, dos quatro métodos, investigando a preservação da correlação original entre renda e aluguel.

Palavras-chave

Emparelhamento estatístico, suposição de independência condicional (CIA), informação auxiliar, procedimento *hot deck*

PUC-Rio - Certificação Digital Nº 0410318/CA

Abstract

Rodrigues, Nádia Maria Coelho**;** Souza, Reinaldo C.. **Statistical Matching using Stochastic Affiliation.** Rio de Janeiro, 2009. 159p. DSc Thesis - Departamento deEngenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Statistical matching is the art of combining information from two or more datasets, collected in independent samples, but a subset of the variables is common to both samples, to produce both coherent and comprehensive information in a synthetic data file, where variables observed in different samples are jointly recorded. Modern computing can make possible, under techniques described here, an advance in the application of Statistical Matching. It is reported on a simulation that splits a single file into two pieces, statistically matched the pieces using four methodologies and compares the results to the original single file. The Conditional Independence Assumption (CIA) did not seem a valid assumption. To avoid CIA, it is suggested two methods of statistical matching where kind of auxiliary information, based on Stochastic Affiliation relationship between income and rent, and residuals are used, in conjunction with nonparametric regression and hot deck distance. Both are compared not only with the classic method based on the CIA and regression without residuals but also with each other. In a nonparametric approach, with a micro objective, an unconstrained statistical matching is applied using hot-deck nearest neighbor within classes, using the household logarithm income and logarithm rent percentile groups. We are satisfied with the performance of creating a correlation coefficient of Y and Z, as measured using MSE.

PUC-Rio - Certificação Digital Nº 0410318/CA

Keywords

Statistical matching, conditional independence assumption (*CIA*), auxiliary information, hot deck imputation procedure.

Sumário

PUC-Rio - Certificação Digital Nº 0410318/CA

PUC-Rio - Certificação Digital Nº 0410318/CA

Lista de Figuras

PUC-Rio - Certificação Digital Nº 0410318/CA

Lista de Tabelas

PUC-Rio - Certificação Digital Nº 0410318/CA

Lista de Quadros

PUC-Rio - Certificação Digital Nº 0410318/CA

Lista de Siglas

EQM	Erro quadrado médio
IBGE	Instituto Brasileiro de Geografia e Estatística
i.i.d	Independente e identicamente distribuídos
MEC	Ministério de Educação e Cultura
MRJ	Método de replicação *Jackknife*
NAEP	National Center for Education Statistics
NHIS	National Health Interview Survey
PNAD	Pesquisa Nacional por Amostra de Domicílios
POF	Pesquisa de Orçamentos Familiares
SAEB	Sistema Nacional de Avaliação de Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
PPV	Pesquisa sobre Padrões de Vida
CIA	*conditional independence assumption*
SAM	*Social Accounting Matrix*
TNRIr	Transformação não-paramétrica relacional intervalar
TNRIo	Transformação não-paramétrica relacional com interpolação
EMV	Estimador de máxima verossimilhança
SQS	*Split Questionnaire Surveys*
PLRD	*positive likehood ratio dependence*

PUC-Rio - Certificação Digital Nº 0410318/CA