Universidad Politécnica de Cataluña

Fusión de Datos: Imputación y Validación

Tesis Doctoral

Presentada para obtener el grado de doctor por la

Universitat Politècnica de Catalunya

Carlos Alberto Juárez Alonso

Departamento de Estadística e Investigación Operativa
Programa: Aplicaciones Técnicas e Informáticas de la Estadística, la Investigación Operativa y la Optimización.

Dr. Tomás Aluja Banet
Director de tesis

Universidad Politécnica de Cataluña
Departamento de Estadística e Investigación Operativa
Barcelona, España
Octubre 2001 – Noviembre 2004.

Este documento se realizó gracias en parte al apoyo proporcionado por la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES), a través del programa SUPERA y a la Universidad de las Américas Puebla.

- Contenido -

Bibliografía

Aluja Banet, T.; Morineau, A.; Rius R. (1997). La greffe de fichiers et ses conditions d'aplication. Méthode et exemple. *Colloque Francophone sur les sondages.* Rennes.

Aluja Banet, T.; Morineau, A. (1999). *Aprender de los datos: el análisis de componentes principales*. Primera Edición. Barcelona: EUB.

Aluja Banet, T.; Thio, S. (2001). Survey Data Fusion. *BMS*, 72: 20-36.

Anderson, T.W. (1984*). An Introduction to Multivariate Statistical Analysis*. Second Edition. New York: John Wiley & Sons.

Bárcena, M.J. (2000). *Técnicas multivariantes para el enlace de encuestas*. Tesis, Barcelona: Universidad Politécnica de Cataluña

Bárcena, M.J.; Tusell, F. (1999). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. *Qüestiio*, 23 (2): 297-320

Bárcena, M.J.; Tusell, F. (2000). Tree-based algorithms for missing data imputation. *Proceedings of the 15th International Workshop on Statistical Modeling*, (eds: E. Ferreira y V. Núñez-Antón), p. 300-305, Bilbao, 2000.

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Cazals, F. *Nearest-Neighbor Search in High Dimension and Molecular Clustering.* <http://algo.inria.fr/seminars/sem96-97/cazals.html> [Consulta: febrero 2003]

Chang, K.; Ghosh, Joydeep. *Principal Curve Classifier A Nonlinear Approach to pattern Classification.* <http://www.lans.ece.utexas.edu/~kuiyu/paper/kuiyu1998ijcnn.pdf>. [Consulta: junio 2003].

Co V. (1997). *Méthodes statistiques et informatiques pour le traitement des données manquantes*. Tesis, Paris: Conservatoire National des Arts et Métiers.

Cochran, W.G. (1977). Sampling Techniques. Third Edition. John Wiley & Sons.

Cohen, ML (1991), Data fusion: an appraisal and experimental evaluation, Journal of the market research society, 31(2), pp153-212

Comyn, M. (1999). *Modélisation et validation des rapprochements et fusions de fichiers d'enquêtes*. Tesis, Paris: Ecole nationale supérieure des télécommunications.

Coppock, S. (2002). *Multi-Modal Data Fusion A Dissertation Proposal*.
< http://www.ececs.uc.edu/AppliedAI/coppock.proposal.pdf > [Consulta: junio 2003].

Costa, A; Satorra, A; Ventura, E. (2002). Estimadores compuestos en estadística regional: una aplicación a la estimación de la tasa de variación de la ocupación en la industria. *QÜESTIIO*, 26 (1-2): 213-243.

Chavez E., Navarro G., Baeza-Yates y Marroquin J.L., Searching in metric spaces. *ACM Computing Surveys*, 33(3):273-321, 2001

Dasarathy B.V., Neareast neighbour norms:NNpattern classification techniques. *IEEE Computer Society Press*, 1991.

Delanoy, C. (1979) Un algorithme rapide de recherche de plus proches voisins. *R.A.I.R.O. Informatique/Computer Science*, 14 (13): 275-286.

Delicado, P. (2001). Another Look at Principal Curves and Surfaces. *Journal of Multivariate Analysis*, 77: 84-116

D'orazio, M.; Di Zio, M.; Scanu, M. *Statistical Matching: a tool for integrating data in National Statistical Institutes.*
URL<http://webfarm.jrc.cec.eu.int/ETKNTTS/Papers/final_papers/43.pdf>.
[Consulta: julio 2003].

Escofier, B.; Pages, J. (1992). *Análisis factoriales simples y múltiples*. Segunda Edición. Bilbao: U.P.V./E.H.U.

Fisher, N.; Derquenne, C.; Saporta, S. (2001). *A new Method to Match Data Sets Applied to Electric Market*. En: Proceedings NTTS-ETK 2001, Eurostat conference on: New Techniques and Technologies for Statistics, Exchange of Technology and Know-how. <http://cedric.cnam.fr/AfficheArticle.php?id=270>. [Consulta: julio 2003].

Friedman J.H.; Baskettand F.; Shustted L,J. (1975). An algorithm for finding nearest neighbors. *IEEE Transactions on computer*, 24:1000-1006.

Fukunaga, K.; Narendra, P.M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers,* 26: 917-922.

Greenacre, M.J. (1993). *Correspondence analysis in practice*. London: Academic Press.

Hart, P.E. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 4(May): 515-516.

Hastie, T.; Stuetzle W. (1989). Principal Curves. *Journal of the American Statistical Association*, 84 (406): 502-516.

Johnson, R.A.; Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Fourth Edition. Prentice Hall.

Kaufman, L,; Rousseeuw, P.J. (1990): *Finding groups in data. An introduction to Cluster analysis*. New York: Wiley.

Kittler J. (1978). A method for determining k nearest neighbors. *Kibernetes*, l7: 313-315

Law, A.M.; Kelton, W.D. (1991). *Simulation Modeling & Analysis*. Second Edition. Mc Graw Hill.

Lebart L. Lejeune M (1995). Assessment of Data Fusions and Injections. *Encuentro Internacional AIMC sobre Investigación de Medios.* Madrid*,* 208-225.

Lebart L.; Morineau A.; Fénelon J.-P. (1985). *Tratamiento estadístico de datos.* Barcelona [etc.]: Marcombo DL.

Little, R.J.; Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New Jersey: John Wiley & Sons.

Liu, J.S.; Wu,Y.N. (1999). Parameter Expansion for Data Augmentation. *Journal of the American Statistical Association,* 94: 1264-1274.

Longford, N.T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society*, 162:227 – 246.

Martínez-Abarca, M-J.; Aluja, T. (1999). Fusión de datos de audiencia: Metodología y su aplicación en el mercado publicitario. *II Seminario sobre nuevas tecnologías.* AEDEMO, 129-146.

Moreno, F. (2004). *Clasificadores eficaces basados en algoritmos rápidos de búsqueda del vecino más cercano*. Tesis, España: Universidad de Alicante.

Montgomery, D.C. (1997). *Design and Analysis of experiments.* Fourth Edition. New York: John Wiley & Sons.

Morrison D. F. (1990). *Multivariate Statistical Methods* Third Edition. New York: McGraw-Hill.

Rassler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York:Springer-Verlag.

Rencher A C. ;(1995). *Methods of multivariate analysis*. New York: John Wiley & Sons

Rius, R.; Aluja, B.; Nonell, R. (1999). File Grafting in Market Research. *Applied Stochastic Models in Business and Industry, 15:451-460*

Rubin D.B. (1987). *Multiple imputation for non response in surveys.* Second Edition. New York: John Wiley & Sons.

Santini G. (1984). La méthode de fusion sur référentiel factoriel. *Séminaire IREP, mars 1984.*

Saporta G. (2002). Data fusion and data grafting. *Computational Statistics & Data Analysis*, 38: 465-473.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* First edition. London: Chapman & Hall.

Schafer, J.L. (2000). *Multiple Imputation: Fabricate your data well.* Preparado para la Sociedad Estadística de Washington. < http:// www.stat.psu.edu/~jls/wss.pdf>. [Consulta: junio 2003].

Soong, R.; Montigny, M. (2001). *The Anatomy of Data Fusion*. Worldwide Readership Research Symposium (Venice). < http://www.zonalatina.com/Zldata183.htm>. [Consulta: junio 2003].

Soong, R. (2001). *Data Fusion in Latin America*. 47ª conferencia anual de la Fundación de Investigación en Publicidad, N.Y. < http://www.zonalatina.com/Zldata166.htm> [Consulta: junio 2003].

Tennenhaus M. (1998). *La Régression PLS Théorie et Pratique*. Paris: Editions Technip

Tussel, F. (2002). Neural networks and predictive matching for flexible imputation. *DataClean 2002 conference, Jyväskylä (Finland), 30 May 2002.*

Van der Putten, P.; Kok, J.N.; Gupta, A. (2002) Data fusion through statistical matching. Paper 185, Center for eBusiness@MIT, MIT Sloan School of Management. < http://ebusiness.mit.edu/research/papers/185_Gupta_Data_Fusion.pdf> [Consulta: junio 2003].

Wayman, J. C. ;(2003) Multiple Imputation for Missing Data: What Is It And How Can I Use It? "003 Annual Meeting of the American Educational Research Association, Chicago, IL.

Wilson, D.L. (1972). Asymptotic properties of nearest neighbors rules using edited data. *IEEE Transactions on systems, Man and Cybernetics*, SMC 2(3): 408-421.

Yunck, T.P. (1976). A technique to identify nearest neighbors. *IEEE Transactions on systems, Man and Cybernetics*, SMC 6(10): 678-683.