



**Universidad Politécica 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 Politécnic 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 -

	Introducción	1
1	Fusión de datos	
1.1	Introducción	7
1.2	Tratamiento de los datos incompletos	12
1.2.1	Modelos explícitos	16
1.2.1.1	Métodos de imputación por la media	17
1.2.1.2	Métodos de imputación por regresión	18
1.2.1.3	Métodos de imputación por maximización de verosimilitud	22
1.2.2	Modelos implícitos	26
1.2.2.1	Métodos de imputación Hot deck: Muestreo aleatorio simple	27
1.2.2.2	Métodos de imputación Hot deck: Por clase	27
1.2.2.3	Métodos de imputación Hot deck: Secuencial	27
1.2.2.4	Métodos de imputación Hot deck: Vecino más cercano	28
1.2.3	Imputación Múltiple	28
1.3	Técnicas de Fusión	31
1.3.1	Emparejamiento aleatorio intra celular	32
1.3.2	Emparejamiento estadístico	32
1.3.3	Fusión mediante descripción factorial	34
1.3.3.1	Fusión por matrimonio	34
1.3.3.2	Fusión por búsqueda de socios	36
1.3.3.3	Fusión por búsqueda de vecinos más cercanos	37
1.3.4	Fusión mediante Árboles de clasificación	39
1.3.5	Fusión mediante Redes neuronales	42
1.3.6	Fusión mediante Análisis homogéneo	43
1.3.7	Fusión mediante regresión PLS	44
1.4	Validación	46
1.4.1	Propuestas de validación	47

2	Métodos de los k vecinos más cercanos	
2.1	Introducción	51
2.2	Regla de decisión y selección de la distancia	52
2.3	Métodos de búsqueda.	56
2.3.1	K-Dimensional Tree (k-d tree)	58
2.3.2	Algoritmo de Fukunaga/Narendra	59
2.3.3	Vantage Point Tree (vp-tree)	59
2.3.4	Geometric Near-neighbor Access Tree (GNAT)	60
2.4	Algoritmo de Fukunaga/Narendra	62
2.4.1	Búsqueda por el método de Branch and Bound	64
2.4.2	Aplicación del algoritmo	67
2.4.3	Restricciones	71
2.4.4	Costo	74
3	Propuesta	
3.1	Introducción	77
3.2	Métodos de imputación	81
3.2.1	Take One Deterministic Multivariate	82
3.2.2	Take One Stochastic Univariate	85
3.2.3	Take One Stochastic Multivariate	87
3.2.4	Take K Deterministic Multivariate	87
3.2.5	Take K Stochastic Univariate	90
3.2.6	Take K Stochastic Multivariate	95
3.3	Otras opciones	99
3.3.1	Variables Mixtas	99
3.3.2	Redondeo	100
3.4	El parámetro $\pi$	100
3.4.1	Estimadores para áreas pequeñas	101
3.4.2	Imputación TKDM	103
3.4.2.1	Variables continuas	103
3.4.2.2	Variables categóricas	104
3.4.3	Imputación TKSU	105
3.4.3.1	Variables continuas	105

	3.4.3.2 Variables categ3ricas	105
	3.4.4 Imputaci3n TKSM	106
	3.4.4.1 Variables continuas	106
	3.4.4.2 Variables categ3ricas	106
3.5	Validaci3n	106
	3.5.1 Comparaci3n de estadisticos marginales	109
	3.5.1.1 Prueba de igualdad de medias	109
	3.5.1.2 Prueba de igualdad de varianzas	109
	3.5.1.3 Prueba conjunta de igualdad de medias	110
	3.5.1.4 Intervalos de confianza para la media de las variables	110
	3.5.2 Homogeneidad	114
	3.5.3 Exactitud	119
3.6	El Sistema GRAFT	123
4	Simulaci3n	
	Parte I	
4.1	Definici3n del archivo de simulaci3n	129
4.2	Construcci3n de la base de datos (Simulaci3n I)	130
4.3	Ensayos (Simulaci3n I)	141
	4.3.1 Imputaci3n T1DM	146
	4.3.1.1 Resultados	146
	4.3.1.2 An3lisis de los resultados	148
	4.3.2 Imputaci3n T1SM	149
	4.3.2.1 Resultados	149
	4.3.2.2 An3lisis de los resultados	151
	4.3.3 Imputaci3n TKDM	152
	4.3.3.1 Resultados	152
	4.3.3.2 An3lisis de los resultados	154
	4.3.4 Imputaci3n TKSM	155
	4.3.4.1 Resultados	155
	4.3.4.2 An3lisis de los resultados	157
4.4	Construcci3n de la base de datos (Simulaci3n II)	158

4.5	Ensayos (Simulación II)	167
4.5.1	Imputación T1DM	170
4.5.1.1	Resultados	170
4.5.1.2	Análisis de los resultados	172
4.5.2	Imputación T1SM	173
4.5.2.1	Resultados	173
4.5.2.2	Análisis de los resultados	175
4.5.3	Imputación TKDM	176
4.5.3.1	Resultados	176
4.5.3.2	Análisis de los resultados	178
4.5.4	Imputación TKSM	179
4.5.4.1	Resultados	179
4.5.4.2	Análisis de los resultados	181
Parte II		
4.6	Ensayos (parámetro $\pi$ )	182
4.6.1	Ensayos (Simulación I)	182
4.6.2	Análisis de los resultados	185
4.6.3	Ensayos (Simulación II)	186
4.6.4	Análisis de los resultados	189
4.7	Imputación EM	190
4.7.1	Resultados	190
4.7.2	Análisis de los resultados	193
4.8	Imputación PLS	194
4.8.1	Resultados	194
4.8.2	Análisis de los resultados	197
4.9	Tabla comparativa	198
Parte III		
4.10	Prueba del sistema	
4.10.1	Descripción del problema	
4.10.2	Imputación TKSM	
Conclusiones		213
Bibliografía		220

## Bibliografía

Aluja Banet, T.; Morineau, A.; Rius R. (1997). La greffe de fichiers et ses conditions d'application. 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. *Qüestio*, 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. *QÜESTIÓ*, 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](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*, 17: 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](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.