Metodología de Encuestas
Vol 3, Núm 2, 2001, 151-162

EDICIÓN DE ENCUESTAS MEDIANTE REDES DE NEURONAS ARTIFICIALES

María Dolores Cubiles de la Vega
Ana Muñoz Reyes
Universidad de Sevilla

Rafael Pino Mejías
Universidad de Sevilla
Centro Andaluz de Prospectiva

Begoña Buiza Camacho
Instituto de Estudios Sociales Avanzados de Andalucía

RESUMEN

Se presenta un procedimiento de imputación de valores perdidos y un método para la detección y corrección de inconsistencias en las respuestas recogidas como resultado de una encuesta estadística. Para ello se describe el Perceptrón Multinivel, modelo concreto de Redes de Neuronas Artificiales utilizado en nuestro trabajo, y se ilustra el funcionamiento del procedimiento sobre la cuestión "Intención de voto" de una encuesta electoral del Centro de Investigaciones Sociológicas. Sobre estos datos reales, la técnica de imputación construída se basa en un modelo de predicción de la intención de voto a partir de las demás cuestiones, presentando una capacidad de generalización estimada que puede calificarse de perfecta. El modelo de detección y corrección de inconsistencias ofrece un rendimiento bastante satisfactorio, por lo que el perceptrón multinivel, confirmando algunos trabajos existentes con datos simulados, se puede considerar como un método prometedor en las tareas de edición de los registros resultantes de una encuesta estadística.

Palabras clave: *redes de neuronas artificiales, edición de encuestas, perceptron multinivel, imputación de datos*.

Introducción

La recogida de datos de cualquier encuesta está sujeta a riesgos serios de errores, que suelen manifestarse, entre otros, en problemas de falta de respuesta o bien en la existencia de registros inconsistentes. Por tanto, es fundamental la edición de los registros resultantes del proceso de encuestación, entendiendo como edición el proceso orientado a la depuración del conjunto de registros, lo que conlleva en particular la imputación de valores perdidos y la detección de inconsistencias, es decir, respuestas incorrectas (distintas a la real) para una o más cuestiones.

Existen diversos procedimientos de edición de registros, originados sobre todo por el trabajo de formalización realizado por Fellegi y Holt (1976), que aun titulándose automáticos requieren la intervención de expertos en la materia. Esta circunstancia, en el caso de tamaños muestrales elevados, puede conllevar un alto coste de las tareas de edición. En este trabajo se describe una aproximación a la tarea de aumentar el grado de automatización de los procedimientos de edición, utilizando modelos basados en Redes de Neuronas Artificiales (en adelante RNAS).

Las RNAS constituyen un conjunto de modelos matemáticos no lineales, utilizados de forma práctica en muchas áreas de la ciencia moderna (Rumelhart et al., 1994). Su gran flexibilidad, caracterizada por diversas propiedades teóricas que convierten estos modelos en aproximadores universales (Ripley, 1996), y el vertiginoso aumento de las prestaciones de los equipos informáticos, las convierten en una poderosa herramienta apropiada para obtener predicciones multidimensionales a partir de entradas también multidimensionales. Por ello, la literatura recoge un creciente número de aplicaciones de las RNAS: concesión de créditos, procesamiento del lenguaje natural, tratamiento de imágenes, reconocimiento de patrones, predicción de series temporales, etc, convirtiendo a las RNAS en una importante técnica dentro de la investigación científica aplicada.

Entre las referencias sobre aplicaciones de las RNAS en las tareas de edición de registros estadísticos, destacan las experiencias de Nordbotten (1995 y 1996), si bien algunas de las aplicaciones descritas en estos trabajos se basan en simulaciones de encuestas. Recientemente, algunas institucionales oficiales estadísticas comienzan a experimentar con la edición estadística mediante RNAS, como se recoge en algunos documentos de trabajo correspondientes a sesiones sobre la Edición Estadística de datos, de la Conferencia de Estadísticos Europeos (Statistics Denmark, 1999; Eurostat, 2000). Los resultados obtenidos en estos trabajos sugieren posibilidades efectivas de éxito para la depuración automática mediante RNAS, pero también recalcan la necesidad de nuevas investigaciones que permitan ahondar en el conocimiento de todo el proceso de construcción y evaluación de tales técnicas.

Nuestro trabajo entronca así con las directrices expresadas en el marco general de evaluación de la eficiencia de la edición estadística de datos, según se recoge en el material metodológico utilizado por la Comisión Estadística de Naciones Unidas y la Comisión Económica Europea, (Nordbotten, 1999).

En el segundo apartado se describe la arquitectura de RNAS utilizada, el perceptrón multinivel. El tercer apartado presenta una aplicación del perceptrón multinivel como

modelo de imputación de la intención de voto en una encuesta electoral. Para esa misma cuestión, en el cuarto apartado se describe un experimento controlado, diseñado para medir la eficacia del perceptrón multinivel como modelo para la detección y corrección de inconsistencias.

Redes de Neuronas Artificiales: el Perceptrón Multinivel

Redes de Neuronas Artificiales

Una RNAS puede describirse como un sistema compuesto por un número, en general elevado, de elementos de procesamiento, también llamados neuronas artificiales o nodos, interconectados entre sí. Cada enlace o conexión tiene asociado un parámetro, llamado coeficiente sináptico. Cada elemento de procesamiento aplica una función, llamada función de activación, a la información que le llega desde otros nodos a él conectados, y envía la salida resultante a otros nodos. Dado un modelo de RNAS y un conjunto de datos (llamado conjunto de entrenamiento) sobre el que se quiere obtener una aproximación a una determinada función, se intenta asignar a los coeficientes sinápticos un conjunto de valores que produzcan la mejor aproximación posible, bajo algún criterio de error.

En nuestro trabajo se utilizarán RNAS alimentadas hacia adelante. Una RNAS alimentada hacia adelante es una RNAS cuyos elementos de procesamiento están organizados en capas o niveles sucesivos, de forma que, una vez ordenadas las capas de izquierda a derecha, solo existen conexiones entre nodos de niveles sucesivos, en el sentido izquierda-derecha. Uno de los modelos más utilizados dentro de esta clase de RNAS es el perceptrón multinivel.

El Perceptrón Multinivel

Un perceptrón multinivel, o perceptrón multicapas, es una Red de Neuronas Artificiales alimentada hacia adelante con tres o más capas de neuronas. Las capas situadas entre la primera y la última reciben el nombre de capas ocultas o capas intermedias. La primera capa, llamada capa de entrada, consta de p nodos correspondientes a un vector de entradas $(x_1, x_2,..., x_p)'$. La última capa, llamada capa de salida, consta de q nodos, cada uno de los cuales produce una salida y_j, por lo que la salida completa de la red es un vector $y=(y_1,y_2,...,y_q)'$.

El objetivo del perceptrón multinivel es el de aproximar una función $\phi : A \subseteq R^p \rightarrow R^q$ que a cada posible vector p-dimensional del conjunto origen asigne un vector imagen q-dimensional. La aproximación se basa en el entrenamiento o aprendizaje de la red a partir de un conjunto de n ejemplos o patrones de entrenamiento $(x^{(r)}, z^{(r)})$, donde $z^{(r)} = \phi(x^{(r)})$, $r=1,2,...,n$. En la figura 1 se representa de forma gráfica un perceptrón multinivel con tres capas. Los nodos representados mediante círculos transmiten a cada uno de los nodos de la siguiente capa un valor constante igual a 1.

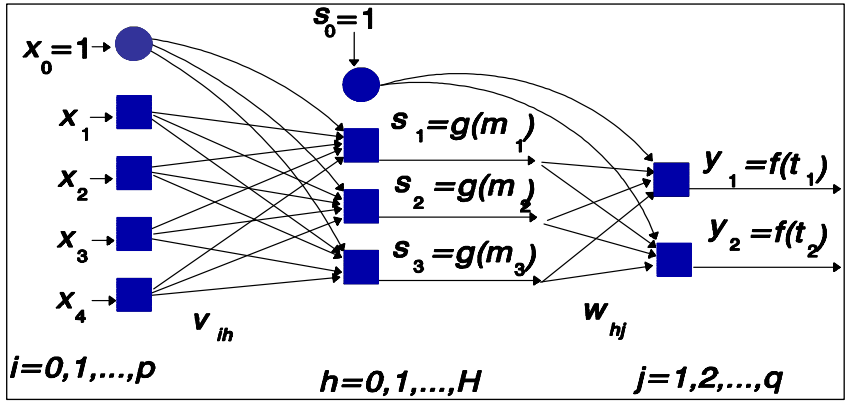

Figura 1: *Perceptrón multicapas con tres niveles.*

En este esquema, H denota el número de nodos ocultos, $\{v_{ih},\ i=0,1,2,...,p,\ h=1,2,...,H\}$ son los coeficientes sinápticos asociados a las interconexiones entre los nodos de entrada y los nodos ocultos, y $\{w_{hj},\ h=0,1,2,...,H,\ j=1,2,...,q\}$ son los coeficientes sinápticos asociados a las interconexiones entre los nodos ocultos y los nodos de salida. La salida de cada nodo oculto, s_h, $h=1,2,...,H$, se obtiene mediante la aplicación de una función de activación, g, a la correspondiente combinación lineal,

$$m_h = v_{0h} + \sum_{i=1}^{p} v_{ih} x_i ,$$

es decir, $s_h=g(m_h)$. Análogamente, los valores asociados a cada nodo de salida, y_j, $j=1,2,...,q$, se obtienen mediante una función de activación f, $y_j=f(t_j)$, siendo t_j la entrada neta al nodo de salida j, obtenida como combinación lineal de las salidas resultantes de las neuronas artificiales de la capa oculta:

$$y_j = f(t_j) = f(w_{0j} + \sum_{h=1}^{H} w_{hj} s_h) = f(w_{0j} + \sum_{h=1}^{H} w_{hj}\, g(v_{0h} + \sum_{i=1}^{p} v_{ih} x_i))$$

Esta última expresión muestra claramente que cada una de las salidas de la red, y_j, $j=1,2,...,q$, es una función anidada, en general no lineal, de los p valores que componen el vector de entrada, $(x_1, x_2, ..., x_p)$. Se deduce además que el número total de parámetros, M, para un perceptrón de tres capas viene dado por la expresión

$$M=(p+1)H+(H+1)q=(p+q+1)H+q$$

El uso del perceptrón multinivel viene respaldado por diversos resultados teóricos, entre los cuales destaca la propiedad de aproximador universal, (Bishop, 1995), donde se

consideran funciones de activación de tipo sigmoidal en la capa oculta y funciones de activación identidad en la capa de salida.

DEFINICIÓN. Se dice que una función $f : R \to R$ es sigmoidal si verifica

$$-\infty < \lim_{x\to-\infty} f(x) < \lim_{x\to+\infty} f(x) < +\infty$$

Ejemplos de funciones de activación sigmoidales son las siguientes:

La función paso: $g(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$

La función signo: $g(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$

La función logística: $g(x) = \dfrac{e^x}{e^x + 1} = \dfrac{1}{1 + e^{-x}}$

La función tangente hiperbólica: $g(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}} = \dfrac{1 - e^{-2x}}{1 + e^{-2x}}$

Reglas de aprendizaje

DEFINICIÓN. Dada una Red de Neuronas Artificiales, se llama algoritmo, método o regla de aprendizaje a cualquier algoritmo que permita obtener una asignación de valores para cada uno de los coeficientes sinápticos.

En nuestro trabajo, al igual que la mayoría de aplicaciones del perceptrón multinivel, se utilizarán procedimientos de aprendizaje supervisado. En este tipo de aprendizaje, se trata de conseguir que la red sea capaz de predecir, a partir de un conjunto de características suministradas como entradas a la red, el valor que tomarán otras características, llamadas características objetivo, habiendo sido observados ambos tipos de características en un conjunto de casos que recibe el nombre de conjunto de entrenamiento.

La mayoría de algoritmos de aprendizaje intentan minimizar el criterio de error cuadrático total que, para un conjunto de entrenamiento *D* y una elección *M*-dimensional *w* de los coeficientes sinápticos, viene definido por la suma de los *nq* residuos cuadráticos:

$$E(D, w) = \sum_{r=1}^{n} \sum_{j=1}^{q} \left(z_j^{(r)} - y_j^{(r)} \right)^2$$

Uno de los problemas prácticos en la utilización de las RNAS es la no existencia de ningún algoritmo de entrenamiento que garantice la convergencia a óptimos globales. En nuestro trabajo se ha empleado el algoritmo de gradientes conjugados, uno de los dos algoritmos de entrenamiento disponibles en SPSS Neural Connection v 1.0. En general, se ha observado una clara superioridad de dicho algoritmo en relación a la otra regla, la

regla delta generalizada con momento. Los textos sobre RNAS suelen incluir amplias descripciones de éstos y otros algoritmos de aprendizaje (Bishop, 1995; Ripley, 1996).

Imputación de la intención de voto en una encuesta electoral

Las encuestas electorales suelen presentar casos donde no se conoce la intención de voto explícita, en general por la negativa del encuestado a responder. Es por ello necesario disponer de mecanismos de imputación que permitan estimar la intención de voto de aquellos encuestados para los cuales, por el motivo que sea, se ignore. En este apartado se describe una aplicación del perceptrón multinivel como modelo de imputación de la intención de voto sobre una encuesta real.

En concreto, se considera la encuesta del Centro de Investigaciones Sociológicas sobre la situación política y social de Andalucía en febrero de 1995. La imputación de la intención de voto en las siguientes elecciones autonómicas puede plantearse como un problema de predicción donde a partir del conocimiento del resto de variables (variables independientes), se desea obtener una estimación del valor que tomará la variable de interés (variable dependiente). El cuadro 1(*a* y *b*) contiene la lista de variables utilizadas para predecir la intención de voto en las siguientes elecciones autonómicas. La tabla 1 recoge las posibles respuestas a la cuestión Intención de voto.

Cuadro 1a: *Variables sociodemográficas*.

Provincia
Tamaño de hábitat
Edad
Sexo
Nivel de estudios
Situación laboral
Relación laboral
Tipo de empresa

Los valores perdidos fueron incluídos en la categoría NS/NC. El problema de la imputación de los valores perdidos puede considerarse como un problema de predicción donde a partir del conocimiento de las variables independientes se desea estimar el valor que tomará la cuestión a imputar. Una vez definido el problema de predicción que se desea acometer mediante el perceptrón multinivel, debe señalarse la necesidad de codificar cada una de las variables independientes y la propia variable dependiente mediante variables auxiliares 0/1, dado que las entradas y salidas de dicho modelo de RNAS son números reales.

De los 1506 casos donde la respuesta a la intención de voto es conocida y distinta de “No sabe/No contesta”, 900 casos, elegidos aleatoriamente, definieron el conjunto de entrenamiento, utilizado para estimar los coeficientes sinápticos mediante el algoritmo de los gradientes conjugados.

Cuadro 1b: *Variables relacionadas con aspectos políticos*.

Ideología política
Recuerdo de voto: elecciones generales (junio 1993)
Aprueba o desaprueba la labor de la Junta de Andalucía
Aprueba o desaprueba la labor del PP
Aprueba o desaprueba la labor del IU
Aprueba o desaprueba la labor del PA
Valoración si gobernase PSOE con mayoría absoluta
Valoración si gobernase PP con mayoría absoluta
Valoración si gobernase PSOE en coalición/apoyo de IU
Evolución de la actuación de IU en Andalucía
Evolución de la actuación de PP en Andalucía
Evolución de la actuación de PSOE en Andalucía
Evolución de la actuación de PA en Andalucía
Partido que mejor defiende los intereses de Andalucía
Partido que mejor representa las ideas de la gente como Ud.
Partido que más confianza le inspira
Partido al que está más unido
Partido que tiene mejores líderes en Andalucía
Partido más capacitado para gobernar en Andalucía
Recuerdo de voto: elecciones autonómicas (junio 1994)
Partido por el que siente más simpatía o más cercano a sus ideas

Tabla 1: *Categorías en la cuestión 'Intención de voto'*.

Código	Descripción
1	AP/PDP
2	CDS
3	PCE
4	PA
5	PSOE
6	Otro dcha.
7	Otro izqda.
90	No votará
999	NS/NC

Sin embargo, para medir el rendimiento que cabe esperar para la red, se requiere un conjunto de casos aparte, al que se le llama conjunto test. En nuestra aplicación, los 606 casos restantes se utilizaron como conjunto test. Así, se construyó un modelo de predicción/imputación basado en el perceptrón multinivel con las siguientes características:

- 3 capas
- 132 nodos de entrada
- 8 nodos de salida
- Función de activación logística en la capa oculta
- Función de activación identidad en la capa de salida
- Tamaño de la capa oculta: 100
- Tamaño del conjunto de entrenamiento: 900
- Tamaño del conjunto test: 606
- Algoritmo de entrenamiento: gradientes conjugados.

La experiencia encontrada en este tipo de aplicaciones del perceptrón multinivel sugiere la necesidad de considerar una capa oculta compuesta por un número elevado de neuronas artificiales, a diferencia de otras aplicaciones como la predicción univariante de series temporales, donde la capa oculta puede ser sensiblemente más reducida (Cubiles de la Vega y otros, 2001).

Las tablas 2 y 3 muestran el espectacular rendimiento de la red construida, que es capaz de predecir, con el 100% de acierto, tanto en el conjunto de entrenamiento como en el conjunto test, la intención de voto a partir de las restantes variables recogidas, definiendo así un modelo de imputación claramente satisfactorio.

Tabla 2: *conjunto de entrenamiento para la predicción mediante perceptrón multinivel*.

	Predicción de intención de voto								
Intención de voto observada	AP/PDP	CDS	PCE	PA	PSOE	Otro dcha.	Otro izqda.	No votará	Total
AP/PDP	111								111
CDS		34							34
PCE			77						77
PA				53					53
PSOE					464				464
Otro dcha.						7			7
Otro izqda.							12		12
No votará								142	142
Total	111	34	77	53	464	7	12	142	900

Detección y corrección de inconsistencias en la encuesta electoral

En el contexto de la encuesta electoral del CIS ya descrita en el apartado 3, nos planteamos la construcción de un modelo basado en el perceptrón multinivel que fuese capaz de detectar e incluso corregir respuestas incorrectas a la cuestión de intención de voto. Un modelo así podría ser útil además, en esta aplicación concreta, a la hora de intentar desvelar la intención de voto real, en algunas ocasiones voluntariamente escon-

dida o falseada por parte del encuestado. Para ello, se requiere un conjunto de entrenamiento apropiado, donde debe aparecer como variable a predecir la intención de voto correcta, mientras que como variables predictoras se tendrán las diversas variables recogidas, incluyendo la intención de voto manifestada, por tanto no corregida.

Tabla 3: *conjunto test para la predicción mediante perceptrón multinivel.*

	Predicción de intención de voto								
Intención de voto observada	AP/PDP	CDS	PCE	PA	PSOE	Otro dcha.	Otro izqda.	No votará	Total
AP/PDP	114								114
CDS		18							18
PCE			67						67
PA				31					31
PSOE					275				275
Otro dcha.						2			2
Otro izqda.							4		14
No votará								95	95
Total	114	18	67	31	275	2	4	95	606

En principio el inconveniente de esta técnica reside en la necesidad de editar previamente los registros a fin de detectar y corregir las posibles inconsistencias. Nordbotten (1995) sugiere para ello el empleo de expertos humanos, si bien realizando el trabajo una vez, de modo que el modelo de edición que se construya basado en el perceptrón multinivel sea posteriormente un mecanismo automático aplicable en posteriores realizaciones de la encuesta.

En nuestro trabajo, sin embargo, hemos optado por la introducción deliberada de inconsistencias, a falta de medios (sobre todo temporales) para realizar una tarea así. Suponiendo que los registros disponibles son correctos, hemos procedido a realizar, con probabilidad 0.1, más alta de lo que sugieren otras experiencias (Nordbotten, 1995) un intercambio de respuestas en aquellos casos donde la intención de voto manifestada era uno de los 5 partidos principales. En los casos seleccionados, la intención de voto se cambia por una de las otras cuatro opciones, elegida a su vez aleatoriamente. Así, el perceptrón multinivel construido responde a las siguientes características:

- 3 capas
- 151 nodos de entrada
- 8 nodos de salida
- Función de activación logística en la capa oculta
- Función de activación identidad en la capa de salida
- Tamaño de la capa oculta: 100
- Tamaño del conjunto de entrenamiento: 900
- Tamaño del conjunto test: 606
- Algoritmo de entrenamiento: gradientes conjugados.

Los resultados que aparecen en las tablas 4 y 5 son muy alentadores. En el conjunto de entrenamiento, el 98,7% de los 75 casos con inconsistencias son corregidos correctamente. Además, sólo un 0,7% de los 825 casos correctos son convertidos en registros inconsistentes. Más aún, y lo que es más importante, en el conjunto test estos porcentajes son del 97,8% y 2,1%, por lo que en definitiva el perceptrón multinivel construido se revela como un modelo muy satisfactorio en la tarea de edición de la cuestión de intención de voto.

Tabla 4: *Casos con inconsistencia en el conjunto de entrenamiento*.

Inconsistencias		Después		
		No	Sí	Total
Antes	No	819	6	825
	Sí	74	1	75
	Total	893	7	900

Tabla 5: *Casos con inconsistencia en el conjunto test*.

Inconsistencias		Después		
		No	Sí	Total
Antes	No	545	16	561
	Sí	44	1	45
	Total	589	17	606

Conclusiones y líneas futuras

Se ha descrito un modelo de imputación, detección y corrección de inconsistencias basado en el perceptrón mutlinivel. La aplicación realizada sobre una encuesta electoral real, refererida a la intención de voto, ha mostrado un rendimiento satisfactorio. Sin embargo, el desarrollo más efectivo de esta línea de trabajo sugiere actuaciones futuras como las siguientes:

- Construcción de modelos de imputación y edición de registros capaces de trabajar con registros completos (en vez de una sola cuestión). El esfuerzo de computación se eleva considerablemente, requiriéndose medios informáticos potentes, sobre todo para cuestionarios de cierta complejidad.
- Estudio de los tamaños necesarios para el conjunto de entrenamiento y el conjunto test, y medidas del error asociado.
- Utilización de otras arquitecturas de RNAS, como las Redes de Base Radial.

- Comparación con técnicas alternativas (otros métodos de imputación, metodología de Fellegi-Holt, sistemas expertos, etc), o incorporación en el proceso de ideas más elaboradas como, por ejemplo, la imputación múltiple (Morales, 2000).

En definitiva, consideramos que las RNAS ofrecen un campo de investigación prometedor en cuanto al desarrollo de modelos de edición de encuestas, como los propios organismos oficiales reseñados en la introducción han puesto de manifiesto también.

Agradecimientos

Los autores agradecen al Instituto de Estudios Sociales Avanzados de Andalucía (CSIC) la colaboración prestada en la realización de este artículo.

Este trabajo ha sido financiado por el Instituto de Estadística de Andalucía, proyecto de investigación "Edición de registros estadísticos mediante redes de neuronas artificiales", código 18.07.02.58.01.

Referencias

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Cubiles de la Vega, M.D, Pino Mejías, R., Moreno Rebollo, J.L., Muñoz García, J. (2001). A Neural Network model for predicting time series with interventions and a comparative analysis. *Journal of Official Statistics* (aceptado para su publicación en el vol. 4 de 2001).

Eurostat (2000). *Editing and Imputation in Eurostat.* Working paper nº21, UN/ECE Work Session on Statistical Data Editing, Conference of European Statistics.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 17-35.

Morales, L. (2000) El efecto de la no respuesta parcial en el análisis de datos de encuesta: una comparación entre la eliminación de observaciones y la imputación múltiple. *Metodología de Encuestas*, 2 (2) 217-238.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks. *Journal of Official Statistics*, 11, 391-411.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics*, 12 (4), 385-401.

Nordbotten, S. (1999). *Evaluating Efficiency of Statistical Data Editing: General Framework*. Material metodológico de la Conferencia de Estadísticos Europeos de 1999.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rumelhart, D. E., Widrow, B., Lehr, M. A. (1994). Neural Networks: Applications in Industry, Business and Science. *Communications of the ACM*, (37)3, 93-105.

Statistics Denmark (1999). *Error Identification and imputations with neural networks.* Working paper nº26, UN/ECE Work Session on Statistical Data Editing, Conference of European Statistics.