APPENDIX B

DATA SETS AND PERTURBATIONS

REPORT ON RESULTS OF EVALUATION

Heather Wagstaff

Office for National Statistics

TABLE OF CONTENTS

Appendix B - Datasets & Perturbations

B.1. Rationale for Selecting the Datasets

B.1.1. Background

The EUREDIT Project aims to develop and evaluate methods for edit and imputation. In order to compare methods objectively, one of the key objectives of the project was: *"To establish a standard collection of data sets for EUREDIT"*. This standard collection will be used to evaluate each of the methods tested. This paper describes the rationale behind the choice of data sets to be included in this standard collection.

In the early stages of the Euredit Project a questionnaire was circulated to all Partners requesting information on potential data sets that they could provide for the project. The results from these questionnaires were summarised and presented at the first project meeting. The initial choice was then made through discussion of the results at this meeting.

B.1.2. Rationale

When deciding upon the datasets to be used in the EUREDIT project, the initial aim was to select 5 datasets that provided a broad and representative coverage of datasets that would be typical of NSIs and other potential users of edit and imputation methods. These datasets would obviously need to be suitable for the evaluation of edit and imputation techniques and cover a range of data sources, such as social surveys, business surveys, time series, censuses and registers. Within each dataset a range of error types would be required, allowing the data to exhibit inconsistencies, non-response (item and unit), outliers and missingness.

In all cases the use of a particular dataset can be justified by three major factors:

1. The dataset is of a type relevant to end-users of the Euredit Project but sufficiently different to the other datasets chosen for the project.

2. The pattern of missingness and errors in the pre-edited dataset can be recreated in the datasets distributed to the partners in the project for evaluation purposes.

3. The availability of edit rules to allow the participants to check for inconsistencies in the dataset.

Table B1.1 below summarises the main aspects of each dataset selected. This displays the wide variation in types and sizes, both in terms of number of variables and number of records, of dataset being used in the Euredit Project, thus highlighting, in general terms, the reasoning behind the choice of these particular datasets.

Table B.1.1 - Summary of datasets selected for EUREDIT

Dataset Name	Type of Data	Type of Variables	No. of Variables	No. of Records
1 Danish Labour Force Survey	Administrative records with pattern of missingness from social survey	Continuous Ordinal Nominal.	14	15579
2 Sample of Anonymised Records from U.K. 1991 Census	Population Census	Categorical Ordinal	35	494024
3 UK Annual Business Inquiry	Business Inquiry Questionnaire	Mostly continuous (rounded to nearest £1000 sterling) 1 nominal (industry classification).	35	9580
4 Swiss Environment Protection Expenditures	Environmental Questionnaire	Continuous (rounded to nearest 1000 CF) Binary Categorical.	70	1239
5 German Socio-Economic Panel Survey	Panel Survey	Nominal Ordinal Continuous	169	5383
6 Times Series for Financial Instruments	Time Series	Continuous	124 time series	522 in each time series

The data sets are described in detail in this Appendix from Chapter B.4 onwards.

B.2. Variants of the Euredit Datasets

B.2.1. Development Datasets

For some methods, particularly neural networks, there is a need to estimate parameters from clean data. In real life situations such networks would learn from data that had been meticulously manually edited – usually a previous survey of the same type or a sample of the actual data. In order to develop and test prototype systems, six development datasets based on a small subset of each original dataset were provided for use with these methods.

Table B.2.1.1 displays the notation used to describe the different versions of any single dataset. It should be noted that, in the context of the Euredit Project, a missing value is not an error. Missing values are easily identified in the data and are the targets for imputation.

Table B.2.1.1 – Notation to describe versions of datasets

Errors?	**Missing?**	
	Yes	**No**
Yes	Y_3	Y_1
No	Y_2	Y^*

However, each of the datasets were available in three versions:

- True data (Y*)
- Data with missing values but no errors (Y_2)
- Data with both errors and missing values (Y_3)

For the purposes of the EUREDIT evaluations, 'true data' means data that the NSI provider considered to be satisfactorily cleaned by their edit and imputation procedures. One could also consider this as 'target data'.

No development dataset was provided containing a Y_1 version, errors but no missing values, as discussion at a Euredit Work Session decided that this would not represent a realistic situation.

Table B.2.1.2 - Versions of Euredit Datasets

Version	**Euredit Dataset**					
	LFS	ABI	SARS	EPE	GSOEP	Time Series
Y*	✓	✓	✓	✓	✓	✓
Y_2	✓	✓	✓	✓	✓	✓
Y_3	✗	✓	✓	✓	✗	✓

Partners were advised that they could use the development datasets in any way they saw fit for the purposes of developing software or expertise in the use of software packages for edit and imputation. However, this usage was subject to the agreed confidentiality constraints. In addition Partners could use any additional non-Euredit dataset to which they had access. This included application of the error generation software to create additional versions of the datasets and use of the evaluation software to assess results

B.2.2. Evaluation Datasets

Similarly to the development datasets, a series of six evaluation datasets was produced. The Danish Labour Force Survey and German Socio-Economic Panel Survey datasets each have only two versions, Y*, Y_2, as they are to be used solely for imputation. The other four datasets have three versions: Y*, Y_2, Y_3, where Y_2 and Y_3 have different observation numbers for individual records to prevent potential disclosure of errors.

The Y* data were retained by ONS, with a subset released for the training of neural networks, and the perturbed data, Y_2, Y_3, were distributed to Partners for edit and imputation.

B.3. Generation of Errors and Missingness

B.3.1. Perturbation of 'true' data

Errors and missingness were imposed on the 'true' datasets by a perturbation process. Thus datasets were created which simulated observed patterns of errors and missingness to facilitate the development and evaluation of methods. The DLFS, EPE data were perturbed by the Danish and Swiss NSI's and similarly the Time Series data by QANTARIS. The ABI, SARS and GSOEP were perturbed by ONS using the Data Perturbation Program which is briefly described below.

The Data Perturbation Program (DPP) simulates errors in a data set using methods developed from the ISTAT ESSE Program. A number of different methods of perturbation are available and the user may specify the conditions for any number of the methods. The DPP supports a range perturbation types, dependent on the type of variable under consideration, these are described in Tables B.3.1.1 - 3 below..

Table B.3.1.1 Methods for categorical and numeric variables

Perturbation Type	Description
Missing completely at random (MCAR)	Where a missing value of variable *y* does not depend on the value of any other variable.
Misplacement errors	The value entered is taken from an adjacent variable, providing the variables are of the same type. For categorical data, misplacement errors only occur when the two variables have the same number of categories.
Interchange of values	The transposition of, or repetition of, digits that occur in the true value. If the true value only consists of one digit then it is replaced by a random value.
Routing errors	When all the conditions are true the value is set to missing, otherwise it is set to a random value in the permitted range
Interchange errors	The true value is replaced by a random value within the permitted range.

Table B.3.1.2 Methods for numeric variables only

Perturbation Type	Description
Loss or addition of zeros	Miss-keying of data so that there are either extra or insufficient zeros. For example, 10 may be keyed as 100 or 10,000 may be keyed as 1,000. This only applies for values ending in a zero.
Under reporting	The supplied value is less that the true value.
Outlier generation	The generation of extreme values.
£'000 errors	This occurs when the units of the variable are ignored. For example, the value of £68,000 is entered as such when £68 should have been keyed. The user can specify different probabilities of such an error occurring for different ranges of the true value.
Transcription error	Simulates a transcription error in a user specified range.
Addition and subtraction of a random number	Simulates an error where a business wrongly includes an item of expenditure in one part of the questionnaire and excludes it from another. The random number is in a range specified by the user and must be no larger than the value from which it is subtracted. The value is added to one or more specified variables and subtracted from one or more other specified variables within a case.

Table B.3.1.3 Global methods (any data type)

Perturbation Type	Description
Keying errors	Digits are changed at random. This applies to all digits of all numbers and is applied after any of the non-global methods. The probability applies to each digit.
Interchange between rows (within a household)	This applies only to hierarchical data sets, for example where some of the details for one individual within a household have been interchanged with those of a different individual within the same household. In this case the values for a set of specified variables for one case are interchanged with those for the same set of variables for a different case with the same household identifier. The probability is proportional to the number of individuals in the household.

B.3.2. Banding errors

Certain methods can rank identified "errors" in order of importance or likelihood of being a true error. In real life a decision is taken on cost grounds as to how many of the potential errors are to be "corrected" by imputation or re-approaching the respondent. In the Euredit Project if this were left to each experimenter to decide for themselves it would be possible that the results of evaluation could depend on where the cut-off is formed between what constitutes an error or not. For this reason a series of bands has been provided so that these methods can be compared on an equal footing. The banding of errors for the edit datasets is described here. The records in the lower numbered bands are more likely to fail more than one edit rule and these are more likely to be the fatal, or hard, rules.

B.4. Danish Labour Force Survey (DLFS)

B.4.1. Overview of the dataset

The Danish Labour Force Survey (DLFS) is a rotating panel survey which samples approximately 15,600 individuals each quarter from the Danish Central Population Register. Random samples of 5,000 and 10,600 people are selected from two strata, unemployed and employed respectively. The selection of individuals, rather than households, yields estimates with smaller variances. Panel members are interviewed in two consecutive quarters followed by a third interview, a year after the second.

B.4.2. File pre-processing

There are no missing values for records in the original Population Register. Contact details, obtained from the Labour Force Survey, were added for each record and the value of income was set to missing for cases of unit non-response in the survey. Thus the missingness was created by using real non-response for those individuals who did not respond to the survey and hence is missing at random (MAR). The linkage and pattern of missingness also facilitates comparison of true and imputed values..

B.4.3. Characteristics of the data file

The data files are supplied as comma separated variables (csv) containing 15,579 records and 14 variables. There is a single row for each record and variable names appear on the first row. Only INCOME has missing values. Apart from AGE, all other variables are categorical.

Tables B.4.3.1 List of Variables in DLFS sample

Mnemonic	Variable	Values	Data Type
NUMBER	Case Number	1 to 15579	Identifier
RESPONSE	Telephone/postal interview or neither	1 = A telephone interview or postal response 0 = Otherwise	Nominal
PANEL	No. of times interviewed	1 = Interviewed for first time 2 = Interviewed for second time 3 = Interviewed for third time	Ordinal
PHONE	Telephone number. Found and contact made	0 = Telephone number found and haven't tried to call the respondent 1 = Telephone number found and have tried to call the respondent	Nominal
LETTER	Postal follow up	1 = After trying to call the respondent and not getting an interview, Dst have sent a postal follow-up 0 = Otherwise	Nominal
SEX	Male/female	1 = Male 2 = Female	Nominal
AGE	Age of respondent	15 to 67	Continuous
MARRIAGE	Marital Status	0 = Not Married 1 = Married	Nominal
EDUCATION	Longest education	1 = Primary school only 2 = Craftsman, Skilled labour, High school only 3 = Long education, schoolteacher, university etc. -9 = No information	Ordinal
BUSINESS	Last employment	1 = Private Industry 2 = Other private business 3 = Government employed -9 = Not applicable	Nominal
UNEMPLOY	Current employment	0 = Not registered as unemployed 1 = Registered as unemployed	Nominal

Tables B.4.3.1 List of Variables in DLFS sample (continued)

Mnemonic	Variable	Values	Data Type
CHILDREN	Any children at home	0 = No children below 19 years living at home 1 = 1 or more children below 19 years living at home	Nominal
COHABITE	Living with another adult	0 = Not living with another adult 1 = Living with another adult	Nominal
AREA	Area of residence	11 = Metropolitan Area (Centre) 12 = Metropolitan Area 13 = Metropolitan Area 14 = Metropolitan Area (Outskirts) 21 = Provincial City 22 = Provincial City 23 = Provincial City 24 = Provincial City (Smallest) 31 = Rural area (with some villages) 32 = Rural Area 33 = Rural Area 34 = Rural area (without villages)	Nominal
INCOME (DK)	Income before tax	If less than 0 then truncated to 0 If 0-500,000 then rounded to nearest 10,000 If 550,000- 900,000 then rounded to nearest 50,000 If over 900,000 then truncated to 900,000	Continuous (with some rounding)
INCOME (DK2)	Income before tax	If less than 0 then truncated to 0 If more than 1,000,000 then truncated to 1,000,000	Continuous (with some rounding)

B.4.4. Perturbation characteristics

Only INCOME contains missingness based on the unit non-response to the DLFS. Some 15,579 individuals were identified as members of the survey sample, of which 11,404 responded, giving a non-response rate of 26.8%. No edit rules were applied.

B.4.5. Key characteristics of the evaluation dataset

The evaluation dataset was created for use only with imputation methods.

B.5. UK Census 1991, Sample of Anonymised Records (SARS)

B.5.1. Overview of the dataset

The SARS dataset consists of a 1% sample of household records from the 1991 UK Census that have been anonymised to avoid disclosure. The dataset is hierarchical in terms of households and persons within households. The Euredit dataset contains 31 of the SARS variables.

The 1% sample was created by taking a 10% simple random sample of households. The selected households were ordered heirarchically by U.K.administrative area then a 10% systematic random sample was taken. The sampling design ensures that the 1% household SAR approximates to a simple stratified random sample of households whilst minimising the risk of disclosure.

B.5.2. File pre-processing

The perturbed data were prepared using the NAG perturbation program.

B.5.3. Characteristics of the data file

The file consisted of comma separated variables, of which the first row contained variable names. The evaluation data dataset contains 492,024 records whilst the development dataset contains 47,870 records. Any 'not applicable' responses are represented by '-9'. Missing values created for the Y2 and Y3 datasets are represented by a blank cell. Note that the variables *distwork* and *isco2* are not consecutively numbered.

Table B.5.3.1 - Variables in the SARS Datasets

Question Number	Mnemonic	Description	Categories
		SAR area	1: Northern England 2: Yorks and Humberside 3: East Midlands 4: East Anglia 5: Inner London 6: Outer London 7: Rest of South East England 8: South West England 9: West Midlands 10: North West England 11: Wales 12: Scotland
H1	Roomsnum	Number of rooms	1 to 15+

Table B.5.3.1 - Variables in the SARS Datasets (continued)

Question Number	Mnemonic	Description	Categories
H2	Hhsptype	Household type	1: detached 2: semi-detached 3: terraced 4: flat-residential 5: flat-commercial 6: converted flat 7: converted flatlet 8: not self contained (s/c) flat 9: not s/c rooms 10: not s/c bedsit 11: other not s/c flat 12: other not s/c rooms 13: other not s/c bedsit 14: non-permanent structure
H3	Tenure	Tenure of household	1: owns outright 2: owns buying 3: private rented (furnished) 4: private rented (unfurnished) 5: rented with job/business 6: rented (housing association) 7-10: rented public sector (7 England & Wales 8-10: Scotland)
H4	Bath	Bath or shower?	1: exclusive use 2: shared 3: none
H4	Insidewc	Inside WC?	1. exclusive use 2. shared 3. none
H4	Cenheat	Central heating	1: all rooms 2: some rooms 3: none
H5	Cars	Number of cars	0 to 3+
2	Sex	Sex	1: male 2: female
3	Age	Age	0 to 90 91 (91 to 92) 93 (93 to 94) 95 (95 +)
4	Mstatus	Marital Status	1: single, 2: married, 3: remarried, 4: divorced, 5: widowed

Table B.5.3.1 - Variables in the SARS Datasets (continued)

Question Number	Mnemonic	Description	Categories	
5	Relat	Relationship to household head	0: household head 1: spouse 2: cohabitee 3: son/daughter 4: child of cohabitee 5: son/daughter in law 6: cohabitee of son/daughter 7: parent 8: parent in law 9: brother/sister 10: brother/sister in law 11: grandchild 12: nephew/niece 13: other related 14: boarder/lodger 15: joint head 16: other unrelated	
6	Residsta	Whereabouts on census night	1: present resident 2: absent resident 3: visitor	
7	Urvisit	Usual residence of visitor	1 to 12 as for Household SAR area 13: outside GB -8: not stated -9: not applicable	
8	Termtim	Term time address of students	1: this address 2: elsewhere in region 3: elsewhere not stated 4: elsewhere out of region -9: not applicable	
9		Area of former residence	1-12 as for SAR area 13: Outside GB -8: not stated -9: not applicable	
10	Cobirth	Country of birth	1: England 2: Scotland 3: Wales 4: Northern Ireland 5: Other UK 6: Irish Republic 7: Australia 8: Canada 9: New Zealand 10: Kenya 11: Nigeria 12: Uganda 3: Other Africa Cmwth 14: Jamaica 15, Other Caribb Cmwth, 16: Bangladesh 17: India 18: Pakistan 19: Sri Lanka 20: Hong Kong 21: Malaysia	22: Singapore 23; Cyprus 24:Gib/Malta/Gozo, 25: Other New Cmwth, 26: France 27: Germany 28: Italy, 29: Spain 30:Bel/Den/Lux/Neth, 31 Portugal/Greece 32: Poland, 33:Al/Bu/Cz/Hu/Ro/Y, 34: Other Europe 35: Turkey/U.S.S.R. 36: South Africa 37: Other Africa 38: U.S.A. 39: Other America 40: Middle East 41: Other Asia 42: Rest/sea/air.

Table B.5.3.1 - Variables in the SARS Datasets (continued)

Question Number	Mnemonic	Description	Categories	
11		Ethnic group	1: White 2: Black Caribbean 3: Black African 4: Black other 5: Indian,	6: Pakistani 7: Bangladeshi 8: Chinese 9: Other-Asian 10: Other-other
12	Ltill	Long term illness	1: yes, 2: no	
13		Employment status	1: employee, 2: manager 3: supervisor 4-5: self-employed -9 = not applicable	
14	Hours	Hours worked weekly	0 to 70 71: 71 to 80 81: 81 +	
15	Econprim	Occupation	1 to 358 -9 = not applicable	
17	Workplce	Workplace	1: at home 2: inside SAR area 3: outside SAR area 4: outside GB 5: inside GB, not stated -8: not stated -9: not applicable	
18		Transport to work	0: works at home 2: British Rail train 3: other rail 4: motor cycle 5: car – driver 6: car – passenger 7: pedal cycle 8: on foot 9: other -8: not stated -9: not applicable	
19	Qualnum	No. of higher ed. Qualifications	1 to 2 2 = 2<=	
19	Qualevel	Level of highest qualification	1 to 3 -9 = not applicable	
19	Qualsub	Subject of highest qualification	1 to 88 -8 = not stated -9 = not applicable	

B.5.4. Perturbation characteristics

Mnemonic	Y3 (for editing and imputation)								Y2 (imputation only)	
	MCAR		Misplacement Errors		Interchange of Values		Interchange Errors		MCAR	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Number										
Hnum										
Areahh	28331	5.73					29162	5.90	29734	6.02
Bath	38137	7.72	22926	4.64	22314	4.52			39534	8.00
Cenheat	40972	8.29	26237	5.31	25900	5.24	27237	5.51	44462	9.00
Insidewc	29140	5.90			28744	5.82			29432	5.96
Cars	49251	9.97			0	0.00			48865	9.89
Hhsptype	33702	6.82			37830	7.66			34360	6.96
Roomsnum	47911	9.70			32205	6.52			49164	9.95
Tenure	34626	7.01							34538	6.99
Persinhh										
Pnum										
Age	37468	7.58					33879	6.86	39620	8.02
Cobirth	38349	7.76			37664	7.62			39291	7.95
Distwork	11958	2.42			9910	2.01			12357	2.50
Empstat									20167	4.08
Hours	15649	3.17					19354	3.92	16816	3.40
Ltill	33670	6.82			28593	5.79			34695	7.02
mstatus	48120	9.74			27894	5.65			49571	10.03
migorgn	3569	0.72					3095	0.63	3678	0.74
occpatn										
qualnum	28801	5.83			28606	5.79			29461	5.96
qualevel	3164	0.64			3192	0.65			3194	0.65
qualsub	3224	0.65			3084	0.62			3310	0.67
relat	28327	5.73					29303	5.93	29613	5.99
residsta	37413	7.57					29271	5.93	39405	7.98
sex	33633	6.81			33282	6.74			34711	7.03
soclass	16478	3.34					20102	4.07	16993	3.44
segroup	16384	3.32					20081	4.06	17474	3.54
termtim	5447	1.10					52271	10.58	10654	2.16
urvisit	838	0.17			515	0.10			798	0.16
workplce	12111	2.45			14119	2.86			12469	2.52
isco										
onsclass										
econprim									23746	4.81

B.5.5. Key characteristics of the evaluation dataset

Due to the heirarchical nature of the data, the data file is large in order to allow sufficient donors for the imputation process. A consequence of this is that processing time can be long for standard methods.

B.6. Annual Business Inquiry (ABI)

B.6.1. Overview of the dataset

The Euredit datasets are samples taken from the retail section of the 1997 and 1998 U.K. Annual Business Inquiry (ABI). For each business sampled the datasets contain information on aspects of purchase costs and employment costs.

Businesses are sent one of 2 questionnaires: a long form which contains 17 retail questions; or a short form which asks for summary information and contains 5 retail questions.

The Euredit datasets contains responses to selected questions from the ABI for 2 sectors and 2 years, 1997 and 1998. A high proportion of the questions are the same for both sectors and years. However there are some differences in the treatment of the "other" categories with more detail in 1998. Sector 2 has a question on excise duties that is not present for Sector 1.

The ABI is conducted in two parts: one dealing with employment; the other with financial information. The financial inquiry covers about two thirds of the UK economy including production; construction; distribution and service industries; the coverage of the employment inquiry is wider.

The ABI dataset is a stratified random sample of about 69,600 businesses taken from the register of legal units. The ABI population is stratified by the Social Iindustrial Classification 1992 (SIC92) and employment using the information from the register. The sampling scheme is designed to give best estimates of the population totals for a given sample size and involves selecting all the largest businesses with a progressively reducing fraction of smaller businesses. This method ensures the sample size is kept to a minimum.

The register used for the ABI is the Inter-Department Business Register (IDBR), which consists of companies, partnerships, sole proprietorships, public authorities, central government departments, local authorities and non-profit making bodies. The main administrative sources of the IDBR are HM Customs and Excise for Value Added Tax (VAT) details and the Inland Revenue for Pay AsYou Earn (PAYE) taxation details.

The original file used for Euredit consists of 9582 records containing: mainly continuous variables which have been rounded to the nearest £000; and two categorical variables Industry Classification derived from SIC92 and size of industry in terms of the number of employees.

B.6.2. File pre-processing

Since the short form asks only summary information, the responses to the additional 12 questions on the long form are set to '-9'. The ABI data contains a high proportion of responses which are totals summed from other responses.

B.6.3. Characteristics of the data file

Details of the variables in each dataset and for both forms are shown in Table B.6.3.1.

Table B.6.3.1 Data items across ABI data sets

Mnemonic	Variable	Sec1 1997		Sec1 1998		Sec2 1997		Sec2 1998	
		Long	Short	Long	Short	Long	Short	Long	Short
REF	Unique Identifier	Y	Y	Y	Y	Y	Y	Y	Y
CLASS	Anonymised hierarchical industrial classification: the first digit indicates the higher level of classification. Numerically adjacent digits do not indicate similarity between classes/subclasses e.g. 3.1 is similar to 3.2 but not necessarily closer to 3.2 than 3.3.	Y	Y	Y	Y	Y	Y	Y	Y
PARTNO	Partition number for experimentation	Y	Y	N	N	N	N	N	N
WEIGHT	Design weight (N/n) for category (1st digit of CLASS) and employment size band (EMPREG)	Y	Y	Y	Y	Y	Y	Y	Y
TURNOVER	Total turnover	Y	Y	Y	Y	Y	Y	Y	Y
EMPWAG	Wages and salaries paid	Y	N	Y	N	Y	N	Y	N
EMPNI	Employers NI contributions	N	N	Y	N	N	N	Y	N
EMPNIOTH	Employers NI contributions and other employment costs	Y	N	N	N	Y	N	N	N
EMPENS	Contributions to pension funds	N	N	Y	N	N	N	Y	N
EMPRED	Redundancy and severance payments to employees	N	N	Y	N	N	N	Y	N
EMPTOTC	Total employment costs	Y	Y	Y	Y	Y	Y	Y	Y
PUREN	Purchases of energy, water and materials	N	N	Y	N	N	N	Y	N
PURENOTH	Purchases of energy and other goods for own consumption	Y	N	N	N	Y	N	N	N
PURCOTH	Purchases of other goods and materials for own consumption	N	N	Y	N	N	N	Y	N
PURESALE	Purchases of goods bought for resale	Y	Y	Y	Y	Y	Y	Y	Y
PURHIRE	Payments of hiring, leasing or renting	Y	N	Y	N	Y	N	Y	N
PURINS	Commercial insurance premiums paid	Y	N	Y	N	Y	N	Y	N
PURTRANS	Purchases of road transport services	Y	N	Y	N	Y	N	Y	N
PURTELE	Purchases of telecommunication services	Y	N	Y	N	Y	N	Y	N
PURCOMP	Purchases of computer and related services	Y	N	Y	N	Y	N	Y	N
PURADV	Purchases of advertising and marketing	Y	N	Y	N	Y	N	Y	N
PUROTHSE	Other services purchased	Y	N	Y	N	Y	N	Y	N
PUROTHAL	All other purchases of goods and services	N	Y	N	Y	N	Y	N	Y
PURTOT	Total purchases of goods and services	Y	Y	Y	Y	Y	Y	Y	Y
TAXRATES	Amounts paid for national non-domestic rates	Y	N	Y	N	Y	N	Y	N
TAXDUTY	Amounts paid for export duty	N	N	N	N	Y	Y	Y	Y
TAXOTHE	Other amounts paid for taxes and levies	Y	N	Y	N	N	N	N	N
TAXOTHD	Other amounts for taxes and levies excluding duty	N	N	N	N	Y	N	Y	N
TAXTOT	Total taxes paid	Y	Y	Y	Y	Y	Y	Y	Y
STOCKBEG	Value of stocks held at beginning of year	Y	Y	Y	Y	Y	Y	Y	Y
STOCKEND	Value of stocks held at end of year	Y	Y	Y	Y	Y	Y	Y	Y
ASSACQ	Total cost of all capital assets acquired	Y	Y	Y	Y	Y	Y	Y	Y
ASSDISP	Total proceeds from capital asset disposal	Y	Y	Y	Y	Y	Y	Y	Y
CAPWORK	Value of work of a capital nature	Y	Y	Y	N	Y	Y	Y	N

Table B.6.3.1 Data items across ABI data sets (continued)

Mnemonic	Variable	Sec1 1997		Sec1 1998		Sec2 1997		Sec2 1998	
		Long	Short	Long	Short	Long	Short	Long	Short
EMPLOY	Total number of employees	Y	Y	Y	Y	Y	Y	Y	Y
TURNREG	Registered turnover	Y	Y	Y	Y	Y	Y	Y	Y
EMPREG	Employment size group from register: 1 = 0 to 9 2 = 10 to 19 3 = 20 to 49 4 = 50 to 99 5 = 100 to 249 6 = 250+	Y	Y	Y	Y	Y	Y	Y	Y
FORMTYPE	1 = long form 2 = short form	Y	Y	Y	Y	Y	Y	Y	Y

Table B.6.3.2 displays the number of records and variables in the 12 Euredit ABI data files.

Table B.6.3.2 - ABI data files number of records and variables

Sector	Year	File Type	Number of Records	Number of Variables
1	97	Y *	6099	31
		Y2	6099	30
		Y3	6099	30
	98	Y2	6233	33
		Y3	6233	33
2	97	Y *	4325	31
		Y2	4325	31
		Y3	4325	31
	98	Y *	5594	34
		Y2	5591	34
		Y3	5594	34

Data provided were collected in two years: 1997 and 1998. The former were to be used for training purposes in the neural network experiments. The 1998 data (Sec1 1998 Y2 and Sec1 1998 Y3) were used for evaluation purposes. Files contained comma separated variables with a single row of data for each record, with variable names on the first row. Including the unique identifier there were 38 variables of which 32 were perturbed.

There were only two categorical variables, the rest being continuous and measured to the nearest £000's. Missing data were indicated as: '-9' for not applicable (long form) with blank cells for missing values.

B.6.4. Perturbation characteristics

The perturbation process ensured that if an individual value is changed then the corresponding total may or may not have been changed appropriately. Derived variables have not been perturbed. Table B.6.4.1 displays the perturbation characteristics of the ABI data. The cell values are percentages.

Table B.6.4.1 - Perturbation characteristics of ABI data

Mnemonic	Long Form	Short Form	Perturbation Method					
			Missing (set to blank)	Random add or subtract from variables *	Random addition or subtraction from variables & adjust corresponding total	Add to one variable & subtract from another	VAT (deduct amount & add to other taxes)	Swap
REF	Y	Y						
CLASS	Y	Y						
WEIGHT	Y	Y						
TURNOVER	Y	Y	4	1 (long); 0.6 (short)				
EMPWAG	Y	N		0.6 (also applies to variables that appear in the short question'ire)	4			4 (not apply to variable 19 and 20)
EMPNI	Y	N						
EMPENS	Y	N						
EMPRED	Y	N						
EMPTOTC	Y	Y						
PUREN	Y	N			4	3	2 (applies only to variables 12 and 20 in the short form)	
PURCOTH	Y	N						
PURESALE	Y	Y						
PURHIRE	Y	N						
PURINS	Y	N						
PURTRANS	Y	N						
PURTELE	Y	N						
PURCOMP	Y	N						
PURADV	Y	N						
PUROTHSE	Y	N						
PUROTHAL	N	Y						
PURTOT	Y	Y						
TAXRATES	Y	N			4	3		
TAXOTHE	Y	N						
TAXTOT	Y	Y						
STOCKBEG	Y	Y						
STOCKEND	Y	Y						
ASSACQ	Y	Y						
ASSDISP	Y	Y						
CAPWORK	Y	N						
EMPLOY	Y	Y		1				
TURNREG	Y	Y						
EMPREG	Y	Y						
FORMTYPE	Y	Y						

* Zero values are replaced by a number, the absolute value of which is < £1 million.

B.6.5. Key characteristics of the evaluation data set

For the ABI data, the variables of most importance are the totals as well as a variable relating to turnover of the business. These variables are:

	Variable name
Turnover	Total turnover
Emptotc	Total employer costs
Purtot	Total purchases of goods and services
Taxtot	Total taxes paid
Assacq	Total costs of all capital assets acquired
Assdisp	Total proceeds from capital asset disposal

The majority of the other variables are subdivisions of these and are therefore of secondary importance. Many are part of the short form and not the long form.

B.7. Swiss Environment Protection Expenditure Survey (EPE)

B.7.1. Overview of the dataset

In 1992/1993 the Swiss Federal Statistical Office (SFSO) carried out four pilot surveys to obtain information on the Environment Protection Expenditure (EPE) of the public sector as well as the private sector. SFSO gave a mandate to INFRAS, a private consulting firm, to develop the appropriate definitions and carry out the survey in close co-operation with SFSO. Three surveys were carried out in the public sector. They covered the federal, the cantonal and the communal factors. One survey was carried out in the private economy: this last survey is the source of the SFSO data for Euredit. The methodology of the surveys is described in a technical report by INFRAS: *Umweltschutzausgaben in der Schweiz, Technischer Bericht zur Piloterhebung, INFRAS 21. August 1995*. In the present documentation the technical report by INFRAS shall be referred as ITR. The results of the surveys are published in "Umweltausgaben und -investitionen in der Schweiz 1992/1993, Ergebnisse einer Pilotstudie" (1996), SFSO.

The dataset consists of information on expenditure relating to environmental issues which was collected by a questionnaire distributed to enterprises in Switzerland in 1993. These were chosen according to class of economic activity.

Characteristics of the original sample
The population unit is the enterprise. The sample is a subset of the Swiss enterprises chosen according to class of economic activity. Furthermore the main survey was carried out for enterprises with 20 or more employees. However in some activity classes a small sample of enterprises with less than 20 employees was chosen. Some 1239 enterprises participated.

The sampling weights were not registered on the original data set but were reconstructed. The sample design is a two way stratification: the first is according to economic activity (actcl); the second according to size of the enterprise in 5 classes (*sizecl*) measured by number of employees. The classes are:

Size of Enterprise (*sizecl*)	**Number of Employees**
0	0-19
1	20-49
2	50-99
3	100-249
4	250 +

The original sample design did not cover size class zero for which the population sizes are unknown. There are also cases with unknown industry or size class. The population sizes have been taken from ITR, page 190/191. However, industry 11 (*ENERGIE*) was not documented and the number of enterprises per size class was taken from the enterprise census of 1991. The sheet weights gives the net sample sizes (ns) and population sizes per stratum (np). Where the stratification is known the weight is np/ns. Where only industry is known, i.e. for size class 0, an overal weight of the activity class was imputed. Thus the weight for size class zero is the population size of the activity class divided by the total sample in the activity class. Where neither activity class nor size class is known (2 cases) the weight is set to 1.

The sum of the weights is 19,968.53. We cannot calibrate this number to a population number since the population size is not documented in ITR. More precisely, the population in size class 0 is unknown. We could calibrate on the number of employees in the population: 2,615,842.92 by calibrating the estimated number of employees on this figure. The estimated number of employees in the population is 2,592,536. Thus an overall correction factor of 1.00899 would result. Since it is close to 1 and since a more precise calibration is done with the ratio weights, we did not apply this correction factor to the weights.

Ratio Weights
Ratio estimators have been applied at the level of aggregates of strata, called groups. The group indicator is *grind*. The matching with the stratum indicator *stratind* and the English names of the groups are on sheet Strata. The weight correction due to the ratio estimator, *gweight*, is the estimate of the total of number of employees (sampling weighted total of variable *empl*) divided by the corresponding number of employees in the population. The population number of employees for the groups is known in theory but looking at the strongly varying g-weights there might have been more enterprises in the population than covered by the survey in some groups.

Aggregate Values
The Analysis Groups are aggregates of the poststratification groups, i.e. of the groups used for the ratio estimators. They are documented in the worksheet *Groups*. The matching of *grind* with the code for the analysis groups *gcode* is also on the worksheet *Groups*.

B.7.2. File pre-processing

The questionnaire asked for the exact expenditures or, in the absence of exact amounts, for a good guess of the amount. Just below the space where to fill in the amount, the respondent could check a box, indicating that the amount filled in was in fact a guess. This was taken as the missing value indicator and the corresponding values were left blank. In that sense there was no stochastic mechanism imposed: it is the actual questionnaire that created the missingness. Outliers were created arbitrarily, with some being representative outliers and some non-representative.

B.7.3. Characteristics of the data file

The true data set consists of comma separated variables that contain responses to an environmental questionnaire with 70 variables. These are responses to the questionnaire plus additional general business questions. In the file there are 1239 records each consisting of 64 variables containing information obtained form the questionnaire and also general business questions. All monetary values are rounded to the nearest 1000 CF.

The initial column is a reference number (I.d.) variable. Among the 54 numerical variables 35 are independent; the other 19 can be derived from them and therefore lead to the creation of 23 edit rules. Five more edit rules arise from the binary variables and *exp93*, finally giving a total of 28 edit rules.

The reference numbers for the Y_2 and Y_3 datasets are different. To allow comparison between Y_2 and Y_3 data with the additional records, and the clean additional records, the codes for these are given in *codes200.xls*. The codes for the additional records correspond to those for the Y_3 data. 200 clean records were provided for training purposes for processing by neural networks. The dataset file contains raw data including errors, outliers and missing values.

The overall the unit response rate is 56%. There are a small number of cells in the *act* and *actcl* variable where a missing value is represented by '-9' otherwise there are no true missing values but a large number of values are estimated rather than being observed.

The responses to the variable exp93 shows that 520 enterprises have not filled in the questionnaire because they didn't have any EPE at all (code 2). This means that is a high proportion of true zero values in the data.

Table B.7.3.1 Data items in EPE data set

Mnemonic	Details	Responses	Data Type
Exp93	Any expenditure in 1993	1 = Non zero expenditure 2 = Zero total expenditure 3 = No response	Categorical
Netinv	Any net-investment in 1993	Yes / No	Binary
Eopinvwp	End of pipe investment: water protection	CF (to nearest 1000)	Continuous
Pininvwp	Process integrated investment: water protection	CF (to nearest 1000)	Continuous
Othinvwp	Other Investment: water protection	CF (to nearest 1000)	Continuous
Totinvwp	Total investment: water protection	CF (to nearest 1000)	Continuous
Eopinvwm	End of pipe investment: waste treatment	CF (to nearest 1000)	Continuous
Pininvwm	Process integrated investment: waste treatment	CF (to nearest 1000)	Continuous
Othinvwm	Other Investment: waste treatment	CF (to nearest 1000)	Continuous
Totinvwm	Total investment: waste treatment	CF (to nearest 1000)	Continuous
Eopinvap	End of pipe investment: air protection	CF (to nearest 1000)	Continuous
Pininvap	Process integrated investment: air protection	CF (to nearest 1000)	Continuous
Othinvap	Other Investment: air protection	CF (to nearest 1000)	Continuous
Totinvap	Total investment: air protection	CF (to nearest 1000)	Continuous
Eopinvnp	End of pipe investment: noise protection	CF (to nearest 1000)	Continuous
Pininvnp	Process integrated investment: noise protection	CF (to nearest 1000)	Continuous
Othinvnp	Other investment: noise protection	CF (to nearest 1000)	Continuous
Totinvnp	Total investment: noise protection	CF (to nearest 1000)	Continuous
Eopinvot	End of pipe investment: Other	CF (to nearest 1000)	Continuous
Pininvot	Process integrated investment: Other	CF (to nearest 1000)	Continuous
Othinvot	Other investments: Other	CF (to nearest 1000)	Continuous
Totinvot	Total investment: Other	CF (to nearest 1000)	Continuous
Eopinvtot	End of pipe investment: Total	CF (to nearest 1000)	Continuous
Pininvtot	Process integrated investment: Total	CF (to nearest 1000)	Continuous
Othinvtot	Other investment: Total	CF (to nearest 1000)	Continuous
Totinvto	Total investment: Total	CF (to nearest 1000)	Continuous
Curexp	Any current expenditures in 1993	Yes / No	Binary
Curexpwp	Current expenditure: water protection	CF (to nearest 1000)	Continuous
Taxexpwp	Taxes: water protection	CF (to nearest 1000)	Continuous
Totexpwp	Total expenditure: water protection	CF (to nearest 1000)	Continuous
Curexpwm	Current Expenditure: waste management	CF (to nearest 1000)	Continuous
Taxexpwm	Taxes: waste management	CF (to nearest 1000)	Continuous
Totexpwm	Total expenditure: waste management	CF (to nearest 1000)	Continuous
Curexpap	Current Expenditure: air protection	CF (to nearest 1000)	Continuous
Taxexpap	Taxes: air protection	CF (to nearest 1000)	Continuous
Totexpap	Total expenditure: air protection	CF (to nearest 1000)	Continuous
Curexpnp	Current expenditure: noise protection	CF (to nearest 1000)	Continuous
Taxexpnp	Taxes: noise protection	CF (to nearest 1000)	Continuous
Totexpnp	Total expenditure: noise protection	CF (to nearest 1000)	Continuous
Curexpot	Current Expenditure: other	CF (to nearest 1000)	Continuous
Taxexpot	Taxes: other	CF (to nearest 1000)	Continuous
Totexpot	Total expenditure: other	CF (to nearest 1000)	Continuous

Table B.7.3.1 Data items in EPE data set (continued)

Mnemonic	Details	Responses	Data Type
Curexptot	Current expenditure: total	CF (to nearest 1000)	Continuous
Taxexptot	Taxes: total	CF (to nearest 1000)	Continuous
Totexpto	Total current expenditure: total	CF (to nearest 1000)	Continuous
Subsid	Any subsidies in 1993	Yes / No	Binary
Subwp	Subsidies: water protection	CF (to nearest 1000)	Continuous
Subwm	Subsidies: waste management	CF (to nearest 1000)	Continuous
Subap	Subsidies: air protection	CF (to nearest 1000)	Continuous
Subnp	Subsidies: noise protection	CF (to nearest 1000)	Continuous
Subot	Subsidies: other	CF (to nearest 1000)	Continuous
Subtot	Subsidies: total	CF (to nearest 1000)	Continuous
Receipts	Any receipts in 1993	Yes / No	Binary
Recwp	Income: water protection	CF (to nearest 1000)	Continuous
Recwm	Income: waste management	CF (to nearest 1000)	Continuous
Recap	Income: air protection	CF (to nearest 1000)	Continuous
Recnp	Income: noise protection	CF (to nearest 1000)	Continuous
Recot	Income: other	CF (to nearest 1000)	Continuous
Rectot	Income: total	CF (to nearest 1000)	Continuous
Deliv	Delivered	1 to 3	Categorical
Id	Identifier	1 to 1239	
Act	Economic activity	36 categories	Nominal
Emp	Number of employees	2 to 42822	Continuous
Sizecl	Size class (no of employees)	0 = 0-19 1 = 20-49 2 = 50-99 3 = 100-249 4 = 250+	Ordinal
Nuts	Region	0 to 7	Categorical
Actcl	Original industry. These are the analysis groups - see 'grind' below	28 groups	Categorical (non numeric)
Lang	Original language	1 = German 2 = French 3 = Italian	Categorical
Stratind	Stratum indicator	11-14, 21-24, 31-34, 41-44, 52-54, 61-63, 71-73, 80,83,84, 91-94, 101-103, 110-114, 120-123, 131-134, 141-144, 151-154, 161, 163, 171-174 181-184, 191-193, 202-204, 210-214, 220-224, 231-234, 241-244, 251-254, 261-264, 271-273, 281-284	Categorical
Weight	Survey weight, determined by stratind	1 to 353.75	Continuous

B.7.4. Perturbation characteristics

Table B.7.4.1 displays the level of missingness for each variable in the EPE data set.

Table B7.4.1 - Distribution of missingness in the EPE data set

Mnemonic	Missing	
	n	%
Exp93	-	-
Netinv	-	-
Eopinvwp	67	5.41
Pininvwp	75	6.05
Othinvwp	28	2.26
Totinvwp	66	5.33
Eopinvwm	62	5.00
Pininvwm	61	4.92
Othinvwm	33	2.66
Totinvwm	70	5.65
Eopinvtot	75	6.05
Pininvtot	69	5.57
Othinvtot	34	2.74
Totinvto	109	8.8
Curexp	-	-
Curexpwp	146	11.78
Taxexpwp	65	5.25
Totexpwp	111	8.96
Eopinvap	62	5
Pininvap	61	4.92
Othinvap	24	1.94
Totinvap	70	5.65
Eopinvnp	42	3.39
Pininvnp	35	2.82
Othinvnp	15	1.21
Totinvnp	43	3.47
Eopinvot	22	1.78
Pininvot	23	1.86
Othinvot	27	2.18
Totinvot	36	2.91

Mnemonic	Missing	
	n	%
Curexpwm	189	15.25
Taxexpwm	77	6.21
Totexpwm	145	11.7
Curexpap	122	9.85
Taxexpap	11	0.89
Totexpap	89	7.18
Curexpnp	35	2.82
Taxexpnp	2	0.16
Totexpnp	29	2.34
Curexpot	69	5.57
Taxexpot	9	0.73
Totexpot	53	4.28
Curexptot	0	0
Taxexptot	0	0
Totexpto	208	16.79
Subsid	-	-
Subwp	0	0
Subwm	3	0.24
Subap	1	0.08
Subnp	0	0
Subot	1	0.08
Subtot	4	0.32
Receipts	-	-
Recwp	5	0.4
Recwm	78	6.3
Recap	9	0.73
Recnp	0	0
Recot	5	0.4
Rectot	53	4.28

Note that the frequencies have been calculated over all 1239 units, but in fact the 520 cases corresponding to *exp93=2* have no missing values, therefore the frequencies could have been computed only on 719 units and their values would have been increased by a 1.723 factor.

B.7.5. Key characteristics of the evaluation dataset

Since there are a lot of outliers in the data an ad hoc robust ratio estimator has been used. Furthermore there are lots of zero values in the data. Therefore Hulliger and Kassab (1998*) Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes*, Technical Report, Swiss Federal Statistical Office, developed the robust estimators on the parallel survey with communes further. Other estimation procedures might be more appropriate. The population totals of number of employees for the ratio calibration groups are given in sheet Groups. The totals refer to the population of enterprises with more than 20 employees but for the Euredit Project the same totals are used for the whole population.

Table: B.7.5.1 - Characteristics of the EPE evaluation dataset

Mnemonic	Details	Responses	Data Type
Grind	Group indicator/industry, equivalent to OSORTZUS	1 = Energy 21,22 = Construction 31 = Transport 41,42 = Chemistry 51 = Concrete 61 = Brickworks 71 to 74 = Stones 81 = Galvanics 91 to 93 = Metal 101 = Machinery 111 = Paper 121 to 134 = Wood 141 = Textile 151 to 152 = Graphics 161 to 162 = Plastics 171 to 172 = Repairs 181 = Electronics 191 = Watches 201 = Other Industry 211 = Finance 221 = Trade 231 = Other Services 241 = Foundry	Nominal
Gweight	Group weight applied for ratio estimation. Survey weighted total of ANZBEU divided by number of employees in the population.	0.123837 to 67.7875	Continuous

B.8. German Socio-Economic Panel Survey (GSOEP)

B.8.1. Overview of the dataset

The dataset is a selection from the German household survey. There are three initial identification variables and for each year, the individual and household income. In addition, there are 30 education and employment variables for each participant. Also there are 2 variables that will be identical for each year, but are included for clarification purposes.

The data are for people who participated in the survey over the years 1991 to 1996. All persons taking part in wave 1 of the survey (respondents as well as their children) are to be surveyed also the next year: at the same address as well as after a move within Germany (covering residential mobility).

Since all persons are to be personally interviewed once they reach the age of 16, the next generation is automatically taken into account (demographic development).

This dataset represents a closed system with respect to new partners. If a person has a partner in 1991, this partner will usually remain in the survey for the 6 years, unless they dropped out. In which case no information other than their i.d. is given. New partners, not already part of the survey, are given partner numbers but no additional information.

Persons and households which could not be successfully interviewed in a given year are followed until there are two consecutive temporary drop-outs of all household members or a final refusal. In case of a successful interview after a drop-out there is also a small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status).

Two types of information are collected:

1. Standard Components (measured (bi yearly)

- Demography and Population
- Labour Market and Occupation
- Income, Taxes, and Social Security
- Housing
- Health
- Household Production
- Education, Training, and Qualification
- Basis Orientation (preferences and values), Participation, and Integration

2. Special topical modules of the distributed data

1991	Wave 8 (A,B)	Family and social services
	Wave 2 (C)	Family and social services (shortened version plus repetition of subjective indicators and labour market indicators of wave 1 base questions)
1992	Wave 9 (A,B)	Social security and poverty (partly repetition Wave 4)
	Wave 3 (C)	Social security and poverty, labour market indicators and biographical information (Bio)
1993	Wave 10 (A,B)	Further education or training (short. repetition Wave 6)
	Wave 4 (C)	Further education or training, labour market
1994	Wave 11/5	Neighbourhood, values, and expectations
	Wave 1 (D1)	Same as Wave 11/5 plus immigration history and biography
1995	Wave 12 /6	Partial repetition of Wave 7 - use of time and preferences, increased range of income questions
	Wave 1(D2)	Same as Wave 12/6 plus immigration history and biography
1996	Wave 13/7/2	Repetition of social network questions (1991)

This dataset was formed by taking a 95% sample from the original dataset. It represents a closed system with respect to new partners: if a person has a partner in 1991, this partner will usually remain in the survey

for the 6 years, unless they dropped out in which case, no information other than their i.d. is given. New partners, not already part of the survey, are given partner numbers but no additional information. Individual and household weights are shown in the file *weights.xls*

Table B.8.1.1 - Starting Sample Size in Wave 1

Sample	Year	Household (net)	Persons (gross)	Respondents (net)	Children (net)
A and B	1984	5624	15397	11610	3711
C	1990	2071	5818	4229	1510

Key

A "West-German" residents: started in 1984
n=4528 or 4298 households*
Head of household is either German or of another nationality than those in Sample B.

B "Foreigners": started in 1984
n=1393 (for 100% version)or 1326 households (for 95% version)
Head of household is either Turkish, Italian, Spanish, Greek, or Yugoslavian.

C "East-Germans": started in 1990
n=2179 or 2071 households*
Head of household at the time of the survey was a citizen of the GDR.

D "Immigrants": started in 1994/95 in two different subsamples
1994: Subsample D1 with n=236 households
1995: Subsample D2 with n=295 households
Total in 1995 n=522 households (D1 and D2)
At least one household member has moved from abroad to Germany after 1984.

B.8.2. File pre-processing

The perturbed data were created using the NAG perturbation program. If an individual's income was set to missing then the corresponding household income was also set to missing.

B.8.3. Characteristics of the data file

The dataset is hierarchical, with repeated individual records within each household and repeated variables for each of the 6 years for each individual. Hence, each record contains 12 variables: two per year for each of the six years. The data files have 1 row for each record, with the variable names on 1st row. The number of records (individuals) for the training data is 704. That for evaluation is 5383. The files contain comma separated variables.

While identifiers like HHNR, PERSNR, HHNRAKT, SPELLNR, ERHEBJ are not allowed to be missing at all, survey variables might be missing for different purposes. A person can refuse to answer to a question (very often income-related questions) or just might not know an answer. Otherwise a question might not apply to a person or a household; e.g. the rent to be paid when the household is an owner-occupier.

The GSOEP data differentiates three kinds of missing values:

Code	Meaning
-1	no answer / do not know
-2	does not apply
-3	after checking for plausibility a given value was found to be implausible and was finally deleted (thus, this code is to be interpreted like -1)

Note that the GSOEP public use version (95% sample) might use a different coding of the missing values:

Code 100% sample	Code 95% sample
-1	A
-2	B
-3	C

If the person is classed as 'Not Employed' under 'Employment Position', the responses for work related variables (ERLJOB to BRANCH and TREIM to UEBSTD) are 'Not Applicable' and represented by -8.

If the partner indicator suggests 'No Partner', the 'Partner Person Indicator' is 'Not Applicable' and represented by '-8'.

There are a large number of 'Not Applicable' responses for educational qualifications outside Germany and these again are represented by '-8'.

Table B.8.3.1. - Data items in GSOEP data set

Mnemonic	Description		Type
General variables			
Person I.D.	Person number	201 to 5030832	Nominal
Sex	Sex	1 = Male 2 = Female	Nominal
Y.O.B.	Year of birth	1899 to 1975	Ordinal
Variables for each year (the year suffix here = 91: will change for each year)			
The following 2 variables will be identical for each year:			
PERSON ID 91	Person number	201 to 5030832	Nominal
ORIGINAL HHOLD 91	Original household number (From start of survey)	27 to 521787	Nominal
The following variables will take different values for each year			
Income 91	Person income for current year	0 to 739000	Continuous
House income 91	Household income for current year for current year household	0 to 791000	Continuous
HHOLD ID 91	Current year household number	27 to 530859	Nominal
NCHILD91	Number of children in household	0 to 7	Ordinal
ERWTYP91	Employment position	1 = Not employed 2 = Employed	Ordinal
ERLJOB91	Working in occupation trained for	1 = Yes 2 = In Training 3 = No Skills Acquired 4 = No -8 = Not Applicable	Ordinal
BETR91	Size of the company	1=Between 5 and 20 3 = Between 20 and 200 4 = Between 200 and 2000 5 = More than 2000 6 = Less than 5 2 = Self employed alone -8 = Not Applicable	Ordinal
OEFFD91	Civil Service	1 = Yes 2 = No -8 = Not Applicable	Nominal

Table B.8.3.1. - Data items in GSOEP data set (continued)

Mnemonic	Description		Type
AUSB91	Required training for job	1 = No Training 2 = Introduction to job 3 = On the job training 4 = Courses 5 = Vocational training 6 = Technical School 7 = College -8 = Not Applicable	Ordinal
ISCO91	Three digit ISCO occupation code	0 to 999 (see additional info.sheet) -8 = Not Applicable	Nominal
ISCOU91	Two digit ISCO occupation code	0 to 109 (see additional info.sheet) -8 = Not Applicable	Nominal
ISCOH91	One digit ISCO occupation code	0 to 10 (see additional info.sheet) -8 = Not Applicable	Nominal
BRANCH91	Branch of industry - ZUMA	1 to 45 (see additional info.sheet) -8 = Not Applicable	Nominal
PARTZ91	Partner indicator	0 = No partner 1 = Spouse, clearly 2 = Partner, clearly 3 = Spouse, probably 4 = Partner, probably 9 = Partner, who?	Ordinal
PARTNR91	Partner person number	204 to 5307402 = 99999 when partner indicator = 9 -8 = Not Applicable	Nominal
NATION91	Nationality	1 = German 2 = Other	Nominal
PSBIL 91	School-leaving degree	1 = Sec. School degree 2 = Non-class school degree 3 = Tech. School degree 4 = High school degree 5 = Other degree 6 = No school degree	Nominal
PBB01 91	Vocational degree received	1 = Apprenticeship 2 = Vocational school 3 = Healthcare school 4 = Technical school 5 = Civil Service Training 6 = Other training -8 = Not Applicable	Nominal
PBB02 91	Completed college education	1 = Technical college 2 = University technical college 3 = College not in FRG -8 = Not Applicable	Nominal
PBB03 91	No vocational degree	1 = No vocational degree -8 = Not Applicable	Nominal
PSBIA 91	School-leaving degree outside Germany	1 = School, no degree 2 = School with degree 3 = Continuing school -8 = Not Applicable	Nominal

Table B.8.3.1. - Data items in GSOEP data set

Mnemonic	Description		Type
PBBIA 91	Vocational degree outside Germany	1 = On the job training 2 = Vocational training 3 = Vocational school 4 = College 5 = Other Training -8 = Not Applicable	Nominal
FAMSTD 91	Marital status in survey year	1 = Married 2 = Married, separated 3 = Married (Spouse abroad) 4 = Divorced 5 = Widowed 6 = Single	Ordinal
TREIM91	Treiman occupation prestige scale (from ISCO)	0 to 78.90 (See additional info.sheet) -8 = Not Applicable	Continuous
WEGEN91	Wegener occupation prestige scale	0 to 186.80 -8 = Not Applicable	Continuous
BILZEIT 91	Amount education / training (yrs)	7 to 18	Continuous
ERWZEIT 91	Length of time with firm	0 to 57.8 -8 = Not Applicable	Continuous
TATZEIT 91	Actual work time per week	0.2 to 80 -8 = Not Applicable	Continuous
VEBZEIT 91	Agreed upon work time per week	0.2 to 74 -8 = Not Applicable	Continuous
UEBSTD 91	Overtime per week	0.2 to 23.1 -8 = Not Applicable	Continuous

B.8.4. Perturbation characteristics

This dataset is applied for imputation only.

Year	Percent Missing
1991	31.30
1992	32.25
1993	32.12
1994	30.89
1995	31.12
1996	31.73

Only the variable *income* was perturbed. The perturbation program nominally created 10% missing but there were subsequent, manual edits as follows:

1. If the personal income was missing then the corresponding household income was set to missing.
2. If the household income had been perturbed and was missing then the personal income for each member of the household was also set to missing.

B.8.5. Key characteristics of the evaluation dataset

This dataset is for imputation only. The data are hierarchical and has values for the same variable (income) over six successive years.

B.9. Times Series for Financial Instruments

B.9.1. Overview of the dataset

The dataset consists of a number of daily time series for financial instruments and indices between the beginning of 1995 and the end of 1999. It contains prices for various financial instruments of four major financial markets from different countries (USA, United Kingdom, Japan and Germany), as well as a variety of indices. The information includes UK and US shares, exchange rates, financial indices, UK Options and UK Gilts. There should be correlation between the instruments and indices that can be used for edit or imputation. The data will have a trend, which is close to the inflation rate (or perhaps slightly higher), although one must observe a long time series to see the upward trend. All data are logged to base 10.

Six types of indicator are contained in the data. Within each indicator there are a number of time series, containing some missing data. For example in the 'Shares' indicator there are Shares from 15 companies as described below. For each of the other indicators there is also a large amount of time series information. The types of indicator are now described.

B.9.2. Financial Instruments

B.9.2.1. Shares

Ownership of a corporation is represented by shares which are a claim on the corporation's earnings and assets. Shareholders participate in the corporate earnings usually by the yearly paid dividend. Common Stocks usually entitle the shareholder to vote in the election of directors and other matters taken up at the shareholder meeting. Preferred Stocks generally do not confer voting rights, but have a prior claim on assets and earnings. Dividends must be paid on preferred stocks before any can be paid on common stocks.

The shares for a number of USA and UK companies are displayed in Tables B8.1.1 and B8.1.2.

Table B.9.2.1. - Shares for companies in the USA

Name	Symbol	Sector
Delta Airline	DAL	Airline
US Airway	U	Airline
United Airlines	UAL	Airline
Bank America	BAC	Financial services
First Union	FTU	Financial services
American Express	AXP	Financial services
Texas Utilities Co	TXU	Utility
Consolidated Edison Inc	ED	Utility
Columbia Energy Group	CG	Utility

Table B.9.2.2 - Shares for companies in the United Kingdom

Name	Symbol	Sector
British Airways PLC	AWS	Airline
British Petroleum	BP	Oil
Cable and Wireless	CnW	Telecommunication
Glaxo Wellcome	GXO	Pharmaceuticals
Lloyds TSB Group PLC	TSB	Financial services
Rolls Royce	RR	Automobiles

There was a split of shares of BP at 4. Oct 1999 at a ratio of 1:2. The prices before this date are adjusted by that ratio to take this fact into account. The following indicators are tabulated below along with some additional information.

Table B.9.2.3 - Indicators with some additional information.

Exchange Rates	Indices	Derivatives	Bonds
US$ in German marks	GB: FTSE 100	AWS (high1,2 mid1,2, low 1,2)	UKGB202106070800
	German: Dax 30		UKGB200206070700
US$ in Japanese yen	USA: Dow Jones Industrial Index	BP (high1,2 mid1,2, low 1,2)	UKGB200712070725
			UKGB200312070650
GB £ in German marks	USA: NYSE Composite	CnW (high1,2 mid1,2, low 1,2)	UKGB202812070600
			UKGB200912070575
GB £ in Japanese yen	USA: NYSE Transportation	GXO (high1,2 mid1,2, low 1,2)	UKGB200001280850P
			UKGB200406070500
	USA: NYSE Utility	TSB (high1,2 mid1,2, low 1,2)	
	USA: NYSE Finance		
	Japan: Nikkei	RR (high1,2 mid1,2, low 1,2)	

B.9.2.2. Currency Exchange Rates

These are rates at which one country's currency can be converted into another. A wide range of factors influences exchange rates, which change slightly each trading day. Some rates are fixed by agreement.

There are artificial currencies and so artificial exchange rates as well. The ECU was one of the two international currency substitutes, until 1.1.1999 when the Euro became the official European currency, the other being the Special Drawing Rights (SDRs) of the International Monetary Fund (IMF). Since this date all exchange rates between former EU(European Union) currencies, like the French Franc or German Mark, and non-EU currencies are calculated via the Euro and a fixed conversion factor. These national currencies will cease to exist at the end of 2001 when the last stage of the introduction of the Euro will be completed.

The selected exchange rates are those between the currencies of US-Dollar (USD), Japanese Yen (JPY), British Pound (GBP) and the German Mark (DEM). The introduction of the Euro at the 1.1.1999 makes it necessary to transform either the German Mark into Euro or the other way around. As the Euro period is only one fifth of the time range the Euro was transformed into the old currency, DEM.

B.9.2.3. Indices

These are statistical composites that measure changes in the economy or financial markets. In the case of financial markets they measure value changes in representative financial instrument groupings, in particular stocks, bonds and futures.

An average is simply the arithmetic mean of a group of prices (usually weighted by relevant factors), whereas an Index is an average expressed in relation to an earlier established Base Market Value.

Some indices and averages have Sub-Indices, representing a selected group of the index members of a certain business-sector.

The selection of indices contains major indices in the USA (NYSE Composite, Dow Jones Industrial Index), Japan (Nikkei), Germany (DAX 30) and UK (FTSE 100). There are also some sub-indices for the American market, the NYSE – financial, NYSE – transportation and the NYSE - utility index.

B.9.2.4. Derivatives

The chosen derivatives are Call options traded at the LIFFE (London International Financial Futures and Options Exchange) and have as underlyings the UK Shares from section 2. Derivatives are not the series of one instrument but a combination of various short series.

For each underlying UK Share, six time series of derivatives were created, divided into three pairs: One with strike prices 10% below, one 10% above and one with strike prices approximately the price of the underlying three month before maturity (e.g. BP_Clow1, BP_Clow2, BP_Cmid1 and so on).

Options at LIFFE are traded in cycles of quarterly years, e.g. Jan-Apr-Jul-Oct. On the third Wednesday of these months an option matures. The pair of time series is constructed in the way that one contains every second cycle, e.g. Jan-Jul and the other the remaining cycles, e.g. Apr-Oct. In this way they overlap to receive some kind of continuous series of observations.

A derivative instrument (short derivative) is a contract whose value is based on the performance of an underlying financial asset, index or other investment. There are different kinds of derivatives like Call Options, Put Options or Future Contracts.

B.9.2.5. Future Contracts

These are an agreement between two parties to buy or sell a specific amount of a commodity or financial instrument at a particular price on a stipulated future date. Future contracts are normally traded on an exchange. As the two parties to the contract do not necessarily know each other, the exchange provides a mechanism which gives the two parties a guarantee that the contract will be honoured.

Future contracts, for example on currencies exchange rates, are used to eliminate a company's risk in trading with a foreign partner.

B.9.2.6. Option

An Option is in general the right to buy or sell property that is granted in exchange for an agreed upon sum. If the right is not exercised within a specified period, the option expires and the option buyer forfeits the money used to purchase the money. In financial markets the above mentioned property could be any kind of financial instrument like stocks, indices, exchange rates or an other derivative like a future contract. Commodity prices are also a possible underlying for options. Options are traded on many exchanges. There are two main types:

1. A call option gives the buyer the right to buy a certain number of shares of the underlying financial instrument at a fixed price (strike or exercise price) before a specified date in the future (expiration date, exercise date or maturity). For this right, the call option buyer pays the call option seller, called the writer, a fee called a premium, which is forfeited if the buyer does not exercise the option before the agreed-upon date.

2. The opposite of a call option is a put option, which gives the buyer the right to sell a specified number of shares of a financial instrument at a particular price within a specified time period.

American options can be exercised at any time up to the expiration date. European options can be exercised only on the expiration date itself. The terms American and European do not refer to the location of the option or the exchange. Most options traded are American options.

In practice, most call and put options are rarely exercised. Instead, investors buy and sell options before expiration, trading on the rise and fall of the underlying prices.

A second excel spreadsheet (OptionAnnex.xls) describes the composition of each derivative time series using the labelling convention shown in the example below. The Option table in the annex describes the

composition of each time series: on which date an option is included, what kind of option and when it matures. An example of the labelling convention used is:

Example: BP_Apri96C500

BP - Ticker symbol for British Petroleum (see share table)
Apr96 - Maturity date, third Wednesday of April 1996 (corresponds with the end date in the Option table)
C - Call option (*P* for Put)
500 - Strike price 500

The label therefore contains important information that could assist the imputation of the hoes in the derivative series. Please note that the derivative time series are **not** the series of one instrument, but a "puzzle" of short series.

B.9.2.7. Bonds

These are any interest-bearing or discounted government or corporate security that obligates the issuer to pay the bondholder a specified sum of money (coupon), usually at specified intervals, and to repay the principal amount of the loan (corpus) at maturity. Some bonds are callable, which means they are redeemable by the issuer before the scheduled maturity. The issuer must pay the holder a premium price if such a security is retired early.

A bond without the payment of coupons is called a Zero-Coupon bond. To compensate for the lack of steady payments, the zero-coupon bond is issued with a deep (large) discount. Sometimes zero-coupon bonds are generated out of coupon bonds by selling the corpus and the coupons separately (Coupon stripping).

The chosen bonds are also from the UK market, the so-called gilt market. All bonds in the data set are non-callable. The title of the series provides information about the series.

Example: UKGB200611070675

UKGB - United Kingdom Government Bond
2006 - Year of redemption
11 - Month of redemption
07 - Day of redemption
0675 - Coupon of 6.75%

Again, the title will be of assistance for imputation. The coupons of the UK are payable half-yearly on the same dates as the redemption will be. In the example above, coupons will be paid on 7th of November and 7th of May at a rate of GBP3.375 per GBP100 nominal of stock.

Table B.9.x.x. - Instrument descriptions

Type of Series	**Symbol**	**Series Name**	**Further information**
Exchange rates	USD in DEM	US$ in German marks	The introduction of the Euro at the 1.1.1999 makes it necessary to transform either the German Mark into Euro or the other way around. As the Euro period is only one fifth of the time range the Euro was transformed into the old currency, DEM
	USD in JPY	US$ in Japanese yen	
	GBP in JPY	GB £ in Japanese yen	
	GBP in USD	GB £ in US Dollars	

Table B.9.x.x. - Instrument descriptions (continued)

Type of Series	Symbol	Series Name	Further information
Indices	FTSE 100	GB: FTSE 100	The selection of indices contains major indices in the USA (NYSE Composite, Dow Jones Industrial Index), Japan (Nikkei), Germany (DAX 30) and UK (FTSE 100). There are also some sub-indices for the American market, the NYSE – financial, NYSE – transportation and the NYSE - utility index.
	DJIA	USA: Dow Jones Industrial Index	
	NYSE Composite	USA: NYSE Composite	
	NYSE Utility	USA: NYSE Utility	
	NYSE Transportation	USA: NYSE Transportation	
	NYSE Finance	USA: NYSE Finance	
	Nikkei	Japan: Nikkei	
	DAX	German: Dax 30	
UK shares (3rd column shows sector)	AWS	British Airways PLC	Airline
	BP	British Petroleum	Oil
	CnW	Cable + Wireless	Telecommunication
	GXO	Glaxo Wellcome	Pharmaceuticals
	TSB	Lloyds TSB Group PLC	Financial services
	RR	Rolls Royce	Automobiles
US shares (3rd column shows sector)	DAL	Delta Airline	Airline
	U	US Airway	Airline
	UAL	United Airlines	Airline
	BAC	Bank America	Financial services
	FTU	First Union	Financial services
	AXP	American Express	Financial services
	TXU	Texas Utilities Co	Utility
	ED	Consolidated Edison Inc	Utility
	CG	Columbia Energy Group	Utility
Bonds	UKGB199505010300 to UKGB202812070600	UK Government Bond, redemption 1/5/95, 3% UK Government Bond, redemption 7/12/28, 6%	The chosen bonds are also from the UK market, the so-called gilt market. All bonds in the data set are non-callable. The title of the series provides information about the series e.g.: UKGB200611070675 UKGB - United Kingdom Government Bond 2006 - Year of redemption 11 - Month of redemption 07 - Day of redemption 0675 - Coupon of 6.75%

B.9.3. Characteristics of the original sample

The data contain daily closing prices from 2.1.1995 to 30.12.1999 and hence form samples of stockmarket data which fluctuate on a continuous basis. There are some "holes" in the time series given by incomplete data of the providers and national holidays. Missing values are marked as "#NV".

B.9.4. File pre-processing

Outlying observations were generated by a gross error model as described above.

Note that a lowest value of "0" means the numbers cannot be negative, the minimum values found in the data set are generally greater than 0. In the options datasets there are a small number of 0 values. When the data are logged these are shown by #NUM!

B.9.5. Characteristics of the data file

The dataset consists of 124 time series. There are 6 types of indicator with between 4 and 61 individual series each. All files are comma separated. The time series was split into two parts: the first three years where both the missing (Y2), missing with error (Y3), and true values (Y1) datasets were provided, and the final two years of data where only the missing (Y2) and missing + errors (Y3) datasets were provided to partners. The Y1 for the last two years was withheld. All partners, including QANTARIS, could only use these data. Evaluations were performed on the last two years of the time series.

All time series have titles to identify them. These appear as headers for each column in the Excel file as well as file name for the text files. The Excel file consists of various tables representing groups of instruments. These correspond with the folder structure of the text files.

Data files resulting from the merging of shares data (UK and USA) and the UK Gilts (bonds) have 520 observations and 51 variables (columns) plus one header line with variable name. UK Options data files have 520 observations and 36 variables (columns) plus one header line with variable name.

Throughout, for the time series/panel data:

W	Refers to the UK shares data
X	Refers to the US shares data
Y	Refers to the UK Gilts data (bonds)
Z	Refers to the UK options data
WXY	Refers to the horizontally merged W,X and Y data

Both the WXY and Z data sets contain log-return data. These represent the last two years of the 5 years of financial time series/panel data. Missing values in the two true data files, QWXYStar.txt and QZStar.txt, are represented by '-9', and there will be a '-9' in the corresponding position in the missing and imputed data files. missing values in the missing data files are represented by no value in the position.

B.9.6. Perturbation characteristics

All the shares, bonds and options time series in the data QANTARIS supplied (87 in total), have missing values. The auxiliary index variables (12 in total) are complete. These index variables include major exchange rates and major stock indices, and are based on information that is regularly observed in the financial markets, so it is realistic to assume that they are complete.

In order to create the perturbed data, a simple gross error term was added to each of the time series. The values to be perturbed were chosen at random (with probability 1/50). When chosen, their values were multiplied by either 0.1, 0.5, 2, or 10 (this multiplicative factor was also chosen at random). The auxiliary variables were not perturbed.

To create missing observations, again the observations were chosen at random. The probability that a value was chosen was inversely proportional to the value of the VDAX implied volatility index (note that this index has not been supplied with the TS data), thus there is a slight degree of clumping in the missingness pattern. The logic behind doing this was that at times of high trading volumes, time series volatilities increase, and the occurrence of missing values is less frequent. The overall rate of missingness was varied across the time series so that the effect of degree-of-missingness on imputation effectiveness could be examined. The actual missingness rates for each time series are given below.

Table B.9.6.1. - Level of missingness (%)

US Shares		UK Shares		UK Options		UK Gilts	
BAC	9	BP	37	BP_Clow1	44	UKGB199505010300	5
FTU	23	CnW	2	BP_Clow2	3	UKGB199506211025	2
AXP	29	GXO	15	BP_Cmid1	29	UKGB199511151275	13
DAL	73	RR	1	BP_Cmid2	14	UKGB199601221400	29
UAL	5	TSB	7	BP_Chigh1	73	UKGB199605031525	6
U	17	AWS	7	BP_Chigh2	20	UKGB199605151325	56
CG	1			CnW_Clow1	3	UKGB199611151000	9
ED	15			CnW_Clow2	4	UKGB199701221325	1
TXU	15			CnW_Cmid1	18	UKGB199702211050	20
				CnW_Cmid2	2	UKGB199708060700	19
				CnW_Chigh1	13	UKGB199709010875	9
				CnW_Chigh2	9	UKGB199710271500	1
				GXO_Clow1	9	UKGB199801190975	0
				GXO_Clow2	8	UKGB199803300725	11
				GXO_Cmid1	9	UKGB199809301550	9
				GXO_Cmid2	6	UKGB199811201200	12
				GXO_Chigh1	12	UKGB199901150950	2
				GXO_Chigh2	6	UKGB199903261225	10
				AWS_Clow1	10	UKGB199905191050	34
				AWS_Clow2	10	UKGB199908100600	31
				AWS_Cmid1	1	UKGB199911221025	68
				AWS_Cmid2	4	UKGB200001280850	3
				AWS_Chigh1	20	UKGB200003030900	3
				AWS_Chigh2	9	UKGB200007141300	61
				RR_Clow1	35	UKGB200012070800	27
				RR_Clow2	24	UKGB200102261000	6
				RR_Cmid1	3	UKGB200107120950	1
				RR_Cmid2	6	UKGB200108100975	0
				RR_Chigh1	74	UKGB200111060700	1
				RR_Chigh2	10	UKGB200204111000	85
				TSB_Clow1	12	UKGB200206070700	1
				TSB_Clow2	1	UKGB200206140950	4
				TSB_Cmid1	21	UKGB200208270975	4
				TSB_Cmid2	46	UKGB200211190900	8
				TSB_Chigh1	13	UKGB200305070975	32
				TSB_Chigh2	1	UKGB200306100800	1

In all the datasets genuine missing values are represented by -9. When missing values are created for imputation purposes these values are represented by an empty cell.

B.9.7. Key characteristics of the evaluation dataset

An overall criterion for the chosen instruments is that there is at least one obvious interdependency to one of the other instruments, for example the share of "Delta Airline" and the "NYSE – transportation index".

The data patterns of the time series are related to each other in a complex way, so several variables can be used to help predict the others, as well as lagged values of these variables. It is not recommended that naive users try using or implementing complex imputation procedures. One needs not only to consider the key

characteristics of the data, but also the key characteristics of the analysis, and it is the second of these where good statistical expertise becomes necessary. The key points are:

- The time series data have to be continuous
- There has to be a reasonably long history: e.g. a minimum of 200 observations, and preferably more
- To be effective there should be good auxiliary time series available that can be used to model the true values with at least a reasonable degree of accuracy
- The time lag of temporal dependencies in the time series should be fairly limited. Cross-sectional dependencies should be taken advantage of as much as possible.
- If there is a strong non-linear relationship between the value to be imputed and the auxiliary variable, some kind of method should be employed to make this relationship more linear, e.g. by inverting a pricing formula for financial series, or by appropriate pre-transformation of the data, or through the use of a non-linear technique like MLPs
- The assumptions of the statistical model underlying the imputation technique need to be understood and their validity should be checked. In particular, the distributional assumption should be met.
- If the distributional assumption is not met, this could be due to the presence of outliers in the data. Their effect should be curtailed by some appropriate method, e.g. down weighting, Winsorising, removal from estimation, use of robust method, etc. If not, the quality of the imputed values could be seriously degraded.