Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada's Symposium 2005
Methodological Challenges for Future Information Needs

IMPUTING DISTRIBUTIONS IN ADMINISTRATIVE TAX DATA

Rong Huang, Dominique Ladiray[1]

ABSTRACT

Administrative tax data are widely used at Statistics Canada to reduce survey sample sizes and thereby ease the response burden for businesses. While they provide 100% coverage in principle, they have the usual flaws – partial non-response, atypical values, and so on – and have to be analyzed and adjusted.
For example, detailed breakdowns of totals, which are often optional, have to be imputed. The current ratio imputation procedure is based on an *a priori* division of the population of respondent businesses into groups defined by industry code and business size. We propose an alternative method in which the data themselves define homogeneous classes, which are determined by an ascending hierarchical classification of the values of the observed details. The problem is then to assign non-respondent businesses to the appropriate classes. Several methods of doing so are compared, using raw and discretized data: parametric and non-parametric discriminant analyses, log-linear models, etc. The resulting assignment rules are evaluated and compared through simulations and then are validated on the previous year's data.

KEY WORDS: Administrative data, ratio imputation, ascending hierarchical classification, non-parametric discriminant analysis.

1. INTRODUCTION

Statistics Canada made a firm commitment some years ago to use administrative tax data to reduce survey sample sizes and thereby ease the response burden for businesses. In principle, tax data, which are supplied by the Canada Revenue Agency, provide 100% coverage. However, they have the usual flaws – partial non-response, atypical values, and so on – and have to be analyzed and adjusted before they can be used by client divisions. In particular, tax returns tend to emphasize totals rather than detailed breakdowns, which are often optional. When an item in the tax return shows nothing more than the total (or "generic") revenue or expenses, an imputation procedure is used to estimate the breakdown by sub-item ("details").
The current procedure is based on an *a priori* division of the population of respondent businesses into groups defined by industry code (NAICS code) and business size. The marginal distribution of details for respondent businesses within a class is then used for the businesses in the same class that fail to provide details. This ratio estimation process is based on a strong assumption: businesses in the same class are assumed to behave in very similar ways, which means that the average breakdown can be applied to all of them. This can lead to some strange imputations.
We propose an alternative method in which the data themselves define the classes, which are determined by an ascending hierarchical classification of the observed detail values. The next problem is to assign each non-respondent business to an appropriate homogeneous class. Several assignment procedures, based on explanatory variables available in the tax return, are compared using raw and discretized data: parametric and non-parametric discriminant analyses, log-linear models, etc.
The resulting assignment rules are evaluated and compared through simulations using actual data and then are validated on the previous year's data.

2. CURRENT METHOD: DESCRIPTION AND PROBLEMS

The tax return completed by businesses (GIFI) contains some 685 variables, of which some are totals (generics) and others are details. An example of the return's structure is provided in the table below. The *block* is composed of a

[1] Rong Huang, Statistics Canada, Business Survey Methods Division, R.H. Coats Building, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 (e-mail: rong.huang@statcan.ca); Dominique Ladiray, INSEE, Département des comptes nationaux, 15 boulevard Gabriel Péri, 92240 Malakoff, France (e-mail: dominique.ladiray@insee.fr)

generic (code 2700 in grey) and six *details* (codes 2701 through 2706). The taxfiler is free to report whatever level of detail he chooses in the return. If he is uncertain of the detail code, he may decide to include the amount in the generic part. Hence, one taxfiler may choose to fill in a generic code; a second taxfiler may complete the detail codes for the same generic; and a third may fill in both.

Table 1: Generic and details

BT2680	**miscellaneous taxes payable block total amount**
	SHORT TERM DEBT FLDS 2700 TO 2706
2700	short term loan and debt amount
2701	current Canadian bank loan amount
2702	current security sold short liability amount
2703	current security sold under repurchase agreement amount
2704	gold and silver certificate amount
2705	items in transit and check amount
2706	current lien note payable amount

The idea behind this imputation method is quite simple: the generic value is broken down according to the detail distribution observed in the returns of all businesses that report details only.[2]

The table showing the breakdown of the amounts reported by businesses in the same subgroup is a contingency table. We will denote that table as (n_{ij}), $i = 1 \cdots r$, $j = 1 \cdots c$. n_{ij} represents the value reported by business *i* for detail *j*.

We have $n = \sum_i \sum_j n_{ij}$, and the total value reported by business *i* is $n_{i.} = \sum_{j=1}^{j=c} n_{ij}$. Within the subgroup, the total amount reported for detail *j* is $n_{.j} = \sum_{i=1}^{i=r} n_{ij}$.

The basic idea behind the current allocation procedure is to use the observed marginal distribution for businesses that report only details, i.e., $\left\{ \frac{n_{.1}}{n}, \frac{n_{.2}}{n}, \cdots, \frac{n_{.c}}{n} \right\} = \{ f_{.1}, f_{.2}, \cdots, f_{.c} \}$. It is implicitly assumed that the distributions for the various businesses are essentially the same, which makes the marginal distribution a good estimator of the individual distributions.

The "homogeneity" of a subgroup (i.e., the similarity of the row distributions) can be measured with the χ^2 distance (for example, see Saporta, 1990): $d^2 = n \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}}$.

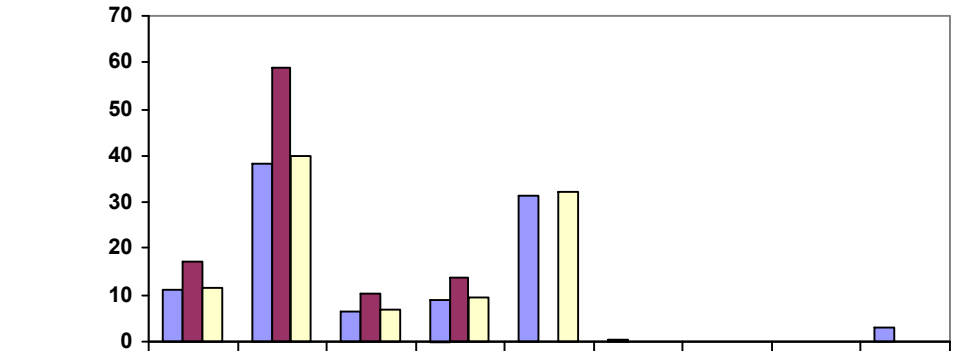
d^2 ranges between 0 and $n \ \inf(r-1, c-1)$. 0 represents total independence between the table's rows and columns, i.e., equality of the detail distributions. The maximum value $n \ \inf(r-1, c-1)$ reflects a perfect functional relation between rows and columns. Unfortunately, it is difficult to compare the values of the d^2 statistics for two different tables, since *n*, *r* and *c* are usually different. Many indicators with a value between 0 and 1 have been proposed in the literature (see Agresti, 1986; Goodman & Kruskal, 1979; SAS Institute, 2001): Pearson's P coefficient, Cramer's V coefficient, Kendall's τ_b, etc.

The homogeneity of the subgroups obtained with the various classifications typically used (12 classifications were studied) was measured with three indicators. For a two-position industry code classification, the three indicators average 0.8, which reflects very different detail distributions.

[2] The algorithm used to calculate the GDA ratios is somewhat more complex in practice (Rondeau, 2003). In particular, the ratios are computed only for subgroups of at least 25 businesses reporting details only; moreover, within a subgroup, only details that have been reported by at least 10% of businesses are considered ("frequent details").

Figure 1 shows, for a given block (9040), the actual marginal distribution and the marginal distributions based on "frequent details" (reported by at least 10% of businesses in the group) and "important details" (the most frequent details, whose cumulative value is greater than 90% of the overall total of the details).

Figure 1: Comparison of the marginal distributions computed with all details, frequent details and important details for block 9130



	1	2	3	4	5	6	7	8	9
Total	11.07	38.33	6.65	9.07	31.15	0.37	0.03	0.1	3.22
Frequent	16.99	58.87	10.21	13.93	0	0	0	0	0
Important	11.49	39.82	6.91	9.42	32.36	0	0	0	0

For block 9130, the "frequent details" distribution omits detail 5, which makes up about 30% of the total, and overestimates detail 2. The use of "important details" eliminates details 6 through 9, which account for less than 4% of the total.

In these circumstances, how can we improve the adjustment quality when the business leaves out the details? Several different methods immediately come to mind.

1. We might consider using donor imputation, i.e., using the distribution of a business that reported only details to impute the details of a business that reported only a generic value. Two difficulties arise:
   - First, the subgroups may have very few members. If so, we run the risk of using the same donor repeatedly.
   - Second, there is the businesses' response behaviour. As shown in Table 6, distributions can vary radically within the same block. If they do and we select a poor donor, we run the risk of imputing a distribution that is farther from the actual distribution than the average distribution.
2. The chart of accounts (COA) developed by Statistics Canada contains fewer details than the tax return, and it is those details which are ultimately important to the client divisions. By reducing the number of details, we in principle increase the homogeneity of the distributions and hence the pertinence of the average distribution in a subgroup.
3. Another possibility, which we developed as this project went along, is to automatically form subgroups in which the detail distributions are similar by design. These subgroups of businesses that report only details are defined by an automatic classification based on the observed distributions. A discriminant analysis is then performed to automatically identify the class to which a business that reports only a generic value belongs. The generic value is then broken down into details using the estimated distribution for the class, such as the marginal distribution.

3. AN ALTERNATIVE METHOD

The principle underlying the proposed method is quite simple and can be summarized in three steps:

- Define homogeneous distribution classes using businesses that report only details in their tax returns.
- Determine the class to which a business that reports only a generic for the block in question belongs. This "inference" is based on the variables available in the business's response, i.e., the generics (sum of the details) or details of the other blocks and certain characteristics of the business (industry, size).

- Estimate the distribution by detail using, for example, the current method, i.e., the marginal distribution based on the businesses in the class.

While this method may seem simple when described in broad terms, implementation requires the setting of many parameters in order to determine the best strategy.

3.1 Looking for the best strategy

The data

The presence of atypical values in the data could affect the results of the analysis, especially since we sometimes use non-robust procedures based on averages (calculating marginal distributions, for example). In looking for the best method to use, for example, in defining the class, we initially worked with discretized data based on quantiles of the distribution in 5, 10, 15, 20, 25 and 30 groups, and with raw data.

Classification

In all cases, classification was performed on the raw data for each block using the contingency table of the amounts (businesses in the rows, details in the columns). Two types of classifications were considered:

1. An ascending hierarchical classification (AHC) using a χ^2 metric (adapted for this type of data) and the Ward strategy.[3] This method presents a number of problems:
   - There may be a large number of businesses in a given block (tens of thousands), and it would be difficult to use an AHC in such cases. An initial non-hierarchical classification was carried out[4] to define 500 classes, on whose centres an AHC was performed.
   - The number of classes has to be specified in advance. We determined the number of classes on the basis of the number of variables. For example, in the case of block 8040, we have three COA variables, and therefore our classification will consist of three or four classes.
2. An *a priori* classification, made necessary by the type of response behaviour of businesses that tend to report few details. Accordingly, we defined some classes, named "AttractorXX", assigning to class *i* businesses that contribute more than XX% of the total amount in detail *i*. For block 8040, we defined the variables Attractor80, Attractor90 and Attractor95. In each case, there is an additional class for all businesses for which no reported detail attains the required XX%.

Selecting the explanatory variables

A business's tax return contains some 300 variables that can be used as explanatory variables for determining the class to which a business belongs. From this set of variables, an initial set of "potentially explanatory" variables was selected using a stepwise parametric discriminant analysis and a stepwise logistic regression.[5] The two lists of candidate variables were then merged to produce a smaller list of explanatory variables.

Selecting the models

Even when the list of explanatory variables is reduced to a few dozen, the potential number of models is huge. Again, we automatically selected potential models (with fewer than 30 explanatory variables) using logistic regression and parametric discriminant analysis.

Evaluating the models

In this final step, each model was estimated by several methods:

- Parametric discriminant analysis
- Non-parametric discriminant analysis using the k closest neighbours method (where k=15 or k=20)
- Linear regression model adapted to the qualitative nature of the variable concerned.[6]

---

[3] SAS PROC CLUSTER after transformation of the data so that applying the Euclidian metric to the transformed data is equivalent to applying a χ^2 metric to the raw data.

[4] SAS PROC FASTCLUS.

[5] Strictly speaking, a logistic regression is not warranted here, since the class number is not a numeric variable. However, we used it simply as a way of "sorting" the variables.

[6] This model was estimated with SAS PROC CATMOD.

In addition, each model was evaluated on the basis of its classification error rate and indicators of concordance between the expected class breakdown and the actual breakdown.

3.2 Initial results

A large-scale simulation based on block 8040 data provided some valuable lessons for the method's implementation.

- The classification error rates are in the range of 15%, which is quite reasonable.
- Linear regression models based on qualitative data are often time-consuming to estimate, essentially because the explanatory variables have rather large numbers of categories.
- Non-parametric discriminant analysis generally performs exceptionally well with models that have few explanatory variables.
- The use of discretized explanatory variables is effective, particularly if the number of groups is quite high. In this case, the effect of atypical values is minimized, and the values still seem "continuous".

3.3 Final evaluation of the method

The study's ultimate objective is to impute the detail distributions. We still need to evaluate the quality of the imputation carried out using the above alternative method and compare it to the quality of the current method.

Test file and evaluation procedure

To that end, we will use a "test file" containing a set of businesses for which the actual detail distribution is known. Test files are commonly used in other fields: tests of statistical algorithms (see NIST, 1998), tests of time series forecasting methods (see Makridakis and Hibon, 2000), and so on.

The test procedure in this case is based on 2001 and 2002 data.

- The 2001 data are used to compute the distributions that will be used to impute the 2002 data.
- For each block studied, businesses that reported only details in 2002 are selected. The values are aggregated to define a "fictitious" generic whose detail distribution will be imputed with ratios from the 2001 file.
- Then the actual and imputed distributions are compared: the best method will be the one with the fewest errors.

The evaluation process is summarized in Figure 2.

Statistical comparison indicators

Statistical indicators are required so that the actual and imputed detail distributions can be compared. They are constructed at the enterprise level (micro level) and at the block level (macro level).

Micro level:

For each business, Pearson's contingency coefficient was computed to measure the distance between the actual detail distribution and the distribution estimated by the imputation method.

Another measure, the "Micro_pseudo_CV", is defined for each business as:

$$Micro\_pseudo\_CV_j = \sqrt{\sum_i (x_{ij} - \hat{x}_{ij})^2} \Big/ \sum_i x_{ij}, \quad j = 1, \cdots n$$

where x_{ij} is the i^{th} detail actually reported by business j and $\hat{x}_{ij}$ is the estimate of the i^{th} detail for business j.

Macro level:

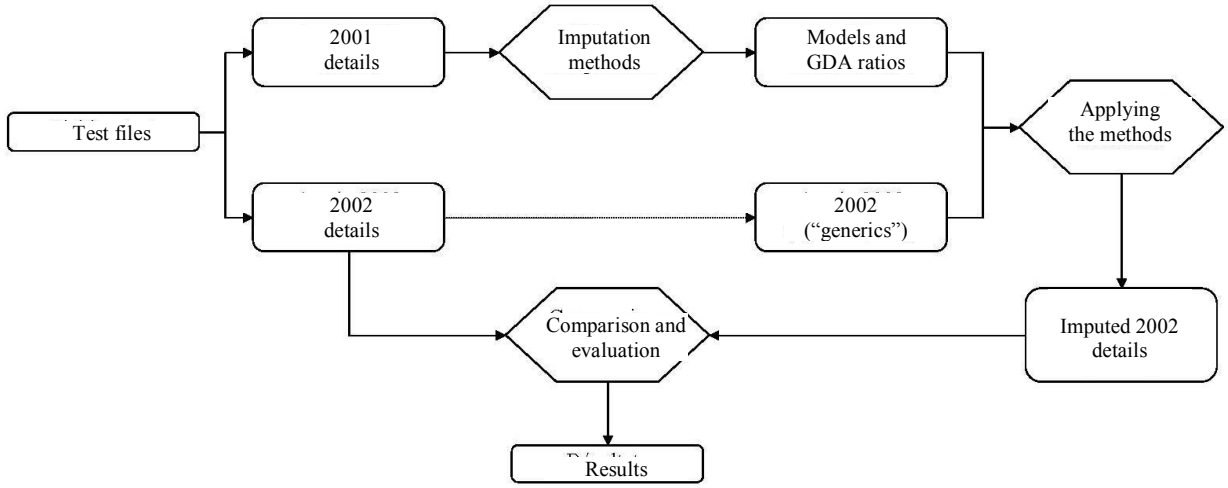
A number of measures can be used at the block level:

The sum of the squares of the errors: $SSE = \sum_i (t_i - \hat{t}_i)^2$, where t_i and $\hat{t}_i$ denote the actual total reported for detail i and the estimated total respectively.

The sum of the squares of the errors in the distribution percentages: $SSEP = \sum_i (\frac{t_i}{\hat{t}_i} - 1)^2$

The "Macro_pseudo_CV": $Macro\_pseudo\_CV = \sqrt{\sum_i (t_i - \hat{t}_i)^2} \Big/ \sum_i t_i$

Figure 2: Diagram of the process for evaluating imputation methods



3.4 Applying the method to block 9760

For 2001 and 2002, details only were reported by 21,782 and 22,041 businesses respectively. This block corresponds to five COA details.
The 21,782 businesses were allocated by automatic classification to two (Cluster2), five (Cluster5) and 15 (Cluster15) classes, and various "attractors": Attractor60, Attractor70 and Attractor90. A final ad hoc classification (Clus_User) was defined using important details.

Homogeneity
As indicated by the Pearson coefficients in Table 2, the groups produced by the various classifications are more homogeneous than the groups defined by the current imputation method.

Table 2: Homogeneity of the groups produced by the various classifications (block 9760, Pearson coefficient)

Cluster	Attractor60
1	0.602
2	0.601
3	0.492
4	0.630
5	0.588
6	0.682

Clus_User	
Maximum	0.711
Q3	0.404
Median	0.381
Q1	0.353
Minimum	0.320

GDA subgroups	
Maximum	0.817
Q3	0.811
Median	0.739
Q1	0.707
Minimum	0.583

Precision
For this block, we consider the 10 best models for the Cluster2, Cluster5, Clus_User and Attractor80 variables, and the 20 best models for the other classification variables (Attractor60, Attractor70 and Cluster15). These models were selected on the basis of their classification error rates.

Figure 3 shows the best models of the eight types of class variables based on previously defined precision criteria: SSE and Pearson coefficients. The GDA procedure is not as good as the models based on the Attractor60 variable for the macro-level criteria and the models based on the Clus_User variable for the micro-level criteria. Its performance is comparable to that of models based on the Cluster2 variable.

What will make the difference is primarily the models' performance in estimating the percentages of the various details. Figure 4 shows the performances of the best models (based on the SSE) of each class variable.

Figure 3: Graph of the best model for each class variable based on the SSE and the median of the Pearson coefficient (block 9760)

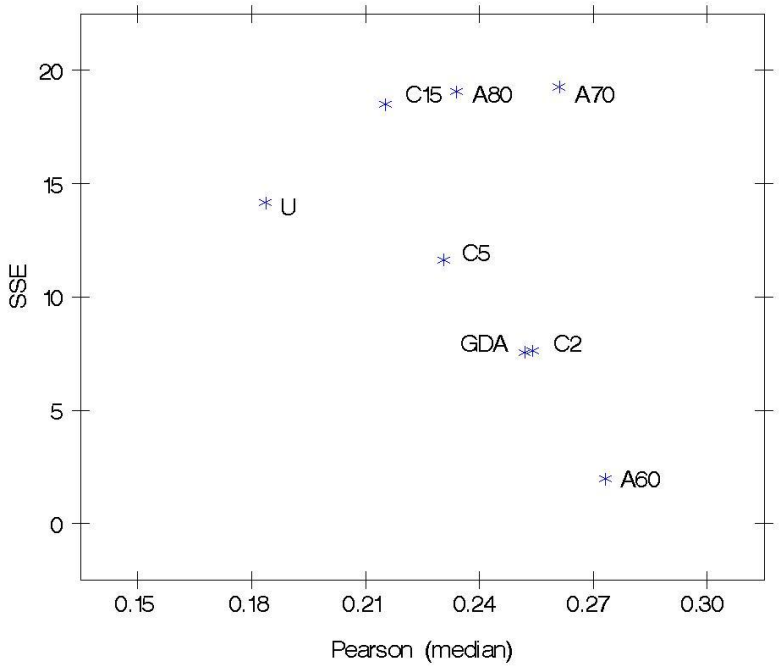
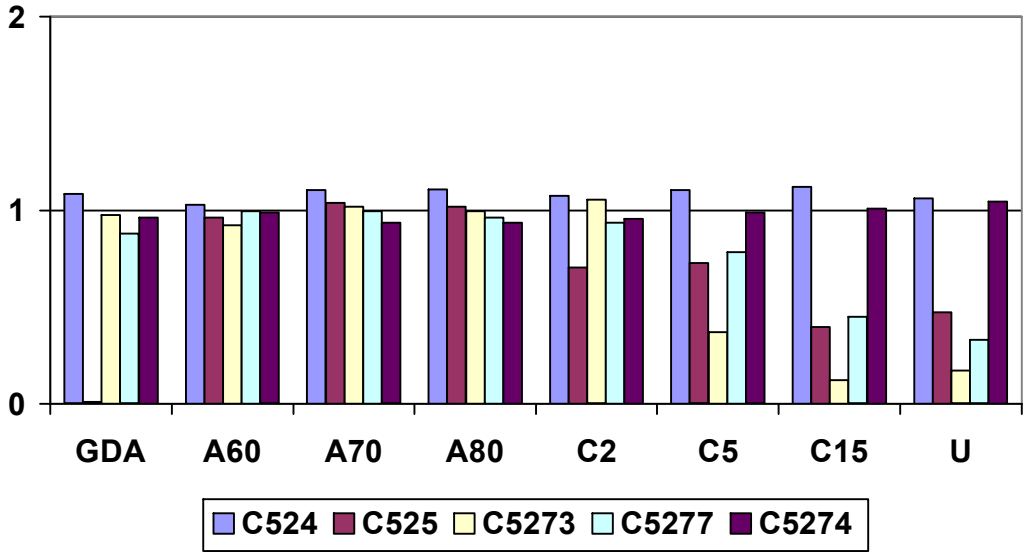


Figure 4: Sum of the squares of the distribution percentages (SSEP) for each detail for the best models of each class variable



The current GDA procedure seriously underestimates detail C525 and overestimates detail C524. The models constructed with the Cluster2, Cluster5, Cluster15 and User_Clus class variables produce very unbalanced estimates of the percentages for each detail.
The best model appears to be the one based on the Attractor60 variable. That is the model we have selected, even though it is not very parsimonious, using 16 explanatory variables.

In summary, for block 9760, the best strategy for imputing the details is based on an *a priori* Attractor60 classification, with the class being predicted by a discriminant analysis using a model with 15 explanatory variables (L9370 L9470 L9541 L9662 L9663 L9760 L9799 L9802 L9804 L9811 L9818 L9819 L9820 L9835 and ratio) discretized into 30 groups, and using a non-parametric estimation method.

4. CONCLUSION

The current procedure for imputing the detail distributions, which is based on the use of marginal distributions of subgroups defined in advance by industry code and business size, can be improved by creating homogeneous distribution classes. This can be accomplished with an automatic classification (AHC) or ad hoc procedures of the "Attractor" type. Non-parametric discriminant analysis can then be used to assign businesses with sufficient precision to classes based on their characteristics. Imputation of the detail distribution through the observed marginal distribution for the class yields better results.
However, the procedure has to be adjusted for each block since the classification method and the optimum model can vary from block to block.

REFERENCES

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Goodman, L. and Kruskal, W. (1979), *Measures of Association for Cross-classifications*, Springer-Verlag, New-York.

Helmer, P., Lafontaine-Sorgo, D., Lalande, D. and Thibault, D. (2001), Generic to Detail Allocation: GIFI 1999 Documentation, Statistics Canada.

Rondeau, C. (2003), Specifications of the GDA ratios calculation, Tax data Division, Statistics Canada.

Rossiter, P. (1998), Modelling Detailed Business Operating Expenses From ABS Economic Collections, Methodology Advisory Committee Paper, Australian Bureau of Statistics.

Saporta, G. (1990), *Probabilités, Analyse des données et Statistique*, Technip.

SAS Institute (2001), *SAS/STAT User's Guide: PROC FREQ*, North Carolina.

Makridakis, S. and Hibon, M. (2000), "The M3-Competition: results, conclusions and implications", *International Journal of Forecasting*, 16, 451-476.

NIST (1998), "StRD: Statistical Reference Datasets for Assessing the Numerical Accuracy of Statistical Softwares", http://www.nist.gov/itl/div898/strd.