Statistica Neerlandica (1998) Vol. 52, nr. 3, pp. 255–257

Large data sets in finance and marketing: introduction by the special issue editor

Philip Hans Franses*

Econometric Institute, Erasmus University Rotterdam, P. O. Box 1738, 3000 DR Rotterdam, The Netherlands

On December 18 and 19 of 1997, a small conference on the ''Statistical Analysis of Large Data Sets in Business Economics'' was organized by the Rotterdam Institute for Business Economic Studies. Eleven presentations were delivered in plenary sessions, which were attended by about 90 participants. The current issue of *Statistica Neerlandica* contains five papers originating from several of these presentations. All papers have been refereed, and revised according to the suggestions made by the reviewers. It is a great pleasure for me to introduce this issue with remarks on the specific field of interest, a brief discussion of the contents of this issue and acknowledgements. I hope the reader finds the issue useful and stimulating for further research.

1 Some introductory remarks

With the advent of improved data collection and storage techniques, one can nowadays get access to data at the transaction level in marketing and finance. In marketing, scanner methods in supermarkets and customer loyalty cards at the household level, for example, can reveal information on any transaction: time, location of the store, and price of the product; but also, age, income and past behavior of the customer. In finance, tick-by-tick trade data have become available for many important stocks, exchange rates and interest rates. Together, this results in the availability of large amounts of data points. Often, statisticians or econometricians are asked to somehow summarize these data, with the intention of deriving forecasts or policy measures.

A first thought is that statisticians should of course be happy with such large data sets. In fact, the number of observations approaches infinity, and hence rather precise statements on the data can be made. However, with this increase of data points, there are also many more statements to be made than one is used to making. Also, it is unlikely that simple linear models with only a few parameters can adequately describe the data. Hence, many observations allow for more detailed analysis using more complicated and advanced models. Some of these models already exist, others need to be developed.

* franses@few.eur.nl

The natural question is whether we really need all these data. Perhaps scanner information at the store level (aggregating over customers) is already sufficient for many questions. And, to trace common patterns across financial markets, one may need weekly returns only. For other questions less aggregated observations are needed. For example, to study the impact of news flashes on stock market volatility, returns measured per second are required. In summary when data are plenty, aggregation and sampling issues become relevant.

2 Contents of this issue

This special issue of *Statistica Neerlandica* brings together five papers, all of which address topics in the statistical or econometric analysis of large data sets in either finance or marketing. The first paper by Clive Granger provides a general view on the issue of extracting information from what he calls mega-panels (in marketing) and the high-frequency data (in finance). Granger discusses which current statistical techniques may become less relevant for such analysis, and he states that new techniques have to be developed. For example, for time series data it may become possible to estimate the conditional distribution (given past observations) instead of the usual conditional mean or variance.

The second paper by Torben Andersen and Tim Bollerslev reviews recent developments in empirical finance concerning the analysis of high-frequency data, and their usefulness for understanding the properties of data collected at lower frequencies. They address modeling intraday returns and volatility, long memory properties, and forecasting volatility (all for univariate time series). It is argued that high-frequency intradaily returns are very useful for many purposes.

The third paper by Sridhar Balasubramanian, Sunil Gupta, Wagner Kamakura, and Michel Wedel is also a survey paper, though now concerning developments in empirical marketing research. These authors review the statistical models used in that area, and how these result in an increased understanding of customer behavior. Also, the models can yield profit functions at the customer level, which can be very useful for the everyday practice of direct marketing.

The last two papers of this issue contain more detailed analyses of some specific issues in empirical finance. The papers have in common the fact that many data are required to perform the statistical analysis. Silvia Caserta, Jon Danielsson, and Casper de Vries argue that large high-frequency data sets are useful to examine extreme events that imply extreme risk. These authors show how such an investigation can be used to evaluate the returns on exotic options. Finally, Christian Hafner and Helmut Herwartz address how multivariate models for sets of excess returns can be interpreted in terms of the effects of positive or negative news.

The general conclusion from all papers in this issue is that large data sets in finance and marketing provide an opportunity to do superior analysis (to quote Clive Granger). Assuming knowledge of how to do this analysis, large data sets can be informative for economic behavior (since transaction level data are available) and for

ready-to-use guidelines on practical problems. It is expected, however, that much further theoretical and empirical research is needed to obtain substantial knowledge of the most appropriate statistical tools for approaching these large data sets.

Acknowledgements

Even a small conference cannot be organized without generous sponsors. I am very grateful to the following institutions:

- Rotterdam Institute for Business Economic Studies,
- Tinbergen Institute (the Rotterdam branch),
- Erasmus Center for Financial Research,
- Erasmus University Trustfund, and the
- Economics Section of The Netherlands Society for Statistics and Operations Research,

while stressing that this list does not correspond with the degree of their generosity.

Thanks are also due to all speakers at the conference, and to the participants who sometimes generated quite some discussion.

The authors of the five papers in this issue are all (no exceptions!) thanked for meeting the (sometimes perhaps too rigid) deadlines. Due to their efforts this issue appears in print less than one year after the conference was being held. It is also worthwhile to memorize that all five papers were specifically prepared for the RIBES conference and for this issue.

All papers were reviewed by referees. I would like to thank (in alphabetical order):

- Ronald Huisman (University of Maastricht)
- Frank de Jong (University of Tilburg)
- Jedid-Jah Jonker (EUR)
- Franc Klaassen (University of Tilburg)
- Frank Kleibergen (EUR)
- Teun Kloek (EUR)
- Ronald Mahieu (EUR)
- Bert Menkveld (EUR)
- Richard Paap (EUR)
- Linda Teunter (EUR)

for their kind help.

Last but not least, I should thank three more individuals. First, Thea de Hoog (secretary of RIBES) provided very accurate, secretarial assistance. Second, Paul de Boer (managing editor of Statistica Neerlandica) was kind enough to ensure a speedy production process. Finally, I am much indebted to Hans van Houwelingen (editor-in-chief of Statistica Neerlandica) for allowing me to put together this special issue.