From Data Analysis to Data Science

Noboru Ohsumi

The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku Tokyo 106-8569, Japan
(e-mail: ohsumi@ism.ac.jp)

Abstract. This paper discusses the significance of the term "data science" to the Japanese Classification Society (JCS) and the international relevance of JCS's research. In 1992, the author argued the urgency of the need to grasp the concept "data science". Despite the emergence of concepts such as data mining, this issue has not been addressed. Discussion will emphasize the history of methods of data analysis proposed by J. Tukey. The interaction between Japan and, particularly, France in the development of data analysis will be emphasized.

1 The research interchange between Japan and France

Because of differences in cultures and researchers' approaches, globalization of the field of statistical science and data analysis remains a future prospect. At the risk of being accused of making an arbitrary interpretation, the author asserts that Japanese researchers looked to France and Germany in the field of mathematics, and toward the UK and the USA in the field of statistical science. The field of data analysis was a rare exception, where Japanese and French researchers collaborated. To our regret, the history of these interchanges is not widely known among statistical science researchers.

One such exchange was between Professor Matusita of the Institute of Statistical Mathematics (ISM) and the late Professor Dugué of the Institute of Statistics at the University of Paris VI. Through their shared interests in traditional mathematical statistics, especially multivariate analysis, they organized the Japanese–French Scientific Seminar on "Data Analytic Methods for Analysing Measurement Datasets". This "bridging seminar" marked the beginning of research exchanges between the researchers in both countries and the subsequent development of data analysis in both countries. It also marked the beginning of an enduring collaboration of Japanese researchers with French researcher Professor J.-P. Benzécri and promising young data analysts of Benzécri's school, including Lebart, Roux, and Jambu.

A group of Japanese researchers led by C. Hayashi was at the hub of this field in Japan at the time, and had achieved considerable advances in data analysis research. Researchers at the ISM knew only partially of concurrent developments in France. It was known, for example, that development of a method similar to Hayashi's Quantification Methods, Type III, had stimulated progress in data analysis in France. However, nothing was known of the work of the "phantom researcher" Professor J.-P. Benzécri.

Such contact included seminars and special lectures at the ISM, the JCS, and other places in Japan to introduce the French philosophies to Japanese researchers. The "*analyse des données*" introduced by Roux was astonishingly new and stimulating to Japanese researchers, as were correspondence analysis (CA) and automatic classification. Roux made an immense contribution to clarifying the similarity of the mathematics between CA and the Type III Quantification Method. Many notable achievements, such as Hayashi's quantification methods and Akaike's information criterion, developed in quick succession. The ISM played an essential part in offering opportunities to put these new theories into practice. It was particularly noteworthy that much of the research based on social surveys in Japan, such as the Survey of Japanese National Character, was undertaken using the Type III Quantification Method.

After Roux visited Japan, Professor L. Lebart, an authority in data analysis and social survey research, was invited to participate in a project that involved Japanese and French researchers from the Japan Society for the Promotion of Science (JSPS), the Centre National de la Recherche Scientifique (CNRS), and the ISM conducting a survey of international attitudes to the "Japanese and French national characters".

2 Later international research interchanges

By 1983, the SFC had started in France, and the Classification Society of North America (CSNA), the Gesellschaft für Klassifikation in Germany, and the British Classification Society (BCS) had been organized. In the same year, Hayashi, together with some researchers who were international members of the CSNA and BCS, founded the Japanese Classification Society. Membership in the International Federation of Classification Societies (IFCS) was gained and, through the great effort of H. Bock and others, the JCS was in the fortunate position of being able to host the Fifth IFCS-96 Conference.

Japanese researchers also promoted international interchange through large conferences such as the meetings of the International Statistical Institute (ISI) and International Biometric Society (IBS) held in Tokyo in the 1980s, which attracted such researchers as Y. Escoufier, J.-P. Nakache, Bouroche, J. Gower, A. Rizzi, and N. Lauro.

The period between 1979 and 1985 marked an important period of close research exchanges between Japan and many European countries.

3 "Analyse des Données" and "Deta Kaiseki"

It is most important to discuss the similarities and differences in the approaches to data analysis between Japan and France with regard to the Quantification Method, especially CA. It is important to emphasize that we agree on the need to develop, through practice, research on the theory

and application of data analysis into a new "data science". Hayashi's Quantification Methods comprise several methods, from Type I to Type VI. In particular, Type III coincides with CA. Hayashi proposed this method in 1952. Underpinning Hayashi's methods was the concept of scaling methods, by which the other methods were unified and discussed. Benzécri's CA (AFC: Analyse Factorielle des Correspondances) appeared about 1962 (Benzécri, 1982). How well it was accepted and what applications were developed from it goes without saying. Benzécri and his school developed elaborate and varied theories of CA and related methodologies. Moreover, considerable research was conducted on automatic classification by many researchers, including Diday, Jambu, Lerman, and Roux. To our regret, however, the "barrier of language" prevented Japanese researchers from gaining true recognition for their achievements. The jargon used in research on the "*analyse des données*" made things even more difficult. Although there has been some improvement, we are still in much the same situation.

In Japan, the term "deta kaiseki" (data analysis) was often misunderstood. The Japanese language used in these papers prevented these achievements from becoming known to international researchers. However, because of publications in Japanese by Lebart and Ohsumi (1994), Japanese researchers are now able to obtain more results of research in France and in other countries. Differences in language, thought and culture make most Japanese researchers more interested in research in English-speaking countries, which presents a great problem for us to solve in the present time. Books in English by Greenacre (1984) and Jambu (1983) are read by many Japanese researchers and students. Those who are interested in *analyse des données* are increasing in number.

Two important results should be remembered in the history of research interchange. In the past, Japanese–French Scientific Seminars were arranged. The first meeting was held at the ISM in Tokyo in 1987 and attracted 180 researchers—an unexpectedly large number. The second meeting was in Montpellier University II in France in 1992. Fewer researchers participated, but the outcome of this meeting was significant: the term "data science" appeared for the first time, and was subsequently used in the preface of a conference publication ("Data Science and Its Applications—La Science des Données et ses Applications": Escoufier et al., 1995).

Researchers in Japan do not all share the same understanding of the concept "data science". The Japan Statistical Society held special sessions on data science at its annual meetings in 1996 and 1997, and drew much interest. However, in the opinion of most researchers, they did not go beyond the general framework of statistical modelling or traditional statistical analysis. One organizer was heard to criticize Japanese researchers for using other researchers' data without paying any attention to the most important problem of data acquisition. What, then, is our "data science"?

What I mean by "data science" includes the most essential studies and concepts on *how to gather data*, including *how to design experiments in data gathering*, and *how to analyse the collected data*. These are the fundamental ways to obtain meaningful findings from many events. How data are gathered is the key to defining the relevant information and making it easy to understand and analyse. In my opinion, this viewpoint on the meaning of data science is fundamentally different from data mining (DM) and knowledge discovery (KD). These concepts are not of practical use because they neglect the problems of "data acquisition" and its practice.

4 Relationship to IFCS: Changing from a linear to a spatial perspective

Japan's foreign relations in the field of data science developed from initial research exchanges with France. The relation was at first a linear one, but more extensive relations followed through foundation of the IFCS, as did exchanges with many other countries. The IFCS was founded in 1983 to federate the classification societies from many countries. The First IFCS International Conference, held at Aachen in Germany in 1987 (organized by Professor Bock), deserves special mention for being the first meeting held by the federation of BCS, CSNA, GfKl, JCS, SFC, and SIS. Japan hosted the Fifth IFCS-96 Conference; this was the culmination of 20 years of international research exchange. In this context, the association between Japan and France may have undergone a marked change from a linear to a spatial relationship.

5 Toward Data Science: as prospects in Data Analysis

The Japanese song "A canary that has forgotten singing" describes the current trend in the field of the data analysis. It appears researchers are seeking mathematical methodologies without considering "what data analysis is" and "what the data acquisition should be". Were we not seeking for a different world of statistical science and data analysis?

Owing to qualitative and quantitative changes in data, it is, indeed, becoming increasingly difficult to grasp all aspects of a dataset in explaining various phenomena. Therefore, new techniques, such as DM, KD, complexity, and neural networks, are being proposed. However, the potential of these methods to solve any of these problems is questionable.

We now have to deal with not only extremely complicated analyses but also greatly altered data. Their characteristics could be categorized as follows:

1. A dataset collected with a definite purpose of explaining phenomena on the basis of statistically appropriate design of experiments or sampling procedures; for example, social survey data including opinion surveys or

attitude surveys. The data acquisition process is transparent, traceable, or reproducible.
2. Laboratorial measurement data gathered with measuring tools or devices. The data include various kinds of measurement units, such as environmental indexes and health indexes, as well as datasets acquired through actual measurements.
3. Data that accumulate gradually in the database by an information processing technique. The purpose or intention of the accumulation cannot be clearly demonstrated. These data include POS data, banking and credit data, and basic financial and personnel databases of corporations.
4. A new qualitative kind of dataset and its database. In particular, textual or non-numerical data extracted from open-ended responses or free format answers and collected systematically. For example, textual data gathered almost automatically through Internet researches, telephony-marketing researches, and call centres for customers.
5. An aggregated dataset generating spontaneously and accumulating automatically in the electronic data collection environments, and its database or data warehouse. A mass of data, the importance of which is not readily known, but which is managed by the high-tech database with a view to extracting some meaning in the future.

When it comes to analysing these datasets, people discuss DM and related techniques. However, the important questions to answer are: what dataset is necessary to explicate a certain phenomenon, why is it necessary, how to design its acquisition, and how difficult the whole process is. This is more important than the dataset itself. Books on DM do contain terms such as "data preparation", "getting the data", "sampling procedures", and "data auditing", but there is an assumption that the dataset is given and the procedure may start with analysis. Fiddling with a dataset once it is collected is merely a self-contented play of data handling. As noted, there are many possible ways to acquire a dataset. Taking this into consideration, one should ask what data analysis should be. To come to the point: I have discussed the paradigm through which we should discuss the concept of data science. Unfortunately, although it is such a basic and fundamental concept, I doubt whether data analysts have been well aware of its importance.

A decline in statistical science was brought to our attention long ago. Nevertheless, no marked improvement has been made. No university in Japan has a department of statistics. In the field of statistical science and data science, we have only one specialized research institute, ISM. Our only statistical science course in graduate school is also at the ISM. Recently in Japan, there has been a great deal of discussion over the guidelines for improving scientific research. In the fields of computational science and informatics, many have thought it necessary to examine how to advance research, and many research projects from other countries are being introduced for the purpose of comparison or benchmarking. Models drawn from large-scale national research

centres in European countries, such as INRIA, the organization of CNRS, the Max Planck Gesellschaft, and MPI-Institut für Informatik have drawn considerable interest. In the field of data analysis, a large-scale National Institute of Informatics was partially commenced in 1999, and is planned as part of a structural reorganization program.

At present, however, there is no clear direction for change; we must determine that direction, each re-examining and revitalizing our separate attitudes. For that purpose, we might have to seek collaboration with other fields, or even consider the possibility of re-organization and integration. We might have to abandon such terms as statistical science or data analysis or similar concepts, and choose, for example, "data science" as a new paradigm. We do believe that such a concept can help to guide and foster a fruitful and expanding relationship among many countries in the future. We very much hope this new age of "data science" will come to fruition, and that what we have achieved in the history of "data analysis" will be of enduring benefit to the coming science and to future research.

Acknowledgements

I wish to thank all staff and members of the Organization Committee of the IFCS-2000 Conference for giving me the opportunity to present this report, and to the researchers from each country belonging to the IFCS, for making great efforts toward the development of data science.

References

DIDAY, E., LEBART, L., PAGES, J.-P. and TOMASSONE, R. (Eds.) (1979): *Data Analysis and Informatics.* North-Holland, Amsterdam.

DIDAY, E., JAMBU, M., LEBART, L., PAGES, J.-P. and TOMASSONE, R. (Eds.) (1983): *Data Analysis and Informatics III.* North-Holland, Amsterdam.

DIDAY, E. and others. (Eds.) (1986): *Data Analysis and Informatics IV.* North-Holland, Amsterdam.

ESCOUFIER, Y., HAYASHI, C., FFICHET, B., DIDAY, E., LEBART, L., OHSUMI, N. and BABA, Y. (Eds.) (1995): *Data Science and its Applications* (La science des données et ses applications). Academic Press, Tokyo.

GREENACRE, M. J. (1984): *Theory and Applications of Correspondence Analysis.* Academic Press, Boston.

HAYASHI, C., DIDAY, E., JAMBU, M. and OHSUMI, N. (Eds.) (1988): *Recent Developments in Clustering and Data Analysis* (Developpements recents en classification automatique et analyse des données). Academic Press, Boston.

HAYASHI, C., OHSUMI, N., BOCK, H.-H. and others (Eds.) (1997): *Data Science, Classification and Related Methods.* Springer-Verlag, Tokyo.

JAMBU, M. and LEBEAUX, M.-O. (1983): *Cluster Analysis and Data analysis.* North-Holland, Amsterdam.

OHSUMI, N., LEBART, L. and others (1994): *Multivariate Descriptive Statistical Analysis* (in Japanese). JUSE Press, Ltd., Tokyo.