From Data Analysis to Data Science

An Overview and Future Prospects: the Research Interchange in Data Analysis Between Japan and France

Noboru Ohsumi

The Institute of Statistical Mathematics
4-6-7 Minami-Azabu
Minato-ku, Tokyo 106
Japan

Summary: The objective of this report is to present a brief overview of the research interchange in the field of Data Analysis between Japan and France, from the point of view. This paper will not attempt to focus on any specific theories or methods; rather, the main purpose is to introduce you to the history of the interaction between France and Japan from my perspective over quite a long period. I hope my remarks will serve as a suitable celebratory address for this meeting of the 20th Anniversary of Société Francophone de Classification (SFC). The methods of "data analysis" advocated and developed by J.Tukey, and known as Exploratory Data Analysis (EDA), have been widely accepted and used. As an inevitable result, many statisticians and data analysts have focused on research trends in the United States or Britain. However, it must be remembered that in both Japan and France, further fruitful fields of data analysis have been cultivated, separately from the development of EDA. It has been nearly 20 years since the bonds of relationship were formed between Japan and France. We cannot help being surprised at the fact that ideas which turned out to be very similar to each other were born in such distant and different cultures almost at the same time. In both countries, theories were generated which were more powerful than Tukey's methods, and they were also developed earlier. It is a pleasure to have this opportunity to acquaint you with the past, present, and future of this important and productive research exchange in data analysis between the two cultures. It would be expected that this report will encourage all researchers to take a further step forward, and will help to make our relationship still closer.

1. The Beginning of the Research Interchange between Japan and France

The research interchange among the data analysts and statisticians in Japan and France has a longer history than you might imagine. However, this history is not widely known among the researchers here in France. In the fields of mathematics, probability theory, and mathematical statistics, researchers had long been promoting active and close research interchange. However, to our regret, these interactions had involved no cross-connection between those sciences and the field of data analysis. Under such circumstances, Professor Matusita of the Institute of Statistical Mathematics and the late Professor Dugué of the Institute of Statistics at the University of Paris VI came into contact with each other.

Their relationship, which had been formed through their research on traditional mathematical statistics, especially on multivariate analysis, led to a small meeting, the Japanese-French Scientific Seminar on "data analytic methods for analyzing the measurement datasets." It was a notable landmark to start our research exchange, and moreover, it was a memorable event in the development of data analysis in both countries.

The first seminar was arranged by Matusita, and was held in Paris in 1978 with the support of the Japan Society for the Promotion of Science (JSPS). Though it was a small meeting with only ten or so participants, it had major significance, especially for Japanese researchers; we had a chance to meet Professor J.-P. Benzécri, who at that time had been dubbed a "phantom researcher" by Japanese researchers because of our very limited knowledge of him. We also had a chance to meet Professors Lebart, Roux, and Jambu, who were then promising young data analysts of Benzécri's school, and with whom we have been keeping close contact ever since.

In those days, there was a group of researchers in Japan who had achieved considerable advances in data analysis research by developing theories and methods of their own and putting them into practice. The group was led by C. Hayashi, a charismatic researcher, and the founder of the quantification methods. Hayashi then held the post of General Director of the Institute of Statistical Mathematics and was at the hub of this field of research in Japan.

The concurrent developments in France we knew only partially. We had heard that a method similar to one of Hayashi's Quantification Methods, Type III, originally called the pattern classification method, had been developed by a researcher and that the method had stimulated progress in data analysis in France. However, at that time there was no such convenient information device as the Internet available, so we were never informed of the work of the "phantom researcher," or that of Nicolas Bourbaki in mathematics.

Meanwhile, as just mentioned, we had a chance to meet Professor Benzécri at the 1978 seminar, and since then Japanese and French researchers have been in a close and lasting relationship with each other. This was the beginning of the history of the mutual development of the new data analysis in France and Japan.

2. The Dawn of Data Analysis

Thus, a bridge was built between Japan and France. We were at the starting point. Hayashi and some other researchers, including myself, planned to invite Benzécri to Japan. Unfortunately, as it turned out, Benzécri could not come, but he introduced Professor M. Roux to Japan on his behalf. Then, one of his evangelists enrolled at University of Paris VI, as JSPS's invited researcher. I myself, as his coordinator for several months, organized some seminars and special lectures at the Institute of Statistical Mathematics (ISM), the Japanese Classification Society (JCS), and other places in Japan to introduce the French philosophies to Japanese researchers.

The "*analyse des données*" (data analysis) introduced by Roux was astonishingly new and stimulating. We were very interested in the ideas of correspondence analysis, automatic classification, and so on. Roux made an immense contribution by helping us examine how similar the mathematics are between correspondence analysis and the Type-III Quantification Method.

In those days, at the Institute of Statistical Mathematics, we used to have heated discussions about what statistical analysis was practical, what statistical science was useful, and above all, what data analysis was. As Benzécri clearly pointed out (1982), a step forward was needed in the field of data analysis research, and this situation was the same in Japan. Many notable achievements, such as Hayashi's quantification methods and Akaike's information criterion (AIC), had come out one after another. The Institute played an essential part in offering opportunities to put these new theories into practice. It was particularly symbolic that much of the research based on surveys in Japan, such as the Survey of Japanese National Character, was done using the Type-III Quantification Method.

After Roux visited Japan, our plan was realized and we invited Professor L. Lebart, an authority in data analysis and social survey research, with the support of the JSPS, the Centre National de la Recherche Scientifique (CNRS), and the Institute of Statistical Mathematics. Lebart and Hayashi had been keeping in close contact with each other. Then, the opportunity was taken to survey international attitudes to the "Japanese and French national characters," both in Japan and France. Lebart visited Japan several times after that. We think highly of him and his colleagues for their contribution to our survey-based research in Japan.

3. The Foundation of the Japanese Classification Society and Its Relationship to the SFC

French researchers made another great contribution to Japan. They worked hard together with

us toward the foundation of the Japanese Classification Society (JCS) in 1983. By that time, the SFC had already started in France, and the Classification Society of North America (CSNA), Gesellschaft für Klassifikation (GfKl), and British Classification Society (BCS) had been organized in their respective countries. They had continued their activities, holding meetings and conferences of their own, or joint meetings, in the case of CSNA and BCS. Meanwhile, Professor M. Jambu, who was a leading figure in the SFC, contacted us and called on us. We learned a lot from him about the activities of the classification societies in the USA and Europe. Hayashi, some other researchers, and myself, as overseas members of CSNA and BCS, had obtained additional information, and decided to found our own society immediately, by way of a small meeting. Later, in 1983, when we gained membership in the International Federation of Classification Societies (IFCS), we established the modus operandi of the society. The JCS is a small society, with a membership of about 200, but it is highly regarded in Japan as the society representing Japan in the international federation. We later took part in the Fifth IFCS International Conference, which came about through the great effort of H. Bock and others.

4. Fruitful Research Interchange Through Later Meetings

I will mention the Japanese-French Scientific Seminar later. There were other large international conferences held in Japan which presented opportunities to meet many French researchers. For example, at the meetings of the International Statistical Institute (ISI) and International Biometric Society (IBS) held in Tokyo in the 1980s, we had a chance to meet researchers Y. Escoufier, Nakache, Bouroche, and Gower from England, and Rizzi and Lauro from Italy. Thus, international interchange was greatly promoted.

The Japanese Data Analysts' group appeared in a large international conference in France for the first time when a meeting on "Data Analysis and Informatics," organized by E. Diday and some other researchers, was held. At the second meeting in Versailles in 1979, C. Hayashi made a speech as an invited speaker, and some other Japanese researchers, including me, read their papers. Other Japanese researchers — Iwatsubo, Yanai, Ohsumi, Takakura, and Sugiyama — followed them at the third and fourth meetings, held in 1983 and 1985, respectively. In particular, the second meeting in 1979, above all, must be remembered by Japanese researchers as extremely important. It was a starting point for close research exchanges between Japan and many European countries, such as Britain, Italy, Germany, Spain, Switzerland, and so on. Acquaintance with Italian researchers, such as Rizzi and Lauro, also evolved into a productive relationship.

5. "Analyse des Données" and "Deta Kaiseki"

There are similarities and differences in the approaches to data analysis between Japan and France. 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."

On the other hand, researchers have different ideas and adopt different approaches due to their cultural backgrounds. So, to make matters simple, it may be apposite to make a comparison between the Type-III Quantification Method and Correspondence Analysis as typical examples.

Hayashi's Quantification Methods comprise several methods, from Type I to Type VI. In particular, Type III, originating from the pattern classification method, coincides with Correspondence Analysis. Hayashi proposed this method in 1952, 10 years earlier than Benzécri described his method. Hayashi's interests covered the whole range of analysis of qualitative data. Underpinning his methods was the idea of scaling methods. Under this scaling method, the other methods were unified and discussed.

We understand that Benzécri's Correspondence Analysis (Analyse Factorielle des Correspondances) appeared about 1962 (Benzécri, 1982). How well it was accepted and what applications were developed goes without saying. Benzécri and his school, the late Professor

B. Escofier, P. Cazes, Lebart, J.-P. Pages, and other researchers, succeeded in developing elaborate and varied theories of Correspondence Analysis and the related methodologies, for example, multiple correspondence analysis. Moreover, there was a lot of research on automatic classification by Diday, Jambu, Lerman, Roux, and many other researchers, and we saw in their achievements trends emerging which were quite different from those in the United States and Britain. To our regret, however, the "barrier of language" intervened, and Japanese researchers were unable to gain true recognition for their achievements. In addition, the "dialect" or "jargon" used in research on the "*analyse des données*" made things still worse. Though there has been some improvement, we are still in much the same situation now.

In Japan, because of various dialects and the uniqueness of Hayashi's theory, there was a misunderstanding that "deta kaiseki" is very difficult. The Japanese language used in these papers prevented these achievements from becoming known to overseas researchers. However, with publications by Lebart and Ohsumi in Japanese (1994), Japanese researchers are now able to get much more information about the research results in France and in other countries. Differences in language and thought make most Japanese researchers more interested in research in English-speaking countries, which presents a great problem for us to solve. 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. We do hope researchers in both countries will overcome the barriers, and develop a more active research interchange.

Now, on behalf of many Japanese researchers, we would like to give special thanks to our French colleagues Bouroche, Caussinus, Diday, Durand, Escoufier, the late Escofier, Fichet, Holmes, Jambu, Lebart, Lerman, the late Megreditchian, Morineau, Nakache, Schektman, Van Cutsem, and many other French researchers. We value their efforts and understanding highly.

6. The Expansion of Our Research Interchange Through the Japanese-French Scientific Seminar

Two important results that should be remembered in the history of research interchange between Japan and France must be emphasized. In the past, Japanese-French Scientific Seminars were arranged. The first meeting was held at the Institute of Statistical Mathematics (ISM) in Tokyo in 1987. Over 180 researchers took part in the seminar, a number much larger than we had expected. The second meeting was in Montpellier University II in France in 1992.

The first meeting was organized by Jambu, Hayashi, and Ohsumi. Fortunately, we were able to obtain support through the Japanese-French Research Interchange Fund from JSPS and CNRS. The meeting was held under the auspices of those organizations and ISM. The results can be seen in the book titled "Recent Developments in Clustering and Data Analysis" (Hayashi et al, 1988).

The second meeting was organized by two members of the Escoufier group, Hayashi and Ohsumi. Not so many researchers participated, but its outcome was very significant: the term "Data Science" appeared in this meeting for the first time. This was a landmark in the history of data analysis studies. After the scientific seminar, Escoufier, Hayashi, and Ohsumi were engaged as editors of the proceedings, and while writing the preface of the book we used the term "Data Science." I remember that was when we made some arrangements at ICOT-4 held at Morocco in 1994. The book was published under the title of "Data Science and Its Applications —La Science des Données et ses Applications" in 1995, three years after the seminar was held (Escoufier et al., 1995).

In Japan, more and more researchers have come to use this term. However, not all researchers share the same concept of what Data Science is. The Japan Statistical Society planned and held special sessions on Data Science at its annual meetings of 1996 and 1997, and drew much interest: but, in the opinion of most researchers, they didn't go beyond the general framework of statistical modeling or traditional statistical analysis.

What we 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 so on. It is the way that data are gathered that is the key to defining the relevant information and making it easy to understand and analyze. This viewpoint on the meaning of Data Science is fundamentally different.

7. Relationship to IFCS — Changing from a Linear to a Spatial Perspective

Japan's foreign relations in this field began with the research interchange between Japan and France. The relation was at first a linear one, so to speak, but more extensive relations developed afterwards, one of the results of which was the foundation of the IFCS, and another was the actual exchanges with many other countries.

Thanks to the efforts of all concerned, the IFCS was founded in 1983 to federate the classification societies from many countries. H. Bock's devotion to the First IFCS International Conference held at Aachen in Germany in 1987 deserves special mention. It was the first meeting held by the federation of BCS, CSNA, GfKl, JCS, SFC, and SIS. After that, IFCS meetings were held in Virginia, Edinburgh, Paris, and Kobe.

So far, twelve societies have joined the federation. It may be very hard to organize it well, and the number of federated societies is small, but its aim is positive and constructive. Japan was in the fortunate position of being able to host the Fifth IFCS-96 Conference in March 1996, after the fourth in France in 1993. This is certainly one of the greatest results of the twenty years' research interchange between Japan and France. Thus, the association between both countries may have undergone a marked change from a linear to the spatial relationship.

8. Toward Data Science — Prospects in Data Analysis

Generally speaking, it has been a long time now since attention was drawn to a decline and crisis in statistical science. Nevertheless, no marked improvement has been made. We find it very surprising and odd that no university in either France or Japan has a department of statistics. Most of the students of statistical analysis or data analysis are enrolled in such departments as information science, biometrics, psychology, and the like, and there they are engaged in their studies. In Japan, in the field of statistical science or data science, we have only one specialized research institute, ISM, and as for graduate schools, we have only one statistical science course, also at the ISM.

In recent times in Japan, there has been a great deal of discussion over the guidelines for scientific research. In the fields of computational science and informatics, many thought it necessary to look over how to advance the research. In the course of this review process, a lot of research projects overseas are being introduced for the purpose of comparison or benchmarking. Models drawn from large-scale national research centers, such as INRIA or the organization of CNRS have drawn much interest. In the field of data analysis, a large-scale institute of computational sciences or informatics is planned as part of the structural reorganization program. We are hoping that this new development will come to fruition, and, at the same time, we are looking to develop a new system.

At the moment, however, no definite plan or idea is to be found. It is necessary for us to decide in what direction we should progress. We must re-examine our attitude toward our research and make ourselves refreshed. For that purpose, we might have to seek collaboration and cooperation with other fields, or even consider the possibility of organization and integration. We might have to abandon such terms as statistical science or data analysis, and choose, for example, "Data Science" as a new keyword and concept. We believe that such a concept can help to guide and foster a fruitful and expanding relationship between both countries in the future.

We very much hope this new age of "Data Science" will come to fruition, and that what we have done in the history of data analysis research will be of enduring benefit to the coming science and to future research.

Acknowledgements

I would like to express my grateful acknowledgements to all staff and members of the Organization Committee of the 20th Anniversary of SFC for giving me the opportunity to present this report, and to all the Japanese and French researchers for making great efforts toward the development of Data Science.

References

(1) E.Diday, L.Lebart, J.-P.Pages, R.Tomassone (eds) (1979): Second international symposium on data analysis and informatics, October 1979, Versailles. *Data Analysis and Informatics*, North-Holland.

(2) E.Diday, M.Jambu, L.Lebart, J.-P.Pages, R.Tomassone (eds) (1983): Third international symposium on data analysis and informatics, October 1983, Versailles. *Data Analysis and Informatics III*, North-Holland.

(3) E.Diday, Y.Escoufier, L.Lebart, J.-P.Pages, Y.Schektman, R.Tomassone (eds) (1986): Fourth international symposium on data analysis and informatics, October 1985, Versailles. *Data Analysis and Informatics IV*, North-Holland.

(4) E.Diday (ed) (1989): Proceedings of the conference on "*Data Analysis, Learning Symbolic and Numeric Knowledge*," Antibes, September 1989. Nova Science Publishers.

(5) C.Hayashi, E.Diday, M.Jambu, N.Ohsumi (eds) (1988): *Recent Developments in Clustering and Data Analysis* -Developpements recents en classification automatique et analyse des données-, Academic Press, Boston; Proceedings of the Japanese-French Scientific Seminar, March 1987.

(6) Y.Escoufier, C.Hayashi, B.Fichet, E.Diday, L.Lebart, N.Ohsumi, Y.Baba (eds) (1995): *Data Science and its Applications* -La science des données et ses applications-. Academic Press, Tokyo; Proceedings of Japanese-French Scientific Seminar, August 31-September 2,1992, Montpellier, France.

(7) C.Hayashi, N.Ohsumi, H.H.Bock and others (eds) (1997): *Data Science, Classification and Related Methods*, Springer-Verlag, Tokyo (in printing); Proceedings of the IFCS-96 in Kobe, March 1996.

(8) N.Ohsumi (1980): "Analyse des données" in France (in Japanese), *Mathematical Science*, No.204, p56-64.

(9) J.-P.Benzécri (1982): *Histoire et préhistoire de l'analyse des données*, Dunod-Bordas, Paris.

(10) M.Jambu and M.-O.Lebeaux (1983): *Cluster Analysis and Data analysis*, North-Holland.

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

(12) L.Lebart, A.Morineau, and K.M.Warwick (1984): *Multivariate Descriptive Statistical Analysis - Correspondence analysis and related techniques for large matrices-*, John-Wiley.

(13) N. Ohsumi, L.Lebart and others (1994): *Multivariate Descriptive Statistical Analysis* (in Japanese), JUSE Press, Ltd.