Preface

Relative to the basic notions of a descriptive attribute (variable) and an object described, there are two fundamental concepts in *Data Analysis*: association between attributes and similarity between objects. Given the description of objects by attributes the goal of data analysis methods is to propose a reduced representation of the data which preserves as accurately as possible the relationships between attributes and between objects. Mainly, there are two types of methods: *factorial* and *clustering*. *Factorial* methods are geometric. For these, the compression structure is obtained from a system of synthetic axes, called *factorial axes*. The most discriminant of them are retained in order to be substituted for the origin axes. Then, the set of data units (objects and also attributes) is represented by a cloud of points placed in the geometrical space, referring to the new system of axes. *Clustering* methods are combinatorial. The compression structure consists of an organized system of proximity clusters. In our terminology, an equivalent term for clustering is *classification*.

In our approach clustering (Classification) is considered as a central tool in data analysis. The extensive development of this principle has led to a very rich methodology. According to this standpoint the first facet of clustering concerns the organization of the attribute set. This enables us to discover the behavioural tendencies and subtendencies of the population studied from a sample of it, the latter defining the object set. The second facet concerns the proximity organization of the object set or a category set induced from it. Behaviour understanding is provided by the first facet and management control by the second facet. Geometrical factorial analysis is often considered as a special tool of data analysis for attribute set structuration. Clustering attributes is a non-classical subject in the literature on data analysis. Generally, the methods proposed for this problem consist of adapting methods created for clustering an object set. By distinguishing clearly the two dual problems: attribute clustering and object clustering, our approach is essentially different.

This book provides a large synthesis and systematic treatment in the area of clustering and combinatorial data analysis. A new vision of this very active field is

given. The methodological principles are very new in the *data mining* field. All types of data structures are clearly represented and can be handled in a precise way: qualitative data of any sort, quantitative data and contingency data. The methods invented have been validated by many important and big applications. Their theoretical foundations are clearly and strongly established from three points of view: logical, combinatorial and statistical. In this way, the respective rationales of the distinct methods are clearly set up.

As expressed above, the special structure we are interested in for a reduced representation of the data is that obtained by clustering methods. A non-hierarchical clustering algorithm on a finite set E, endowed with a similarity index, produces a partition on E. Whereas a hierarchical clustering algorithm on E produces an ordered partition chain on E. This book is dominated by hierarchical clustering. However, methods of non-hierarchical clustering are also considered (see below).

In Chap. 1 we study some formal and combinatorial aspects of the sought mathematical structure: partition or ordered chain of partitions. More particularly, two sides are developed. The first is enumerative and consists of counting chains in the partition lattice or counting specific subsets in the partition set. In order to relate the partition type and the cardinality of the equivalence relation graph associated with it, we are led to address the set organized of an integer partition. The second important side concerns the mathematical representation of a partition and, more generally and importantly, an ordered chain of partitions on a finite set E. Thereby, the relationships between the latter structure and numerical (rep., ordinal) *ultrametric spaces* are established. In fact, all the algorithmic development of a given clustering method is dependent on the representation adopted. We end Chap. 1 by showing the transition between the formalization of symmetrical hierarchical clustering and that of directed hierarchical clustering, where junctions between clusters are directed according to a total (also said "linear") order on E.

Our method is focused on *ascendant agglomerative hierarchical clustering* (AAHC). However, non-hierarchical clustering plays an important role in the compression of data representation. This methodology addresses the problem of clustering an object set and not that of an attribute set. Its philosophy is different from that of hierarchical clustering. In these conditions, we describe in Chap. 2 two fundamental and essentially different methods of non-hierarchical clustering. These reflect two important families of no-hierarchical clustering algorithms. It is a matter of the *"*central*"* partitions of S. Régnier and that of "dynamic clustering" of E. Diday. The latter is derived from a generalization of the "allocating and centring" k-means algorithm, defined by D.J. Hall and G.H. Ball (see references of the chapter concerned). This method is discussed in this chapter. On the other hand, new theoretical and software developments are mentioned.

For the mathematical data representation the descriptive attributes are interpreted in terms of relations on the object set. Thereby, categorical attributes of any sort are represented faithfully. In these conditions, numerical attributes are defined as valued relations. Whereas classical approaches propose a converse reasoning by assigning, more or less arbitrarily, numerical values to categories.

In Chap. 3 we describe the set theoretic and relational representation of the data description. All types of data can be taken into account. Two description levels are considered: objects and categories. For each of the levels, object description and category description, two attribute types are considered depending on the arity of the representative relation on the object set, unary or binary. Notice that the arity of the representative relation associated with a given attribute can be greater than two. And this, is also considered in our development. Thus, in this framework, we define several structured attributes concerned by observation of real data.

The fundamental concept of resemblance between data units: attributes, objects or categories, is studied in Chaps. 4–7. It is based on a deep development of a similarity notion between combinatorial structures. Invariance properties of statistical nature are set up. These lead to a constructive and unified theory of the resemblance notion. Classical association coefficients such that the Goodman and Kruskal, Kendall and Yule coefficients are clearly stood in the framework of this theory. Two options are considered for normalization of the association coefficients between descriptive attributes: standard deviation and maximum. A probability scale, associated with the first normalization, is built in order to compare association coefficients between attributes or similarity indices between objects (resp., categories). This scale is obtained by associating independent random data with the observed one, the random model respecting the general characteristics of the data observed. This comparison technique is a part of the *likelihood linkage analysis* (LLA) clustering method where an observed value of a numerical similarity index is situated with respect to its unlikelihood bigness. Well-know non-parametric statistical theorems are needed for the application of this approach to the attribute comparison. New theorems are established. Based on the same principle an *index of implication* between Boolean attributes is set up. Also, we show how partial association coefficients between structured categorical attributes are built.

Comparing objects described is not equivalent to comparing descriptive attributes. We show in Chap. 7 how the LLA approach enables similarity indices between objects, described by heterogeneous attributes of different types, to be built. We also show how comparing categories is a specific task.

The fascinating concept of "natural" cluster of objects cannot be defined mathematically. Its realization in real cases is expected as a result derived from application of clustering algorithms. Such a cluster is interpreted intuitively. However, it is important to define it as accurately as possible. This definition is necessarily a statistical one. Nevertheless, statistical formalization of a "natural" cluster is very difficult. In Chap. 8 we address this concept. Statistical tools are established for understanding the meaning of such a cluster. For this purpose, initial description is examined for all types of data. Thus, the analysis of a "natural" cluster is essentially analytical. Another way consists of crossing with the target cluster associated with a "natural" cluster, known and discriminant clusters disjoint logically of it, but statistically linked. A "natural" cluster is a part of a "natural" clustering. Generally, this statistical structure sustains real data. However, it is important to test this hypothesis for the data treated. In these conditions, "classifiability" testing hypotheses are proposed and studied.

Whereas Chap. 8 is focused on the intrinsic analysis of clustering, Chap. 9 is devoted to comparing clusterings or clustering trees on the same finite set endowed with a similarity or dissimilarity index. In this chapter very powerful tools are established for this comparison. In this, the similarity data is either numerical or ordinal. A minute analysis of the comparison criteria for both types (numerical or ordinal) is provided. The criteria proposed have a combinatorial and non-parametric statistical nature and they are extremely general. They are established with respect to a probabilistic independence hypothesis between similarity and clustering structures. This enables us to establish significant and non-biased comparisons.

As mentioned above, AAHC is considered in this book as a main tool for *data analysis*. Starting with similarities or distances between data units (See Chaps. 4–7) we show in Chap. 10 how to build a classification tree on the data set corresponding to an agglomerative technique. Ordinal notion of pairwise similarities is treated first. Natural transition to a numerical version of this notion is shown. Defining a dissimilarity between disjoint subsets of the set to be clustered is a fundamental task in agglomerative hierarchical clustering. This dissimilarity is established from the pairwise dissimilarities of data units. Two families of dissimilarity indices are studied. The first is classical and employs distances and weightings. The second is defined from probabilistic indices obtained in the context of the LLA approach. The numerical dissimilarity indices between disjoint subsets of the data set enable comparisons between the clusters merged to be made. The algorithmic analysis of the clustering tree construction is a very important problem. Fundamental results for this problem are reported in this chapter. Thus, we describe some basic solutions provided for agglomerative hierarchical clustering of large data sets. Their computational complexities are expressed. We end this chapter by showing the transition between the usual symmetric hierarchical clustering and that directed where junctions between the branches of the hierarchical tree are compatible with a total order on the set clustered.

In Chap. 11 we begin by describing the *Classification Hiérarchique par Analyse de la Vraisemblance des Liens* (CHAVL) software. The address of a link is specified in the References section in order to access this software. The latter performs according to the LLA methodology, the AAHC of a descriptive attribute set or, dually, a described object (resp., category) set; and this, for a large family of data table structures. In this chapter the results obtained by the LLA method on many real cases are reported. These are provided from different areas: psychosociology, sociological surveys, biology, bioinformatics, image data processing, rural economy. The LLA hierarchical clustering method is applied in order to discover "natural" clusters and behavioural tendencies in the population observed. The cluster interpretation is based on the coefficients developed in Chaps. 4–8. In some of these cases, comparison of the LLA results with those of the Ward hierarchical clustering method, is expressed. In order to realize the different facets in applying the LLA method, some presentations of the processed real cases are detailed sufficiently.

The book ends with Chap. 12 devoted to a general conclusion in which several routes for future research works are outlined. Moreover, the contribution of the

book to challenges and advances in cluster analysis is clearly specified. Further, in this chapter, the situation of the book content with respect to other books in the same field is described.

The starting point of the project of this book was a reviewed and completed English translation of the French book:

Classification et analyse ordinale des données

published—with the support of the CNRS—by Dunod (Paris) in 1981.

The progress of my research, the works I met and the considerable development of the field concerned have made that a single volume cannot suffice to cover the entire material expressed in the French book.

In the book we propose here, symmetrical synthetic structures for summarizing data are considered. For these structures—defined by partitions or partition chains—if x and y are two elements of the set E to be organized, the role of x with respect to y is identical to that of y with respect to x.

The different steps of the passage from the data table to the synthetic structure (partition or partition chain) on E are minutely studied. Recall that the set E to be clustered may be an attribute set or an object set (resp., a category set).

The book we propose is a *new* book. It corresponds with respect to the earlier French version, to a new writing, a new design and a much larger scope and potential. The intuitive introductions, the examples and the mathematical formalization and analysis of the subjects treated permit the reader to understand in depth the different approaches in data analysis and clustering. Special concern is devoted for expressing the relationships between these approaches. More precisely, the development provided in this book has the following general distinctive and related features:

1. Mathematical and statistical foundations of combinatorial data analysis and clustering;
2. Mathematical, formal conception and properties are set up in order to compare different approaches in the field concerned;
3. Definition of new methods, guided by a few fundamental principles taking into account the formal analysis;
4. Applying new methods to real data.

More specific distinctive features might be listed as follows:

- Formal descriptions and specific mathematical properties of the synthetic structures sought in clustering (partitions, partition chains (symmetrical and directed));
- Emphasizing data description by categorical attributes of different sorts (broad scope);
- Interpreting descriptive attributes in terms of relations on the object set described;
- Set theoretic representation of the relations defined by the descriptive attributes;
- Very clear typology of data description in the most general case;

- Development of a unified association coefficient notion (symmetrical and asymmetrical) between descriptive attributes of different sorts, including all types of categorical attributes;
- Development of a similarity notion between objects or categories for different types of description, including all types of categorical attributes;
- Probabilistic similarity measures between objects, object clusters, categories, category clusters, attributes, attribute clusters, …;
- Clustering numerical or categorical descriptive attributes of different kinds;
- Clustering data units (objects or categories) described by a mixing of descriptive attribute types;
- Dual association between object clustering and attribute clustering;
- Seriation and clustering;
- Combinatorial and non-parametric statistical basis for the association coefficients, similarity indices and criteria in clustering;
- Algorithmic studies.

In the part of the French book not retaken here the synthetic structures summarizing the data are of *asymmetrical* nature. Ordinal considerations take part. The chapters concerned with the latter, which may constitute a second volume, are: 6–10. Let me give briefly the subject of each of them.

- Chapter 6: Principal component analysis and correspondence analysis;
- Chapter 7: Mathematical comparisons between *factorial* analysis and *classification* methods;
- Chapter 8: From combinatorial and statistical *seriation* methods to a family of cluster analysis methods;
- Chapter 9: Totally ordering the whole set of categories associated with a set of ordinal categorical attributes;
- Chapter 10: Assignation problems in *pattern recognition* between geometrical figures where the quality measure of the assignation has to be independent of specific geometrical transformations applied on the figures concerned.

As indicated in the title of the book, our work refers to *Combinatorial and Statistical Data Analysis*. The importance of this methodology has already been underlined in the well-known article "Combinatorial Data Analysis" by Phipps Arabie and Lawrence Hubert, published in 1992.

This book is not conceived *a priori* as a "text book". It is a result of my research led since 1966, with many collaborators (See below). Thus the main orientation is "research". However, the latter is placed in the framework of the entire domain concerned. Moreover, a very important part of this research is oriented towards the foundation and synthesis of different methods in combinatorial data analysis and clustering. Consequently, this book is a *reference book*. It will be very useful to master's and Ph.D. students. Wide parts of this book can be taught to students of computer science, statistics and mathematics. I did it.

Let me now cite, in alphabetic order, the names of different collaborators who have worked with me and participated in this research. Most often, but not always,

they were around preparing theses and subsequent articles. I especially thank them. The theses defended at the University of Rennes 1 can be consulted at the link address: Sadoc.abes.fr/Recherche avancée.

Collaborators

Jérôme Azé, Helena Bacelar-Nicolaü, Kaddour Bachar, Jean-Louis Buard, Thierry Chantrel, Isaac Cohen-Hallaleh, Jean-Louis Cotrieux, François Daudé, Aziz Faraj, Jean-Paul Geffrault, Nadia Ghazzali, Régis Gras, Sylvie Guillaume, Ivan Kojadinovic, Pascale Kuntz, Jean-Yves Lafaye, Georges Lecalvé, Alain Léger, Henri Leredde, Jean Rémi Massé, Annie Moreau, Roger Ngouënet, Fernando Nicolaü Da Costa, Mohammed Ouali-Allah, Philippe Peter, Joaquim Pinto Da Costa, Annick Prod'Homme, Habibullah Rostam, Valérie Rouat, François Rouxel, Abdel Rahmane Sbii, Basavaneppa Tallur and Philippe Villoing.

http://www.springer.com/978-1-4471-6791-4

Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering

Lerman, I.C.

2016, XXIV, 647 p. 54 illus., Hardcover
ISBN: 978-1-4471-6791-4