Jul 1st, 12:00 AM

Ranking regions using cluster analysis, Hasse diagram technique and topology

Guillermo Restrepo

Rainer Brüggemann

Ranking regions using cluster analysis, Hasse diagram technique and topology

G. Restrepo[a,b] and R. Brüggemann[c]

[a] *Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia*

[b] *Environmental Chemistry & Ecotoxicology, University Bayreuth, D-95440 Bayreuth Germany*

[c] *Department Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries,Müggelseedamm 310 , D-12587 Berlin*

Abstract: We developed the theoretical background of the application of Hierarchical Cluster Analysis HCA to the improvement of the understanding of Hasse Diagrams coming from the ranking process developed for the application of the Hasse Diagram Technique HDT. The use of HCA is based on the idea of the reduction of the number of elements considered in a poset (partially ordered set) by the selection of representatives from the original set to rank. We showed that the clusters arisen from HCA can be interpreted as similarity classes where one of its members (the nearest to the centre of the cluster) can be selected as representative of its class. We applied this procedure to the reduction of a set of 59 regions of Baden Wuerttemberg, Germany, monitored with respect to Pb, Cd, Zn and S pollution in the herb layer. After applying HCA we found 26 representatives and we drew the Hasse diagram of this set of representatives. Finally, we describe the mathematical procedure to endow a partially ordered set with a topology taking advantage of the network structure of a Hasse diagram where the comparisons of the elements are considered as the open sets of a topological basis. Using such a basis we calculated the closure of two subsets of the set of 26 regions (big traffic or sensitive ecosystems and industrialised regions). We showed how can be interpreted the results of the closures of these subsets

Keywords: Partially ordered sets, Cluster analysis, Hasse diagram technique, Topology, Mathematical chemistry

1. INTRODUCTION

Ranking of objects is an important tool for decision support systems and due to the fact of its mathematical background it is possible to apply it in whatever field of knowledge where the aim is to endow a set of objects with "priorities." Brüggemann and co-workers [1999,2001a,2003,2004] have shown how chemical and environmental sets of objects can be ranked using the mathematical ideas of partially ordered sets (posets). The advantage of this methodology is the use of a visualization technique called the Hasse diagram technique HDT. An important feature of HDT is the lack of an ordering index that permits to avoid subjectivities in the final rank caused by the selection of the index. Another characteristic of Hasse diagrams is they not only present information on the ranking but also show whether the criteria, characterising the objects, lead to ambiguities in the ranking. For instance, an object might be ranked higher according to one attribute but lower according to another. These two objects are not strictly ordered because their data are "contradictory," which causes an ambiguity in an attempt to construct a total order and usually is hidden when one uses an index for ranking [Brüggemann et al., 2001a]. However, this ambiguity is easy to observe through the HDT for the presence or absence of lines in the diagram [Brüggemann et al, 2001a]. Since HDT is based on diagrams representing the network of order-comparabilities among the objects of a set, the understanding of such relationships can be difficult if the cardinality of the set under consideration is high and if the elements in it have many comparabilities (many links). A way to solve this shortcoming, it is to get more understandable a Hasse diagram, is to try to reduce the number of objects to be compared by deleting objects similar to other ones and calculating the Hasse diagram

over the reduced set of objects [Brüggemann et al., 2001b]. This idea suggests the application of a classificatory system over the original set of objects in order to find similarity classes from which can be extracted representatives. Thus, the Hasse diagram is drawn over the set of representatives and it can be used to extrapolate the order relations found for each representative to the members of their respective classes [Restrepo and Brüggemann, 2005]. In this paper we develop this idea and also we applied the concept to a practical example.

Now, if we concentrate our attention in the interpretation of a Hasse diagram [Brüggemann, 1999,2001a,2003,2004], we realise that in the core of such a interpretation are the order relationships shown by the Hasse diagram. However, another point of view of a Hasse diagram is the use of general topology concepts. A Hasse diagram is a directed graph where each line is an ordered relationship among the points linked by the line. In that case, given a point of the diagram it is possible to know its neighbours in terms of order relationships. In other words, it is possible to know, given a point, which are the points "greater" or "lower" than it. Thus, the set of such a points constitutes the neighbourhood of the selected point. Having these neighbourhoods extracted from the Hasse diagram, it is possible to build up a basis for a topology in order to study some topological properties of particular subsets of the original set of objects. The importance of the topological study of a set is that given such a mathematical structure it is possible to study local parts of the Hasse diagram keeping the order relationships.

2. METHODS

The discussion sets up in this paper considers the application of hierarchical cluster analysis HCA for the reduction of the set of work, the application of the HDT and finally the use of topological ideas for the study of some subsets of interest. For that reason, in the following, we briefly describe the general aspects of each one of these topics and we also show a particular example of application of this procedure to a set of regions of Baden Wuerttemberg, Germany, monitored regarding four pollutants.

2.1 Hierarchical Cluster Analysis HCA

We introduce, for the sake of clarity, some useful definitions.

Definition 1. Let $Q=\{c_i \mid i \in I\}$ the ground set where the index set I is the set of positive integers and every c_i is an object.

Definition 2. Let $IB=\{q_j \mid j \in I\}$ the set of properties where every q_j is a property (The symbol *IB* comes from the name information basis mentioned by Brüggemann et al. [2001a]).

Definition 3. We characterize each object i by a map $\varphi: i \in Q \rightarrow \{q_j(c_i) \mid j \in I\}$ and write simply c_i.

The procedure of HCA considers the application a similarity function (SF) among the c_i's $\in Q$ and a grouping methodology (GM). The final result of this procedure is a graphical representation called dendrogram in which clusters of objects are shown with similarities among their properties q_j. There are several methods for selecting clusters from a dendrogram [Wild and Blankley, 2000]. However, the majority of them consider a dendrogram as a geometrical object where the distance between objects and also between possible clusters is the cornerstone of the methods. Instead of considering a dendrogram as a geometrical object, Restrepo and co-workers developed a method for determining the clusters in HCA [Restrepo et al., 2005], considering a dendrogram as a topological object where the important facts between the elements and possible clusters are not their distances but their similarity relationships. The method for selecting clusters is based on the idea of looking for highly populated and homogeneous clusters. It calculates $1 \leq n \leq |Q|$ possible cuts over the dendrogram measuring the resemblance of each partition of Q generated by each cut with the square partition of the set (further details about this method are found in Restrepo et al., 2005).

Now if we interpret in mathematical terms the process of clustering we find that each cluster is a similarity class [Restrepo and Brüggemann, 2005]. An advantage of the construction of similarity classes is the reduction of the time of analysis because the objects belonging to a particular cluster share their properties (they are similar). In this case, if we need to study the behaviour of the objects in Q, then we can select one object from each cluster as representative of its cluster. According to Bock [1974] a criteria for selecting representatives is the choice of central points of the cluster. Thus, the reduction of the set of work to a set of representatives is one of the methods used to do more understandable Hasse diagrams.

2.2 Ranking representatives through HDT

Once we have the representatives of each class we can rank the c_i's $\in Q$ according to their properties q_j's $\in IB$ as follows:

Definition 4. Let c_1, $c_2 \in Q$ and $q_j \in IB$, then $c_1 \leq c_2$ if $q_j(c_1) \leq q_j(c_2)$ for all q_j is called a product- (or component-wise-) order if it obeys the following axioms of order:

i) reflexivity (an object can be compared with itself)

ii) antisymmetry (if an object c_1 is better than c_2, then c_2 is worse than c_1),

iii) transitivity (if $c_1 < c_2$ and $c_2 < c_3$ then $c_1 < c_3$; $c_1, c_2, c_3 \in Q$).

If only one property is regarded, then for all objects a $\leq$ relation can be established. In general, when more than one property is regarded it is possible to have several properties not perfectly correlating; hence the product order cannot be established for -at least- some objects. We called such objects "incomparable" [Brüggemann et al., 2001a]. A directed graph of the comparrabilities and incomparabilities of the elements of a set, where relations due to transitivity are deleted, is a Hasse diagram. In order to avoid the directed links in the graph, it can be drawn in a way that an object c_x worse than c_y is located above c_y [Brüggemann et al., 2001a; Trotter, 1991].

2.3 Topological interpretation of a Hasse Diagram

Once we have the Hasse diagram it is possible to rank the objects according to the partial order shown by the diagram. But the network structure of the diagram suggests the possibility of studying relational properties of the objects in the dendrogram. These properties are order ones and can be studied defining a topological basis over the whole set of objects. In general, a topological basis contains the information about the relationships among the objects of a set. In our case the relationship is a comparability and it can be used to define neighbourhoods for the c_i's $\in Q$.

Definition 5. Let **B** be a collection of subsets of a non-empty set *Q*, such that:

1. $Q = \cup_{B \in \mathbf{B}} B$

2. If $B_1, B_2 \in \mathbf{B}$, then $B_1 \cap B_2$ is the union of elements of **B**, then **B** is called a *basis for the topology* τ, where $\tau = \left\{ \cup_{B \in F} B \mid F \subseteq \mathbf{B} \right\}$.

The elements of a topological basis are called open sets and in our case each object of the set has a collection of open sets, which we define to consist of all comparable elements. In other words, the open sets of an element in a Hasse diagram are all the possible chains we can extract from the diagram where the element under consideration is present. Formally, we write:

Proposition 1. Given a set *Q* and a Hasse diagram *P* over it, let $\mathbf{B} = \{B \in Q \mid B \text{ is a chain in } P\}$, then **B** is a basis for the topology τ_P, where τ_P stands for a topology over the Hasse diagram *P*■

Having built up the topological basis we can calculate some topological properties of subsets of interest on *Q*, in this paper we consider just one topological property, the closure of a subset.

Definition 5. Let $A \subset Q$ and $c \in Q$, *c* is said to be a *closure point* of *A* iff for every $O \in \tau$, such that $c \in O$, then $O \cap A \neq \varnothing$.

Let $A \subset Q$; the *closure* of *A* is defined as: $\overline{A} = \{c \in Q \mid c \text{ is closure point of } A\}$

The importance of the closure of a set *A* is that *A* can be selected as we want. It is not necessary that the elements of *A* belong or not to a chain in the Hasse diagram. Thus, this topological property allows us to know the comparable objects with reference to a subset *A* of interest. We show in the following the application of this procedure.

3. RESULTS

The set *Q* is a set of 59 regions of Baden Wuerttemberg, Germany, monitored with respect to Pb, Cd, Zn and S pollution in the herb layer (the labels of the regions are 1-54 and 56-60). Thus, each region, the set *Q* and *IB* could be defined as follows:

Definition 6. Let $IB = \{q_j \mid j \in \{\text{Pb,Cd,Zn,S}\}\}$ the *set of properties* (information base).

Definition 7. Let $c_i = \{q_j(c_i) \mid j \in \{\text{Pb,Cd,Zn,S}\}\}$ the i-th *region*, where $q_j(c_i)$ means the property q_j of c_i.

Definition 8. Let $Q = \{c_i\}$ the *set of regions*.

We applied HCA using the Euclidean distance (1) as SF and the unweighted average linkage (2) as GM. In this way we obtained the dendrogram shown in figure 1.

$$d(c_k, c_l) = \left\langle \sum_{j=1}^{4} | q_{jk} - q_{jl} |^2 \right\rangle^{1/2} \qquad (1)$$

$$f(r,s) = \frac{n_A}{n_A + n_B} f(A,s) + \frac{n_B}{n_A + n_B} f(B,s) \qquad (2)$$

A,B and s means the regions c_i to group, r is the reunion of A and B; $f(A,s)$, $f(B,s)$ and $f(A,B)$ are the Euclidean distances between A and s, B and s and A and B, respectively [Otto, 1999].

Then, we looked for the most homogeneous clusters and we found the optimal value for n=4. The similarity classes obtained appear in table 1.

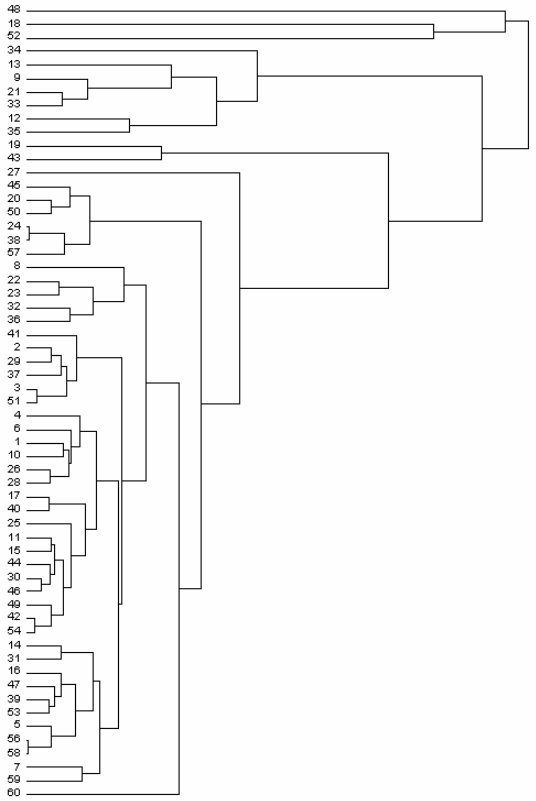

Figure 1. Dendrogram of the 59 regions studied.

Table 1. Classes obtained for the set Q and their representatives.

Class	Repres.	Class	Repres..
{48,18,52}	18	{4}	4
{34}	34	{6}	6
{13,9,21,33}	13	{1,10,26,28}	28
{12,35}	12	{17,40}	17
{19,43}	19	{25}	25
{27}	27	{11,15}	11
{45,20,50}	20	{44,30,46}	46
{24,38,57}	24	{49,42,54}	42
{8}	8	{14,31}	14
{22,23,32,36}	23	{16,47,39,53}	53
{41}	41	{5,56,58}	56
{2,29,37}	29	{7,59}	59
{3,51}	51	{60}	60

The selection of a representative per class was achieved calculating the Euclidean distance (1) between each member of each class and the centroid of the class. We selected as representative of the class the furthest region to the centroid. In the case of having classes of just two regions, we calculated the Euclidean distance (1) among each member of the class and the centroids of the other classes. Once again, we selected the furthest region. The representatives of the 26 classes appear in table 1.

The next step of our study was the construction of a Hasse diagram taking into account the properties q_j of the representatives selected. The objects were organised in four levels (Figure 2). In comparison to the Hasse diagram, containing 59 regions (not shown) a by far more simple Hasse diagram was obtained by applying HCA.

It can be seen that there are 13 maximal and 16 minimal regions. The maximal regions are more contaminated than the others and the contrary can be said about the minimal regions. There are two regions with an intermedia ranking of contamination, they are 29 and 20. It is important to remark that this ranking was developed using 4 pollutants and for that reason it is not possible to say that determined region is more contaminated regarding one of the pollutants. However, it is possible to extract such a particular information from the same diagram adding fictitious regions representing the mean value of the j-th attribute. The application of this procedure is shown by Restrepo and Brüggemann [Restrepo and Brüggemann, 2005].

Now, in order to apply the topological study discussed above, we built up the topological basis **B** for the Hasse diagram.

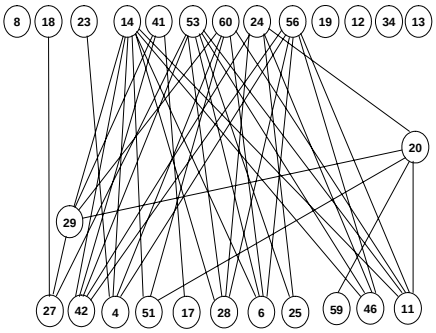

Figure 2. Hasse diagram of the 26 representatives.

$$\mathbf{B} = \left\{ \begin{array}{l} \{27\},\{42\},\{4\},\{46\},\{29\},\{28\},\{51\},\{59\},\{11\},\{17\}, \\ \{6\},\{25\},\{20\},\{8\},\{18\},\{23\},\{14\},\{41\},\{53\},\{60\}, \\ \{24\},\{56\},\{19\},\{12\},\{34\},\{13\},\{27,18\},\{27,29\}, \\ \{27,53\},\{42,14\},\{42,41\},\{42,53\},\{42,24\},\{42,56\}, \\ \{4,23\},\{4,14\},\{4,60\},\{4,56\},\{46,14\},\{46,53\}, \\ \{46,24\},\{46,56\},\{29,14\},\{29,41\},\{29,60\},\{29,20\}, \\ \{28,14\},\{28,53\},\{28,24\},\{28,56\},\{51,14\},\{51,60\}, \\ \{51,20\},\{59,20\},\{11,14\},\{11,53\},\{11,60\},\{11,56\}, \\ \{11,20\},\{17,41\},\{6,14\},\{6,53\},\{6,60\},\{6,56\}, \\ \{25,53\},\{25,24\},\{20,24\},\{27,29,14\},\{27,29,41\}, \\ \{27,29,60\},\{27,29,20\},\{51,20,24\},\{59,20,24\}, \\ \{11,20,24\},\{27,29,20,24\} \end{array} \right\}$$

Where every element of the basis is a chain of the Hasse diagram and its elements are ordered according tho the order relation used to build up the diagram. Thus, the first element of each open set of the basis is lower or less contaminated than the next element.

Now we selected some subsets of *Q* of particular interest.

Definition 9. Let I={49,43,44,1,47,30,3,56} the set of high industrialised regions.

Let *T*={20,59,31,52,30} the set of regions with big traffic (also neighborhood of highways to sensitive ecosystems).

For the sake of simplicity we represent directly the sets (closures) as sub-Hasse diagrams from the original one (Figure 3-4).

The meaning of such a topological property can be explained as follows: Given a subset, every element belonging to its closure is comparable to the elements of the subsets. In other words, every closure point is more contaminated or less than one element of the subset under consideration.

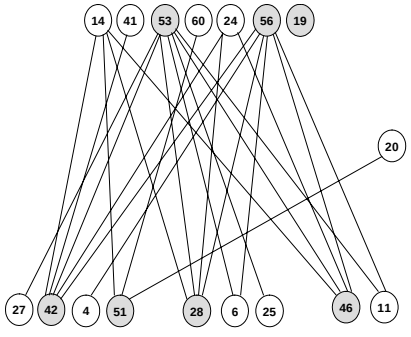

Figure 3. Closure $\overline{I}$ of the set of industrialised regions, in grey the members or the corresponding representatives of *I*.

Regarding the subset of industrialised regions, there are some objects from *Q* {14,41,60,24,6,25,20,11,27,4} which are not considered as industrialized regions but also appear related to them. It is important to remark that 5 out of 10 of these regions appear more contaminated than regions 42, 28, 51, 46 and 19. Note that 19 is considered either a maximal or a minimal object (trivial case) since its lack of relationships.

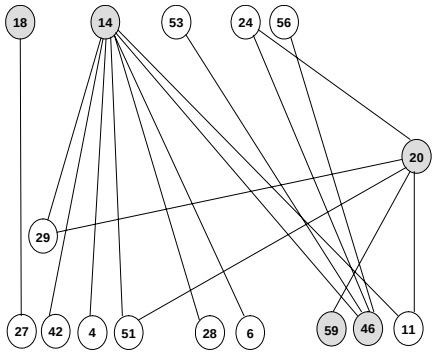

Figure 4. Closure $\overline{T}$ of the set of regions with big traffic or sensitive ecosystems, in grey the members or the corresponding representatives of elements of *T*.

Finally, the subset of regions with big traffic shows a closure of 16 regions where 11 are regions not considered at the beginning as big traffic ones or with sensitive ecosystems. It is important to note that regions {53,56} and {24} are more contaminated than regions 46 and 20, respectively. Especially, region 24 is more contaminated than regions 46 and 20 simultaneously. Since 20 is more contaminated than 59 (a minimal region) then 24 becomes an important closure region of the set of traffic regions and it would be interesting to explore it as well as its ecosystem.

4. CONCLUSIONS

We show how can be combined HCA and HDT in order to do more understandable the Hasse diagrams obtained through the HDT. HCA was used to classify the objects of the set of work Q in order to find similarity classes. Afterwards we selected a representative of each class and we drew the Hasse diagram over the reduced set of representatives. We apply this methodology to the analysis of 59 regions of Baden Wuerttemberg, Germany, monitored with respect to Pb, Cd, Zn and S pollution in the herb layer. In this particular example we dropped the cardinality of the set of work from 59 to 26. We also describe how topological ideas can be applied to subsets of interest of the set of work. This concept was developed considering the network structure of the Hasse diagram as a collection of neighbourhoods where each element has at least one neighbourhood and in the general case each element has a set of neighbourhoods corresponding to all the chains containing it. We

show how those collections of chains constitute a basis for a topology. Furthermore we used such a topological basis as the source of information for calculating the closure of particular subsets of the set of work. We understood by closure all the elements of the Hasse diagram that are comparable to the elements of the subset under consideration, here a close connection Here an interesting connection between the concepts of topology (basis and closure) and concepts of order theory (order ideals and filters as realizations of closures) is opened with interesting practical applications. Finally we apply this topological idea to the study of the Hasse diagram of 59 regions considering two kinds of subsets. The first one was the set of industrialised regions and the other one was the set of regions with big traffic or sensitive ecosystems. The closure of these subsets showed that both sets (industrialised and big traffic regions) have elements comparable to other regions not considered as industrialised or traffic regions. The set of industrialised regions showed 5 regions more contaminated than the less contaminated of this subset. Regarding the set of regions with big traffic, its closure showed there is a region (24) more contaminated than regions 46 and 20 simultaneously (these two considered as big traffic regions). Finally, as a general comment is important to stress the fact that the topological point of view of a Hasse diagram is a versatile interpretative tool of Hasse diagrams since it is based one the comparisons set up in the Hasse diagram and allow to the analyst to study particular subsets of interest and also their order or ranking behaviour.

5. ACKNOWLEDGEMENTS

G. Restrepo particularly thanks to COLCIENCIAS and the Universidad de Pamplona in Colombia for their financial support for developing this research.

6. REFERENCES

Brüggemann, R., and Bartel, H.G. A Theoretical Concept To Rank Environmentally Significant Chemicals, *J. Chem. Inf. Comput. Sci*. 39, 211-217, 1999.

Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., and Steinberg, C.E.W. Applying the Concept of Partially Ordered Sets on the Ranking of Near-Shore Sediments by a Battery of Tests *J. Chem. Inf. Comput. Sci*. 41, 918-925, 2001a.

Brüggemann, R., Halfon, E., Welzl, G., and Voigt, K. Order Theoretical Tools for the Evaluation of Complex Regional Pollution Patterns *J. Chem. Inf. Comput. Sci*. 43, 1771-1779, 2003.

Brüggemann, R., Sørensen, P.B., Lerche, D., and Carlsen, L. Estimation of Averaged Ranks by a Local Partial Order Model. *J. Chem. Inf. Comput. Sci.* 44, 618-625, 2004.

Brüggemann, R., Welzl, G., Order theory meets statistics – Hasse diagram technique In Voigt, K., Welzl, G. Order Theoretical tools in Environmental Sciences Order Theory (Hasse diagram Technique) Meets Multivariate Statistics, Proceedings of the Fourth Workshop November 2001b in Iffeldorf, Bavaria, Germany.

Restrepo, G., and Brüggemann, R. Ranking regions through cluster analysis and posets. *WSEAS Trans. Inf. Sci. Appl.* 7, 976-981, 2005.

Wild, D.J., and Blankley, C.J. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci*. 40, 155-162, 2000.

Restrepo, G., Llanos, E.J., and Villaveces, J.L. Trees (Dendrograms and Consensus Trees) and their topological information, in Basak, S., and Sinha, D.K. (Ed.), Proceedings of the Fourth Indo-US Workshop on Mathematical Chamistry, University of Pune, Pune, India, 2005, pp. 39-62.

Luque, I., Cerruela, G., and Gomez-Nieto, M.A. Clustering Chemical Databases Using Adaptable Projection Cells and MCS Similarity Values. *J. Chem. Inf. Model*. 45, 1178-1194, 2005.

Bock, H.H. Automatische Klassifikation, Vandenhoeck-Ruprecht, 1974.

Trotter, W.T. Combinatorics and Partially Ordered Sets: Dimension Theory; John Hopkins Series in the Mathematical Science; The J. Hopkins University Press: Baltimore, 1991.

Brüggemann, R., Voigt, K., Kaune, A., Pudenz, S., Komossa, D., and Friedrich, J. Vergleichende ökologische Bewertung von Regionen in Baden-Württemberg, GSF-Bericht 20/98 (Final report of a project, ID 209501.02 of the Landesanstalt für Umweltschutz des Bundeslandes Baden-Württemberg, Germany).

Otto, M. Chemometrics: Statistics and Computer Application in Analytical Chemistry; Wiley-VCH: Weinheim, 1999; pp. 148-156.