Acta Œcologica
Œcol. Gener., 1989, Vol. 10, n° 3, p. 245-256.

The analysis of multiple tables in factorial ecology

III. Three-mode principal component analysis: "Analyse triadique complète"

P. M. Kroonenberg
Department of Education,
Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands

ABSTRACT

THIOULOUSE and CHESSEL's (1987) "partial" triadic analysis to handle multiple tables in ecology can be extended to a complete triadic analysis. This method was already developed by TUCKER (1966) under the name of three-mode factor (or principal components) analysis. This technique is applied to THIOULOUSE and CHESSEL's data on the water quality of the Méaudret. The compact and efficient data condensation of the method is emphasized and illustrated.

KEY-WORDS: multi-table analysis, physico-chemical variables, pollution, spatial-temporal structures, three-mode principal components analysis, three-mode factor analysis, triadic analysis.

RÉSUMÉ

L'analyse triadique « partielle » (THIOULOUSE & CHESSEL, 1987) pour analyser des tableaux multiples en écologie peut être étendue à une analyse triadique complète comme l'avait déjà proposé TUCKER (1966). Cette technique, dite analyse en composantes principales à trois modes, est appliquée à des données de THIOULOUSE et CHESSEL concernant la qualité de l'eau du Méaudret. La condensation compacte et efficace des données par cette méthode est soulignée et illustrée.

MOTS-CLÉS : analyse multi-tableaux, variables physico-chimiques, pollution, structures spatio-temporelles, analyse en composantes principales à trois modes, analyse factorielle à trois modes, analyse triadique.

INTRODUCTION

In a recent ecological study of the river Méaudret, THIOULOUSE and CHESSEL (1987) presented an "analyse triadique partielle", following the procedures described in JAFFRENOU (1978). The data (see their table I; originally in DOLEDEC and CHESSEL, 1987) consist of measurements of water quality with nine variables (see table I) at five stations in four different months (June, August, November, and February). In their discussion of the uses and aims of "analyse triadique partielle", they state that there are two possibilities for treating these data:

– *either* as a chronological series of matrices [stations × variables] (one matrix for each sampled month);

– *or* as a series of matrices [months × variables] (one matrix for each station).

They indicate that (1) different results will be obtained according to the point of view taken, and (2) the choice between the two possibilities is strictly one of the aim of the study. THIOULOUSE and CHESSEL prefer to study the structure of the variables, and the chronological evolution of the upstream-downstream gradient as defined by the geographical location of the five stations. This leads them to choose the first treatment of the data. Their aim is to approach the possible modifications of the upstream-downstream gradient in the perspective of functional ecology. The other treatment of the data would be more concerned with the yearly cycle of measurements on the variables and their spatial variability.

In this paper, we aim to show that the data can be analysed in such a way that both perspectives can be treated simultaneously, and in fact one can even include a third perspective by considering the data:

– as a series of matrices [months × stations] (one matrix for each variable).

This point of view would concentrate on the spatial-temporal relations for each variable.

Such an increase in interpretational multiplicity can only be bought with analytic complexity. In fact, the model to be used has been described by JAFFRENOU (1978, chap. IV), who also shows that his model is in fact identical (p. 72 ff.) with three-mode factor analysis (now generally called three-mode principal component analysis) proposed by TUCKER (1966). In TUCKER's method, a principal component analysis is performed on each of the three arrangements of the data alluded to above, and a weight matrix, commonly called the *core matrix* (JAFFRENOU's matrix **H**), is derived which describes in a very compact way the fundamental relationships between the three approaches.

To be a bit more precise, the aim of the procedure is to derive components for each of the ways or "modes" (*i. e.*, stations, variables, and months); say, P components for the first, Q for the second, and R for the third way. The core matrix will have then order P by Q by R, and it contains the weights, g_{pqr}, assigned to each of the possible combinations of components from the three ways. In particular, g_{pqr} indicates the joint weight for the p-th component of the first way, the q-th component of the second way, and the r-th component of the third way, and its squared value indicates the variation accounted for by that combination of components. The complete model for the data points x_{ijk} may be written as

$$x_{ijk}=\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr}+e_{ijk}$$

with $i=1, \ldots, 5$ stations, $j=1, \ldots, 9$ variables, and $k=1, \ldots, 4$ months, and where the e_{ijk} are the random errors. An observed score x_{ijk} is thus "modelled" as a systematic part of sums of multiplicative terms and a random error part. The a_{ip} are the entries of an $I \times P$ matrix A with the components of the first way, here stations, as its columns. In accordance with the French literature, especially that on STATIS (see, for instance, L'HERMIER DES PLANTES, 1976; LAVIT, 1988), a matrix with such components will be refered to as a "compromis", or compromise solution. The b_{jp} and c_{kr} are similarly defined for the second and third way with B and C the compromise solutions for the variables and months, respectively.

The estimation of this "model" is performed via iterative alternating (or conditional) least squares procedures first described in KROONENBERG and DE LEEUW (1980), and is embodied in the program TUCKALS3 (KROONENBERG and BROUWER, 1985) which can be obtained for a fee via the author (see also Appendix A). Because of the least squares approach used to estimate the parameters, it is possible to indicate how well the variability of each station, variable, and month is represented by the fitted three-mode model (see TEN BERGE *et al.*, 1987). This is analogous to the information described by DOLEDEC and CHESSEL (1987; Table III) for a different kind of model. The proportions explained variability are indicated in tables II, III, and IV, which also present the compromise solutions. For further details on the methods used here see TUCKER (1966) and KROONENBERG (1983).

In passing, it should be remarked, that the increase in analytic complexity has the disadvantage of necessitating the use of specialized programs, whereas the attractiveness of THIOULOUSE and CHESSEL's approach is that their analysis can be carried out using standard PCA programs.

In the analysis to be presented, we will concentrate on the global features of the data, *i. e.* on the compromise solutions for the stations, variables, and months, respectively. In addition, one could also look at the intrastructure, *i. e.* at the relationships of the common structure of one way (say, months) with the levels of the two other ways (say, with the stations and variables) to investigate their changes over time in greater detail. This intrastructure could be described, but this would lead to very similar descriptions as the very detailed ones by THIOULOUSE and CHESSEL, and there is no need to repeat them here.

Comparisons between three-mode principal component analysis, STATIS and various other three-way methods can be found in KROONENBERG (1987), LAVIT (1988), KIERS (1988), and CARLIER *et al.* (1989).

We will scale the data in the same way as THIOULOUSE and CHESSEL, even though we prefer to scale the variables across all stations and months, rather than scaling the variables across stations at each month. Their practice removes the month to month variability which might not be adviseable, because this variability is something that should be explained as well. With their scaling, the scores are measured in different standard units across months which makes them less comparable than might be desirable. It should be noted that the variabilities between months differ considerably, so that different scalings may lead to different results. For the sake of comparability, and to avoid unnecessary complications, we use here THIOULOUSE and CHESSEL's scaling. In other words, our analysis is based on the values in their Table II.

RESULTS

First we will briefly look at the means which are removed before the structural three-mode analysis. The subsequent analyses are then based on the scaled deviations with respect to these means. For instance, we do not compare the absolute temperature difference of a station between two months say June and August, but the *relative position of that station with respect to the mean of June and the* relative position with respect to the mean of August. In other words, a station which would be about average in every month on every variable would have a score of about zero after the centring operation, and all the remaining variability would reflect random variation. Such a station could serve as a reference point for all other stations. In the present study no station really fulfills this function; possibly station C comes closest to it.

TABLE I. — Means and standard deviations of the variables for each month, and standard deviations of the variables across stations and months.

	Means					Standard Deviations				
		Months					Months			
Variables	All	June	Aug	Nov	Feb	All	June	Aug	Nov	Feb
1. Water Temperature	8	11	**14**	2	3	0.9	1.0	1.2	0.7	0.4
2. Water Flow	172	200	108	65	**314**	83	98	37	21	128
8. Nitrates	5.6	3.7	6.0	**8.7**	4.1	2.7	0.6	3.5	4.0	1.2
3. pH	8.3	**8.5**	8.0	8.3	8.2	0.2	0.1	0.3	0.2	0.1
4. Conductivity	336	295	**359**	345	343	26	10	22	43	18
5. DBO_5	7.3	4	**10**	**10**	5	8	2	7	13	3
6. Oxydibility	2.3	1.9	**2.9**	2.5	1.9	1.7	0.7	1.4	2.8	0.8
7. Ammonia	2.5	0.8	**4.6**	3.2	1.4	3.2	1.0	3.9	4.8	1.1
9. Orthophosphates	1.8	0.7	**2.7**	2.6	1.2	1.5	0.5	1.6	2.3	0.9

Note: All=Year mean/standard deviations. **Bold** means indicate the highest recorded mean.

The main trends in the means are as follows. The highest (measured) concentration of the pollutants occur in the month of August and they fall off gradually the *following months reaching a low in June, under the assumption that the data are* reasonably representative of the river's characteristics for any year. Furthermore, the means of pH are virtually stable the year around, nitrates are high in November, water flow reaches its peak in February after jumping from its low in November. Finally, not surprisingly, the water temperature is high in summer and low in winter. The analyses to follow should thus be interpreted with respect to these means.

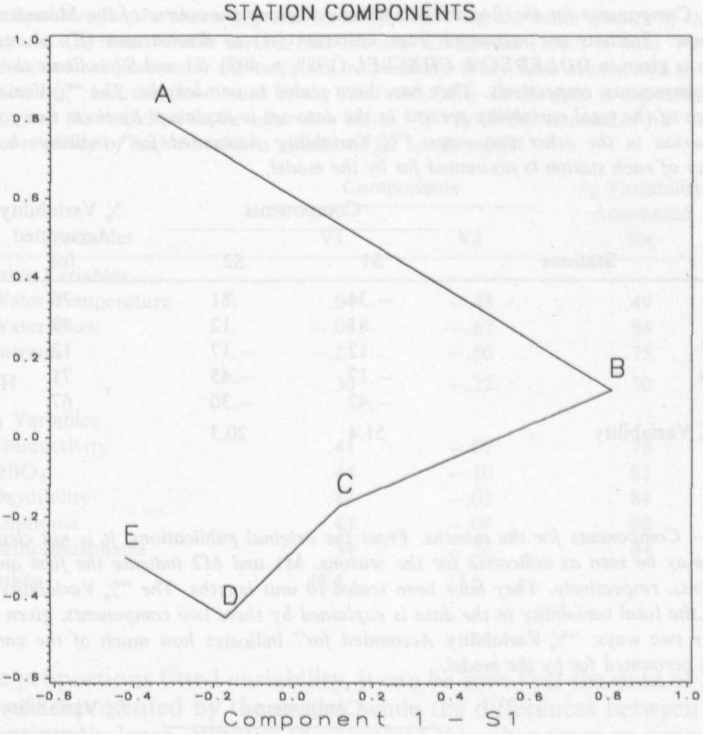

FIG. 1. — Graphical representation of the station components (see table II).

With respect to the standard deviations, there is considerable variation across months (see table I), and these differences can play a role in the analysis. We have performed analyses both with our preferred scaling, and with that of THIOULOUSE and CHESSEL. The two analyses lead to the same overall pattern, but to somewhat different conclusions about what are the important deviations from the major patterns. As mentioned above, we will concentrate on the per month scaled data.

Next we present the results of the three compromise solutions for each of three modes, i.e. stations, variables, and months (tables II, III, IV, and figures 1, 2, 3).

The solution with two components for each way accounts for 72% of the variability, which compares reasonably well with the partial triadic solution for the months of 89% by THIOULOUSE and CHESSEL. Of their 89% variability, 80% is associated with two components each of the other two modes. The variability accounted for in the present analysis is based on the simultaneous reduction in all three modes. Separate independent partial triadic analyses account for 84, 82 and 89% of the total variability, respectively. Incidentally these separate solutions are used as starting points for the iterative procedure presented here, and are thus available for comparison with the final simultaneous solution.

TABLE II. — *Components for the five stations measured along the course of the Méaudret, a tributary of the Bourne: Stations are measured from upstream (A) to downstream (E). A map of the exact locations is given in DOLEDEC & CHESSEL (1987, p. 407). S1 and S2 indicate the first and second station components, respectively. They have been scaled to unit lengths. The "% Variability" indicates how much of the total variability present in the data set is explained by these two components, given the reduction in the other two ways; "% Variability Accounted for" indicates how much of the variability of each station is accounted for by the model.*

	Components		% Variability
			Accounted
Stations	S1	S2	for
A	−.34	.81	76
B	.81	.12	89
C	.12	−.17	12
D	−.17	−.45	71
E	−.42	−.30	67
% Variability	51.4	20.3	

TABLE III. — *Components for the months. From the original publications, it is not clear whether these months may be seen as indicative for the seasons. M1 and M2 indicate the first and second month components, respectively. They have been scaled to unit lengths. The "% Variability" indicates how much of the total variability in the data is explained by these two components, given the reduction in the other two ways; "% Variability Accounted for" indicates how much of the variability of each month is accounted for by the model.*

	Components		% Variability
			Accounted
Months	M1	M2	for
June	.52	.26	75
August	.54	.60	84
November	.56	−.40	83
February	.38	−.64	45
% Variability	67.7	4.0	

The first thing to notice is that the months (fig. 3) are similarly arranged to THIOULOUSE and CHESSEL's fig. 2B on the first axis, but differently on the second one. The differences are due to the simultaneous analysis of the three ways, as only that part of the variability is accounted for that also can be explained by the other two ways at the same time, and this restriction is not operating in THIOULOUSE and CHESSEL's analysis. Note that, except for the month of February, the fit for the months separately is fairly equal.

There is a striking similarity between the compromise solutions of the stations (fig. 2) and that of the variables (fig. 3) with the patterns displayed in THIOULOUSE and CHESSEL's figure 4 constructed by simultaneously plotting the principal components for the stations and the variables from the analysis on their table III. This similarity between components is primarily due to the very strong first component of the months (see TEN BERGE *et al.*, 1987). Because of this similarity with previous analyses, we will not go into a detailed description of the patterns at this moment.

TABLE IV. — *Components for the chemo-physical variables describing the water quality of the Méaudret; for the exact units of measurement see DOLEDEC and CHESSEL* (1987, *p.* 465). *V1 and V2 are the first and second components of the variables, respectively. They have been scaled to unit lengths. The "% Variability" indicates how much of the total variability in the data is explained by the two components, given the reduction in the other two ways; "% Variability Accounted for" indicates how much of the variability of each variable is accounted for by the model.*

	Components		% Variability
			Accounted
Variables	V1	V2	for
Restaurative Variables			
1. Water Temperature	.04	−.48	49
2. Water Flow	−.08	−.62	84
8. Nitrates	−.23	−.50	75
3. pH	−.30	−.22	50
Pollution Variables			
4. Conductivity	.41	−.07	73
5. DBO_5	.44	−.10	85
6. Oxydibility	.44	−.03	84
7. Ammonia	.43	−.09	82
9. Orthophosphates	.34	−.26	64
% Variability	48.1	23.6	

From the proportions fitted variability, it can be seen that the data of Station C are not very well represented by the model, while the differences between the other stations are not overly large. What is the cause of this phenomenon cannot be said without further intimate knowledge of the subject matter. The proportions explaining variability of the variables differ with water temperature and pH at the lower end of the scale. Again the differences do not seem so large as to affect the interpretation seriously, but this is of course a subjective judgement.

As mentioned above, the method used here includes weights for the combination of components from three different modes. These weights are given in table V, and the eight numbers represent an extremely compact description of the 180 data points. Of these eight numbers, only three are sizeable, *i. e.* (S1, V1, M1), (S2, V2, M1), and (S1, V2, M2) account for 48, 20 and 4% of the total variability, respectively. The interpretation of these numbers has to be made by looking at the components to which the weights apply. Thus to interpret (S1, V1, M1), we have to investigate jointly the first components of each of the ways. For (S2, V2, M1), we have to investigate jointly the second components of the first two modes (stations and variables) and the first of the third mode (months), and similarly for (S1, V2, M2). This can be done by referring back to tables II, III, IV, and figures 1, 2, 3.

The (S1, V1, M1) combination indicates that station B [=S1] has all through the year [=M1] (very) high values on the (organic) Pollution variables [=V1], compared to the other stations (especially A, D, and E), which score below the average on the Pollution variables.

The (S2, V2, M1) combination indicates that all year round [=M1] the downstream stations D, E, (C) [=S2] have comparatively higher values with respect to

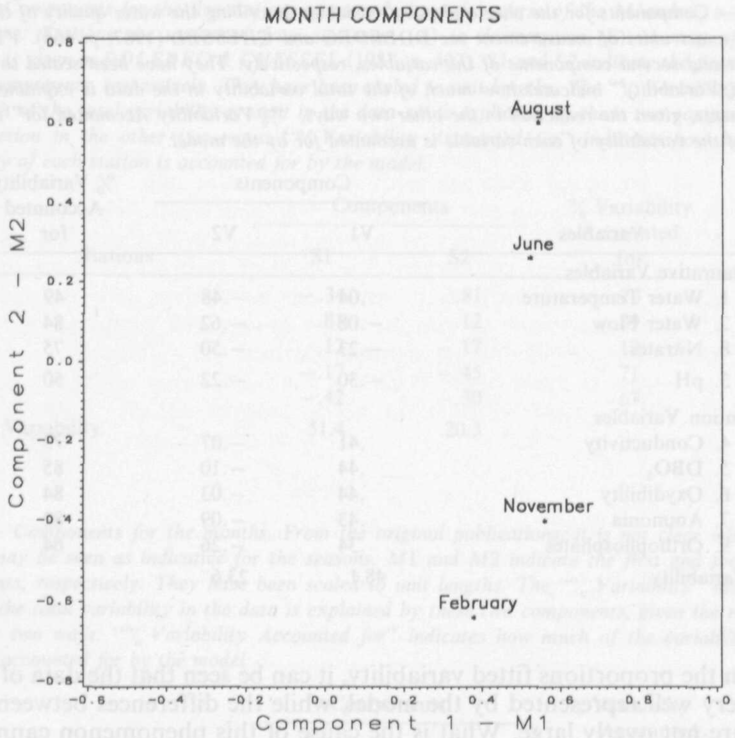

FIG. 2. — Graphical representation of the month components (see table III).

the mean on the Restaurative variables [=V2] compared to the upstream stations A, (B), thus there is a positive gradient from upstream to downstream with respect to temperature, water flow and the nitrate concentration. Note that station A has comparatively low values on both V1 and V2, so that the upper reach of the Méaudret is comparatively least polluted, coldest, and has the smallest water flow all year round.

The second component of the months accounts for only 6% of the variability, and shows primarily a contrast between the summer months and the winter months (see table II). Which factors are primarily responsible for this contrast? The answer is supplied by the (S1, V2, M2) combination of components. It indicates that B, (C) [=S1] has comparatively higher values on the Restaurative values [=V2] in summer than in winter [=M2], while E, A, (D) have comparatively higher values on these variables in winter than in summer. This trend should be seen as a detailing (or moderating) of the general trends described by the other two component combinations. In addition, this trend is less important than either of them, as follows from the smaller value (=2.57 or 4% variability accounted for).

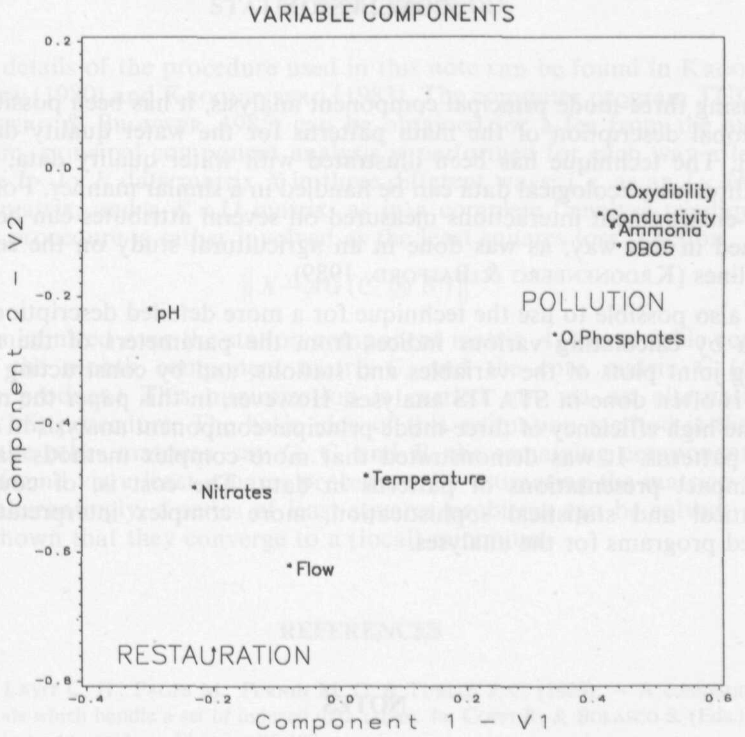

FIG. 3. – Graphical representation of the variable components (see table IV).

TABLE V. — *Core matrix (G) with weights for combinations of components from the stations (S_p), the Variables (V_q) and the Months (M_r). An element g_{pqr} or (S_p V_q M_r) indicates the weight of the combination of the p-th component of the first, the q-th component of the second, and the r-th component of the third way. The "% Variability" indicates how much of the total variability in the data is explained by a combination of components. The values are computed by squaring the g_{pqr} and dividing them by the total sum of squares.*

	Variable components		% Variability	
	V1	V2		
	Pollution	Restaurative		
Variables	Variables	Variables	V1	V2
M1 - Month Component 1: Common trends for all months				
S1 B, (C) vs. E, A, (D)	**9.27**	−.01	**.48**	.00
S2 A, (B) vs. D, E, (C)	.01	**5.99**	.00	**.20**
M2 - Month Component 2: Summer versus winter differences				
S1	.01	**2.57**	.00	**.04**
S2	−.82	−.01	.00	.00

CONCLUSION

By using three-mode principal component analysis, it has been possible to give a very global description of the main patterns for the water quality data of the Méaudret. The technique has been illustrated with water quality data, but other types of three-way ecological data can be handled in a similar manner. For instance, genotype-environment interactions measured on several attributes can be fruitfully approached in this way, as was done in an agricultural study on the selection of soybean lines (Kroonenberg & Basford, 1989).

It is also possible to use the technique for a more detailed description of three-way data by calculating various indices from the parameters of the model, by producing joint plots of the variables and stations, and by constructing trajectory plots, as is often done in STATIS analyses. However, in this paper the main focus was on the high efficiency of three-mode principal component analysis in describing complex patterns. It was demonstrated that more complex methods can lead to more compact presentations of patterns in data. The cost is, of course, more mathematical and statistical sophistication, more complex interpretations, and specialized programs for the analyses.

NOTES

([1]). The technique used by Thioulouse and Chessel is called "partielle", because a full "analyse triadique" would consist of similar analyses on other arrangements of the data (see Jaffrenou, 1978, chap. IV rather than chap. III). Incidentally, their approach has a precursor in Tucker & Messick (1963) who followed an essentially similar procedure for similarity data. In fact, the full analysis is the subject of this paper.

([2]). Thioulouse and Chessel refer to such component matrices as the "interstructure", however within the STATIS framework, that term is generally reserved for the components of a matrix C which has, apart from scaling constants, elements $c_{kl} = \mathrm{Tr}\, S_k S_l$, with $S_t = X_t X_t'$, and X_t is a slice of the three-way data matrix X. Thus in STATIS, the interstructure is based on covariances between covariance matrices, or on the data to the fourth power, which is quite different from the components referred to in Thioulouse and Chessel's paper. However, both Glacon (1981) and Carlier *et al.* (1989) indicate that one may also use the X_t directly in the trace operators to define C, which is what Jaffrenou, and Thioulouse and Chessel do, and which allows the latter to find their interstructure directly using ordinary principal components. Unfortunately, the distinction between the interstructure and the compromise solution becomes a bit blurred, especially in the kind of three-way models we are discussing here, in which there are no triple-subscripted arrays of, say, stations by variables by month components, such as displayed in Thioulouse and Chessel's table III.

STATISTICAL APPENDIX

Most details of the procedure used in this note can be found in KROONENBERG & DE LEEUW (1980) and KROONENBERG (1983). The computer program TUCKALS3 (KROONENBERG & BROUWER, 1985) can be obtained for a fee from the author. In the program, principal component analysis is performed for each way after stringing-out the $I \times J \times K$ data matrix X in three different ways, *i. e.* as an $I \times JK$ matrix, an $J \times KI$ matrix, and a $K \times IJ$ matrix, as in a complete "analyse triadique". The estimation procedure is rather involved as the least squares loss function

$$\| X - AG(C' \otimes B') \|^2$$

has to be minimized over the station component matrix A, the variable component matrix B, the month component matrix C, and the core matrix G ($\otimes$ is the Kronecker product.) This minimization is carried out via an alternating least squares (ALS) procedure. The basic idea of this estimation method is that conditional on the other matrices, say G, C, and B, the remaining component matrix, A, can be found via a least squares procedure. By estimating the matrices one at a time and alternatingly, a series of least squares problems can be solved, of which it can be shown that they converge to a (local) minimum.

REFERENCES

CARLIER A., LAVIT C. H., PAGES M., PERNIN M. O. & TURLOT J. C. (1989). — A comparative review of methods which handle a set of indexed data tables. *In:* COPPI R. & BOLASCO S. (Eds.), *Multiway data analysis*. Amsterdam, Elsevier, 85-102.

DOLÉDEC S. & CHESSEL D., 1987. — Rythmes saisonniers et composantes stationnelles en milieu aquatique. I. Description d'un plan d'observation complet par projection de variables. *Acta Oecologica. Oecologia Generalis*, **8**, n° 3, 403-406.

GLACON F., 1981. — Analyse conjointe de plusieurs matrices de données. Comparaison de différentes méthodes. *Thèse de 3^e^ cycle*, Université Scientifique et Médicale de Grenoble.

JAFFRENOU P. A., 1978. — Sur l'analyse des familles finies de variables vectorielles. Bases algébriques et application à la description statistique. *Thèse de 3^e^ cycle*, Université des Sciences et Technique du Languedoc, Montpellier-II.

KIERS H. (1988). — Comparison of "Anglo-Saxon" and "French" three-mode methods. *Statistique et Analyse des Données*, **13**, 14-32.

KROONENBERG P. M., 1983. — *Three-mode principal component analysis: Theory and applications*. D.S.W.O. Press, Leiden, Pays-Bas, 399 p.

KROONENBERG P. M., 1987. — Multivariate and longitudinal data on growing children. Solutions using a three-mode principal component analysis and some comparison results with other approaches. *In:* JANSSEN J., MARCOTORCHINO F., & PROTH J. M. (Eds.), *Data analysis. The ins and outs of solving real problems*, New York: Plenum, 89-112.

KROONENBERG P. M. & BASFORD K. E. (1989). — An investigation of multi-attribute genotype response across environments using three-mode principal component analysis. *Euphytica*, **44**, 109-123.

KROONENBERG P. M. & BROUWER P., 1985. — *User's guide to TUCKALS3. Version* 4.0. Department of Education, Leiden University, Leiden, Pays-Bas.

KROONENBERG P. M. & DE LEEUW J., 1980. — Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, **45**, 69-97.

LAVIT C. H., 1988. — *Analyse conjointe de tableaux quantitatifs*. Paris, Masson.

L'HERMIER DES PLANTES H., 1976. — Structuration des tableaux à trois indices de la statistique. *Thèse de 3^e^ cycle*, Université des Sciences et Technique de Languedoc, Montpellier-II.

TEN BERGE J. M. F., DE LEEUW J. & KROONENBERG P. M., 1987. – Some new results on principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, **52**, 183-191.

THIOULOUSE J. & CHESSEL D., 1987. – Les analyses multitableaux en écologie factorielle. I. De la typologie d'état à la typologie de fonctionnement par l'analyse triadique. *Acta Oecologica. Oecologia Generalis*, **8**, n° 4, 463-480.

TUCKER L. R., 1966. – Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279-311.

TUCKER L. R. & MESSICK S., 1963. – An individual differences model for multidimensional scaling. *Psychometrika*, **28**, 333-367.