# HOMOGENEITY ANALYSIS OF EVENT HISTORY DATA[*]

Jan DE LEEUW
Peter VAN DER HEIJDEN
Ita KREFT
Department of Data Theory
Leiden University

ABSTRACT

The technique of homogeneity analysis (also called multiple correspondence analysis) is applied to event history data obtained in the National Travel Survey of the Netherlands Central Bureau of Statistics. We introduce the theory of homogeneity analysis by using the idea of optimal quantification or transformation. Optimality is defined in terms of a ratio of quadratic forms, where both quadratic forms are components of the total variance after quantification. This defines a large class of optimal quantification techniques, with the size of the class depending on the number of independent sources of variation in the study design.

## 1 INTRODUCTION

In this paper we discuss the analysis of _event history data_, _panel data_, or _pooled cross-sectional and time series data_. Lazarsfeld (1978) reviews early work on panel data modelling in sociology. Wansbeek (1980) and Dielman (1983) discuss panel data modelling in econometrics. Also compare special issues of the Annales de l'INSEE (30/31, 1978) and of the Annals of Applied Econometrics (1983-1).

Early panel data had a rather special form, because the individuals in the sample were observed at only a very small number of time points

[*] We thank Jan C. Mulder (Institute of Town Planning Research, Technical University, Delft) for constructing our data set from the National Travel Survey base.

or waves, usually two. Recent advances in data collection methods and in statistical modelling have made it possible to study behaviour followed in continuous time. These continuous event history data are also becoming more popular in spatial analysis and other areas of human geography (Carlstein, Parks, and Thrift, 1978). They are often collected in the form of diaries. The example we analyze in this paper will be such time-activity data, with measurements in continuous time. Data were sampled from the National Travel Survey 1980, collected by the Netherlands Central Bureau of Statistics (Moning, 1983). More details about the const^uction of the sample are in Kreft and Mulder (1984).

If there are only a few time points, panel data in sociology have usually been modelled by log-linear or logit models. Econometrics uses regression methods, of course, but econometricians usually assume that they deal with continuous variables. We are mainly interested in discrete or nominal variables, for which regression methods are not appropriate. If there is a large number of discrete time-points, it becomes necessary to take the dynamic structure of the process into account. Sociologists have been using Markov processes for this purpose (Anderson, 1979). In the case of continuous time event histories Coleman, Singer, Spiegelman, Heckman, Hannan, Tuma and others consequently turned to continuous time Markov processes (cf. Tuma and Hannan, 1984, for a review). Although Markov processes are a very natural choice of models from a theoretical point of view, they do not necessarily provide a good starting point for data analysis. The maximum likelihood methods of Tuma and others can be used only if the number of states is relatively small, or if we impose strong stationarity assumptions. In other cases we will encounter the empty-cell problem, that is already familiar from (cross-sectional) analysis of categorical data. If there is not enough prior knowledge to impose strong models, then we are more or less forced to use exploratory techniques. In this paper we illustrate the use of homogeneity analysis (also known as multiple correspondence analysis and as qualitative harmonic analysis) on the National Travel Survey data.

## 2 THREE-WAY INDICATORS

Suppose I is a set of individuals, T is a set of time-points, and S is a set of states. At time t each individual i is in one of the states s. We code our data by using a three-way indicator function g. For data analysis purposes we can suppose, of course, without loss of generality that the sets I and T and S are finite. The indicator function becomes a three-way indicator matrix G, of dimension card (I) x card (T) x card (S). Observe that $g_{it+} = 1$ for all i,t (replacing an index by + means summing over the index). This shows that the index set S plays a role which is somewhat special. Clearly our indicator notation is very general. It can be used irrespective of the number of time points ('continuous time' merely means a very large number of time points), and for any number of states (more particularly the state space can be continuous and/or multivariate).

In the National Travel Survey we have used the sample of 940 individuals (470 husband-wife pairs), who were followed during one whole day. The activities (states) were coded in five categories: (a) work, including school, (b) being at home, (c) shopping, (d) travelling, (e) other, including visits, sports, culture. For each minute, during one whole day, we know which of the activities each of our individuals was engaged in. Thus T has 24 x 60 = 1440 elements, and the three-way matrix G is 940 x 1440 x 5. We know that $g_{it+} = 1$ by definition, but it is of some interest to study the other marginals. Marginal $g_{i+s}$ is a 940 x 5 matrix, which shows for each individual how many minutes (s)he spend on each of the five activities. Marginal $g_{+ts}$ is 1440 x 5, it shows for each minute how many individuals were engaged in each of the five activities. The last set of marginals defines the time-budgets of the sample, the columns of $g_{+ts}$ can be plotted as five different functions of time (adding up to 940 for each minute).

In the usual time-budget studies using diaries these two nontrivial margins are studied in detail. They are investigated for structure by various techniques, and they are related to various exogeneous variables. Using only the marginals has the disadvantage, however, that interactions between individuals and time are eliminated. Homogeneity analysis tries to analyze these interactions, and consequently works directly on the body of the three-way indicator matrix G.

## 3 QUANTIFICATION

Suppose that we replace each $g_{its}$ which is nonzero by a real number $y_{its}$. We can write this as $x_{its} = g_{its}y_{its}$, and we call X the <u>quantified three-way indicator matrix</u>. The idea behind homogeneity analysis and related techniques is to decompose the variation in X as in the analysis of variance. Or, more precisely, using analysis of variance terminology and notation is one way to introduce these techniques. This was pioneered by Guttman (1941), and systematically exploited in a somewhat different context by Abelson (1960).

The sum of squares of the $x_{its}$ (which is a function of the $y_{its}$) can be decomposed in eight different components: the mean; the main effects for individuals, time-points, and states; the three two-factor interactions I × T, I × S, and T × S; and the three-factor interaction I × T × S. We assume that the reader is familiar with this decomposition of the sum of squares of the $x_{its}$ in eight component sums of squares.

The next fundamental idea is that of <u>optimal scaling</u>. This means that we choose the quantifications $y_{its}$ by maximizing a criterion. We define the criterion* by selecting a subset of the eight component sums of squares. The criterion is the sum of the selected sums of squares. It is, of course, not interesting to look for the unrestricted maximum of the criterion. Because is it unbounded, we need some form of <u>normalization</u>. For the normalization we select another subset (usually containing the first one), and we require that the sum of the sums of squares in this second subset is equal to one. This makes the sets of feasible quantifications bounded, and the optimalization problem becomes well defined. Compare De Leeuw (1982) for precise conditions.

A final component defining a technique in this class are the <u>constraints</u> on the quantifications (in addition to the normalization constraints). They are of the form $y_{its} = y_{ts}$ or $y_{its} = w_{is}z_{ts}$ or $y_{its} = a_{is} + b_{ts}$, and so on. The reason for these constraints is clear. Unrestricted quantification is far too general, in fact $x_{its} = g_{its}y_{its}$ implies that the $y_{its}$ corresponding with $g_{its} = 0$ do not influence criterion and normalization, and are completely arbitrary. If summing over an index in a product is indicated by a bar, then 'interaction' $x_{its} = g_{it\underline{s}}y_{it\underline{s}}$ can be made equal to any arbitrary matrix by a suitable choice of the $y_{its}$. Unrestricted quantification gives too much freedom.

A first example from this class is taken from Fisher (1938, section 39.2). For the criterion we use components I and T, for the normalization I, T and I x T. The constraint is $y_{its} = y_s$. Because neither criterion nor normalization involve S, the technique can be interpreted in terms of $x_{it} = g_{it\underline{s}}y_{it\underline{s}} = g_{it\underline{s}}y_{\underline{s}}$. Replace the state labels by real numbers in such a way that the sum of the I and T main effects is maximized relative to the total variance.

We now define homogeneity analysis (Gifi, 1981), multiple correspondence analysis (Cazes a.o., 1977, Lebart a.o., 1977), or qualitative harmonic analysis (Deville and Saporta, 1980, 1983; Saporta, 1981) in this framework. We use the criterion consisting of I, the normalization consisting of the sum of I, T, and I x T, and the constraints $y_{its} = y_{ts}$. Now $\underline{x}_{it} = g_{it\underline{s}}y_{it\underline{s}} = g_{it\underline{s}}y_{t\underline{s}}$, and we maximize the sum of squares between individuals relative to the total variance of the $x_{it}$. For the interpretation we must keep in mind that maximizing the variance between individuals amounts to the same thing as minimizing the variance between time-points within individuals. Thus if we plot the $x_{it}$ as a function of time, one curve for each individual, then homogeneity analysis transforms the states in such a way that these curves become as similar as possible to horizontal lines. The reason for doing this is that if the curve is a horizontal line, then we can characterize the individual by a single number, the height of her/his line.

Thus homogeneity analysis defines curves to be 'satisfactory' if they are close to being constant as a function of time. Many other definitions of being satisfactory are possible, of which low degree polynomiality is perhaps the most natural alternative (De Leeuw, 1972, 1984a). Another related idea is to perform the analysis of variance decompositions not directly on the curves $x_{it}$, but on smoothed versions of these curves. The theory of interpolating splines can be used here. This amounts to the same thing as introducing a weighted metric on the space of curves (Besse, 1979; Ramsay, 1982; De Leeuw, 1984a). We shall not study these additional possibilities in this paper, we restrict ourselves to simple variations of ordinary homogeneity analysis. Of course we can define additional solutions for the quantifications by maximizing the same criterion, with the restriction of orthogononality to all previous solutions added to the normalization requirements.

Orthogonality is defined in the metric used for the normalization. Proceeding in this way we can find various orthogonal dimensions, indeed we can proceed until we have an orthogonal basis for the space of all feasible quantifications.

## 4 ANALYSIS I

The first analysis of our example is standard homogeneity analysis, as explained above. We give the matrix formulation of the problem first. Suppose $G_t$ is the card (I) x card (S) indicator matrix for time t. In the example the $G_t$ are of dimension 940 x 5. Homogeneity analysis is equivalent to a correspondence analysis on the card (I) x (card (S) x card (T)) supermatrix, with the $G_t$ next to each other as submatrices. Standard references for correspondence analysis are Benzécri a.o. (1973, 1980), Nishisato (1980), Gifi (1981), Greenacre (1984). In our example the supermatrix G is of dimension 940 x 7200, which is far too large for most correspondence analysis programs. And even if we could analyze it, by special tricks, the solutions for the quantifications $y_{its}$ would presumably be very unstable (Gifi, 1981, De Leeuw, 1984b). Even for homogeneity analysis there is an empty cell problem in this case.

The solution of this dilemma is, of course, that we impose constraints on the quantifications. In our case we have required that the $y_{ts}$ corresponding with minutes in the same hour must be equal. In fact the actual restrictions were a bit more complicated, because during typical home-work-travel periods we used half-hours, and we did not start recording behaviour until 6.00 a.m.. The details are in Kreft and Mulder (1984). We merely indicate here, that the minutes were divided into 22 time-periods. Quantifications for minutes in the same time-period must be equal. Homogeneity analysis with these additional restrictions amounts to correspondence analysis on a 940 x 110 supermatrix, consisting of 22 submatrices of dimension 940 x 5. Each submatrix corresponds with a time-period, and is formed by adding the $G_t$ for all minutes in the same period. Similar restrictions were used in this context by Deville and Saporta (1980).

The first four singular values (canonical correlations) from the correspondence analysis were 0.6660, 0.5464, 0.4568, 0.4491. We only use the first two in our further remarks. This decision is made because

(a) the 'elbow'-criterion indicates that the remaining singular values are approximately equal, (b) the remaining dimensions are not really 'common' factors but contrast either one time period or one activity with the rest, (c) three-dimensional plots are less attractive. Figure 1 shows the projections of the 940 individuals on the first two singular vectors. The very large cluster in the top right-hand section are the individuals who are mainly at home. The second large cluster, top left, are people who work full-time. Thus dimension one, the horizontal dimension, contrasts 'work' with 'being-at-home'. The second, vertical, dimension contrasts 'working' and 'being-at-home', at the top, with other activities outside the house, at the bottom. These alternative activities are mainly in the category 'other', but also in 'shopping' and 'traveling'. This interpretation of the dimensions becomes beautifully clear if we plot the 5 x 22 = 110 category quanti-

Figure 1: analysis 1, 940 object scores in two dimensions

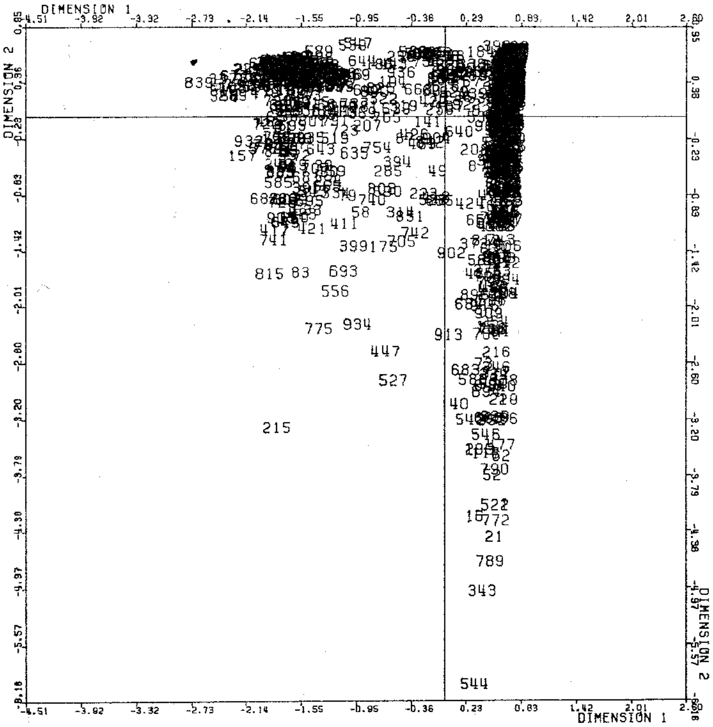
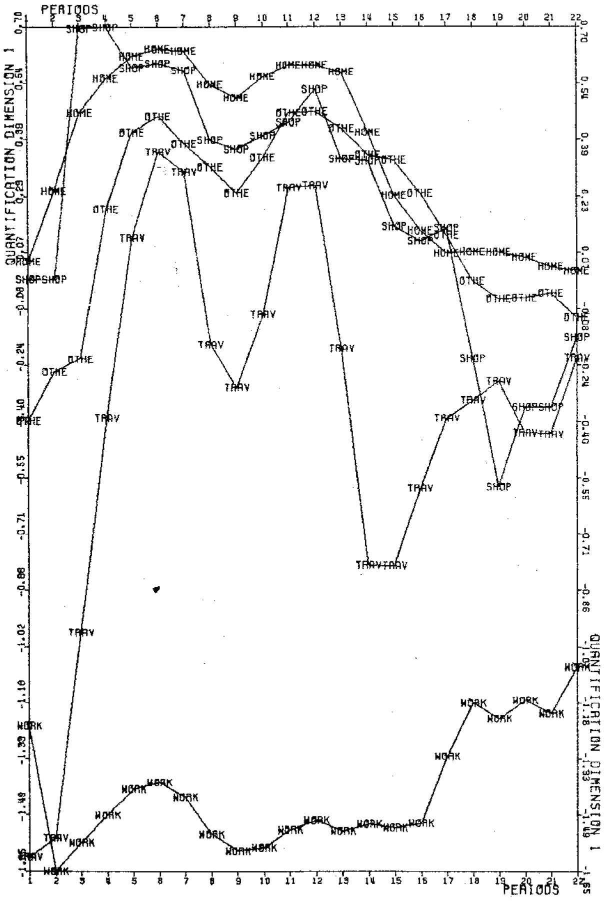
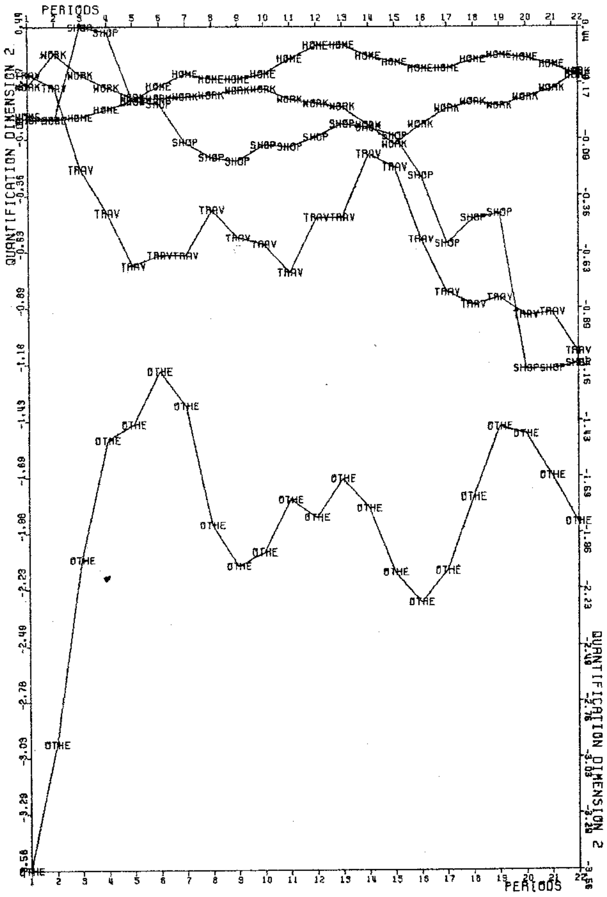


Figure 2a: analysis 1, category quantifications,
dimension one against time



fications. In this analysis we have not plotted them in two dimensions, but we have plotted each dimension separately against time-period. Figure 2a is the plot for the first dimension. We have seen that these category quantifications make maximum discrimination of the individuals possible. They are related in a simple way to figure 1: the score for 'travel' in period 6 is the average score of all individuals who travel in period 6 (on dimension 1, and weighted with the number of minutes they travel). Individuals who work a lot are low on the dimension, individuals who are at home are high. During working hours the

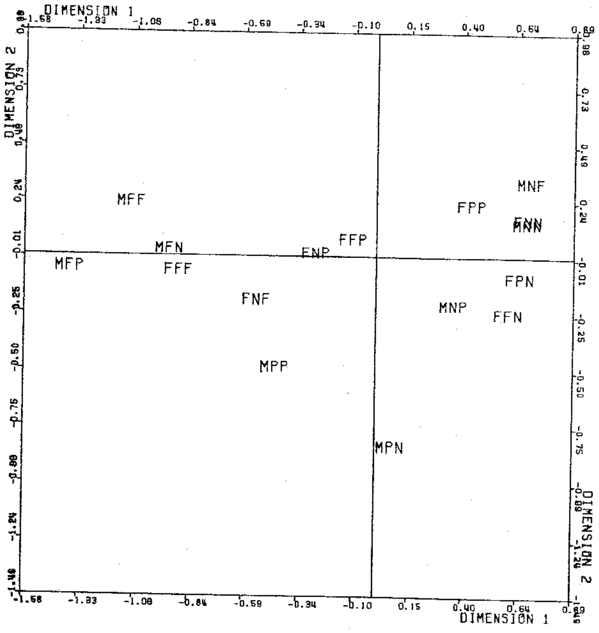
Figure 2b: analysis 1, category quantifications, dimension two against time



average for travelling persons is close to the average for persons at home, during the lunch break. During the early morning and late evening hours it is much closer to the average of the persons who work. The same thing is true for shopping, although working people do not shop a great deal during lunch time, and do not even shop much during late afternoon. Shopping is done by those who stay at home. Figure 2b shows a similar plot for the second dimension. The social and cultural activities are concentrated in the morning, in the early afternoon, and in the early evening. The morning and early afternoon shopping behaves like 'other', in the evening it is quite different. Travelling behaves in the opposite way. If you are going to visit somebody or something, then you have to travel before and after this visit. Thus travelling hours are just before and just after visiting hours. On the first dimension we can best discriminate people during working hours, on the second dimension we can best discriminate them outside working hours. More detailed interpretations of these plots are in Kreft and Mulder (1984).

Another way to interpret the dimensions of the homogeneity analysis is by using passive variables (or supplementary variables). They play no role in the analysis, only in the interpretation. They are used afterwards to label the plots, and to compute centroids of groups of individuals. In our analysis we have used two passive variables: sex combined with work-situation of the family. Work situation has nine possible values: the head of the family can be employed full-time, part-time, or no-time, and the same thing is true for his/her partner. If we combine this with sex we have a new interactive passive variable

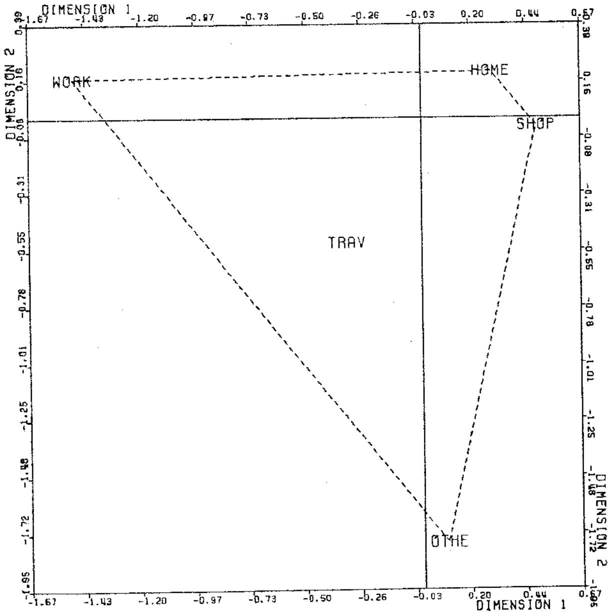
Figure 3: analysis 1, centroids for passive sex x work variable



with 18 categories, of which woman-in-a-family-in-which-head-works-fulltime-and-the-partner-does-not-work is a typical one. These 18 categories can be used to compute 18 centroids from the projections in figure 1. We label them by a three-letter code indicating sex (M/F), working status of head (F,P,N), and of partner (F,P,N). Thus on the left in figure 3 we find MFP, the average position in figure 1 of males from families in which the head works full-time and the partner part-time. If we compare MFF with FFF for instance, we see that females from families in which both partners work full time are more at home on the average, and spend more time on social, cultural, recreational activities. Comparing MFP with FFP shows that in almost all cases the male is the head of the family. There are no PF-families. MFP work more than MFF. Many other interesting details can be found in these plots, and other additional passive variables can be tried. Compare Kreft and Mulder (1984).

It may be of some interest to show what happens if we restrict $y_{its}$ by $y_{its} = y_s$. Thus category quantifications must be the same for all

Figure 4: analysis,1, category quantifications 940 x 5 table



minutes. We have to perform a correspondence analysis on a 940 x 5 table. The singular values are 0.6197, 0.5285, 0.3318, 0.2040. Observe that the first two are quite close to those of our previous analysis. Figure 4, in which the category quantifications from this analysis are plotted in two dimensions, shows that the interpretation of the dimensions is still the same. They are stable, even under these very severe restrictions.

The most remarkable finding from analysis I is, perhaps, that the five different curves in figures 2a and 2b are roughly proportional to each other. In figure 2a all five curves have the same hills and valleys. In figure 2b 'other' is opposed to 'travel' and 'shopping', while 'at home' and 'work' are more or less neutral. A simple explanation for proportionality is given by De Leeuw (1984a). If the diaries are a random sample from a first-order stationary Markov chain, then the optimal $y_{its}$ have, approximately, the form $y_{its} = w_s z_t$. Deviations from this form can be due to sampling errors, and to deviations from Markovity. The proportionality is maintained under various forms of nonstationarity. Thus analysis I suggests that perhaps Markov models can be used quite effectively here, although it is clear that a more detailed analysis shows various systematic deviations from proportionality (in the evenings and at night, for example) and thus from Markovity.

## 5 ANALYSIS II

In a transposed homogeneity analysis of the same data we interchange the role of I and S. Thus we require $y_{its} = y_{it}$ and we look at the induced quantifications $x_{ts} = g_{its}y_{its} = g_{its}y_{it}$. We maximize variance due to S, keeping S plus T plus S x T equal to a constant. Thus we minimize variance between time-points, within states. We have to perform a correspondence analysis on a matrix of 5 x (22 x 940) = 5 x 20680. Again minutes have been grouped into periods, otherwise the analysis would have been on a 5 x (1440 x 940) = 5 x 1353600 matrix. Results of the analysis are a bit disappointing. Singular values are very close to one. This is basically because the matrices $G_t$ are close to being orthogona. for most minutes, and thus the matrix fed into correspondence analysis here, which consists of all $G_t$ on top of each other, is also close to orthogonal. All four dimensions are largely

Figure 5a: analysis 2, 2J680 x 5 table, average object scores for men and women, dimension one against time

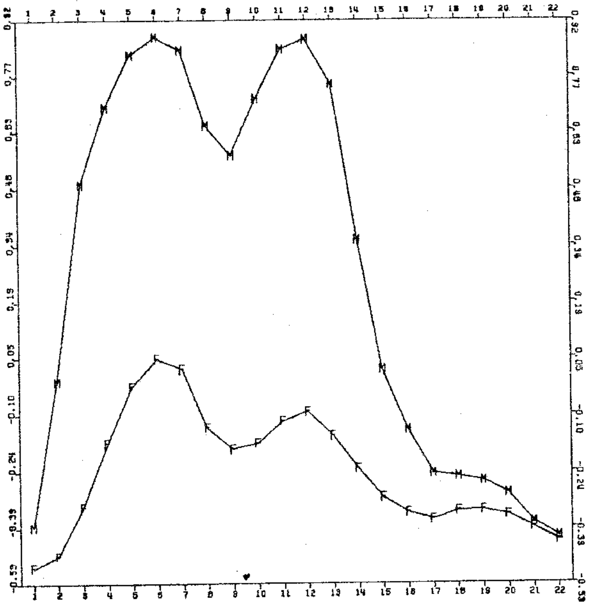
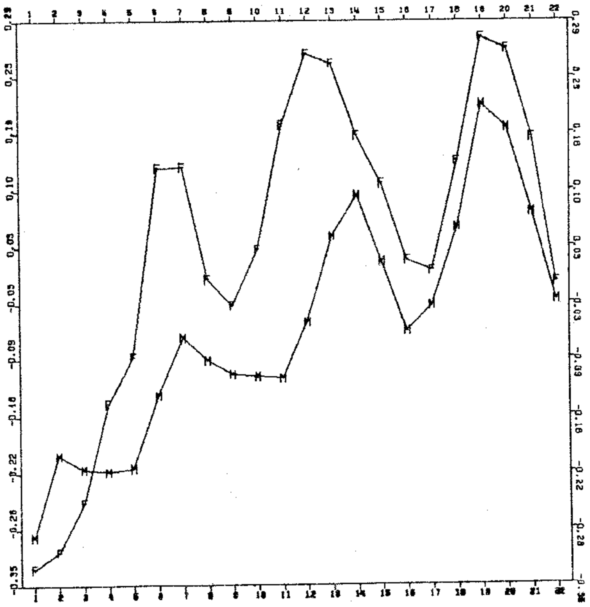


Figure 5b: analysis 2, 2068 x 5 table, average object scores for men and women, dimension two against time

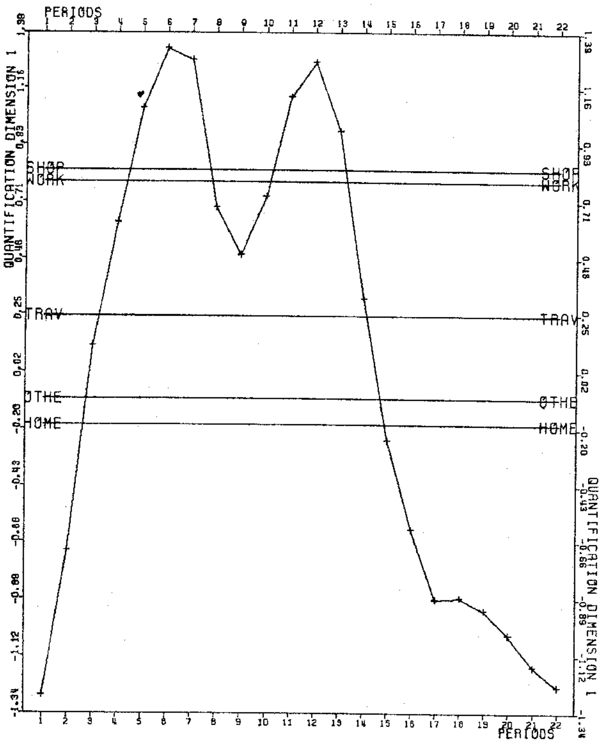


unique, although dimension 1 still contrasts 'work' with 'being at home', and dimension 2 contrasts 'work' and 'being at home' on one side with the other activities on the other side.

Of course we do not present plots of all 20680 values of $y_{it}$ for the first two dimensions. What we have done is compute the average $y_{it}$ value for men and the average $y_{it}$ value for women, on both dimensions. This is plotted, against time, in figures 5a and 5b. We see men working more than women, women working more in the morning, partners being at home together during the night, in figure 5a. We also see, in 5b, that women perform more social, recreational, shopping, and travel activities, especially during the day. The three peaks are visiting, traveling, and shopping times. Only in the morning men outtravel women.

Repeating the same analysis with $y_{its} = y_t$ means performing corres-

Figure 6: analysis 2, category and period quantifications, 22 x 5 table



pondence analysis on a 5 x 22 matrix. Its singular values are 0.3610, .01521, 0.0788, and 0.0509. Figure 6 plots the $y_t$ against time, and shows the five column averages of $x_{ts}$ as horizontal straight lines. We have only done this for the first dimension, with being at home and the other activities on one side with shopping and working on the other side (with travel nicely in between). For the interpretation remember that the score for shopping, for instance, is the weighted average of the period quantifications $y_t$, with weights equal to the number of minutes spent on shopping during the period.

## 6 OTHER ANALYSES

We have already seen in both analysis above, that it is very useful for interpretation purposes to work with passive or supplementary variables. We have already used the fact that our sample consists of 470 man-woman pairs for this purpose. But in many cases the background information can be incorporated in the analysis in a more active way. We mention some examples.

Because our sample consists of paired observations it is perhaps more natural to use the family as a unit of analysis. This means that the data should be coded as two paired three-way indicators, or as one single four-way indicator matrix. The four ways are F(amily), G(ender), T(ime), and S(tate). We have elements $g_{fgts}$, with $g_{fgt+} = 1$ and with $g_{f1++} = g_{f2++}$. By using quantifications we can define 16 components of the total sum of squares, and many corresponding forms of optimal scaling techniques. We can maximize the differences between families, and so on. There are a great many possibilities, and a choice from them will depend on both the particular question we are investigating and the prior knowledge we have about the data and the process that generates them. Of course the number of possibilities increases even more, if we allow for additional active background information such as age, number of children, employment status, and so on.

We shall not present any of these additional analyses, because the basic idea is probably clear. A class of techniques have been introduced which is based on (a) coding the data as a multiway indicator matrix, (b) quantifying the categories, (c) imposing simple restrictions on the category quantifications, (d) maximizing a component of the sum of squares relative to another component. We have seen that Fisher's qualitative ANOVA is one example, qualitative harmonic analysis of Deville and Saporta is another one. Our general approach makes it also possible to introduce 'transposed' analyses and to consider many other possibilities. Interpretability, prior theory, gauges such as the Markov processes, analysis of stability, precise nature of the questions asked are all possible criteria that can be used in the choice of the technique from this class (compare Gifi, 1981; De Leeuw, 1984c).

## REFERENCES

Abelson, R.P. (1960). Scales derived by consideration of variance components in multiway tables. In H. Gulliksen and S.J. Messick (eds.), Psychological Scaling: theory and applications. New York: Wiley.

Anderson, T.W. (1979). Panels and time-series analysis: Markov Chains and autoregressive processes. In R.K. Merton, J.S. Coleman, & P.H. Rossie (eds.), Qualitative and quantitative social research. Papers in honour of Paul F. Lazarsfeld. New York: The Free Press.

Benzécri, J.P. e.a. (1973). L'Analyse des Données. Paris: Dunod. (2 vols).

Benzécri, J.P. e.a. (1980). Practique de l'analyse des données. Paris: Dunod (3 vols).

Besse, P. (1979). Etude descriptive d'une processus. Approximation et Interpolation, Thèse. Université Paul Sabatier de Toulouse.

Carlstein, T., Parks, D. & Thrift, N. (eds.) (1978). Timing space and spacing time. London: Edward Arnold (3 vols).

Cazes, P., Baumerder, A., Bonnefous, S., Pagès, J.P. (1977). Codage et analyse des tableaux logiques. Introduction a la practique des variables qualitatives. Cahiers de BURO, no 27. Paris: Université Pierre et Marie Curie.

De Leeuw, J. (1972). Canonical analysis of multiple time series. University of Leiden: Department of Data Theory.

De Leeuw, J. (1982). Generalized eigenvalue problems with positive semidefinite matrices, Psychometrika, 47, 87-93.

De Leeuw, J. (1984a). Homogeneity analysis of curves and processes. Paper presented at the Table Ronde 'Analyse des Données'. Toulouse: january 9-10.

De Leeuw, J. (1984b). Statistical properties of multiple correspondence analysis. Paper presented at the conference 'New Multivariate methods in Statistics'. Bowdoin College, Maine, june 10-16.

De Leeuw, J. (1984c). Models of data. Kwantitatieve Methoden, 5, 17-30.

Deville, J.C., & Saporta, G. (1980). Analyse harmonique qualitative. In E. Diday (ed.) Data analysis and informatics, Amsterdam: North Holland Publishing Co.

Deville, J.C., & Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. Journal of Econometrics, 22, 169-190.

Dielman, T.E. (1983). Pooled cross sectional and time series data: a survey of current statistical methodology. *The American Statistician*, *37*, 111-122.

Fisher, R.A. (1938). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Gifi, A. (1981). *Nonlinear multivariate analysis*. University of Leiden: Department of Data Theory. New edition: Leiden: DSWO Press, 1984.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. New York: Academic Press.

Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In P. Horst (ed.). *The prediction of personal adjustment*. New York: SSRC.

Kreft, I., & Mulder, J.C. (1984). De analyse van verplaatsingsgedrag met behulp van correspondentieanalyse (in Dutch, with English summary). University of Leiden: Department of Data Theory.

Lazarsfeld, P.F. (1978). Some episodes in the history of panel analysis. In D.B. Kandel (ed.). *Longitudinal research on drug use*. New York: Wiley.

Lebart, L., Morineau, A., Tabard, N. (1977). *Techniques de la Description Statistique*. Paris: Dunod.

Moning, H. (1983). *The National Travel Survey in The Netherlands*. Heerlen: Central Bureau of Statistics.

Nishisato, S. (1980). *Analysis of Categorical data: Dual Scaling and its applications*. Toronto: University of Toronto Press.

Ramsay, J.O. (1982). When the data are functions. *Psychometrika*, *47*, 379-396.

Saporta, G. (1981). *Methodes exploratoires d'analyse de données temporelles*, Thèse. Paris: Université Pierre et Marie Curie.

Tuma, N.B., & Hannan, M.T. (1984). *Social dynamics: models and methods*. New York: Academic Press.

Wansbeek, T.J. (1980). *Quantitative effects in panel data modelling*. Dissertation. Leiden University.

Jan DE LEEUW, Peter VAN DER HEIJDEN, Ita KREFT,
Department of Data Theory FSW/RUL,
Middelstegracht 4, 2312 TW Leiden, The Netherlands.