JOURNAL OF CHEMOMETRICS
J. Chemometrics 2000; **14**: 241–259

Determination of underlying components of a cyclical time series by means of two-way and three-way factor analytic techniques

Pentti Paatero[1]*† and Sirkka Juntto[2]

[1]*Department of Physics, University of Helsinki, Box 9, FIN-00014 Helsinki/University, Finland*
[2]*Finnish Meteorological Institute, Sahaajankatu 20 E, FIN-00810 Helsinki, Finland*

SUMMARY

A technique is presented for determining the underlying components in a cyclical time series which is influenced by one prominent cycle (the diurnal or the yearly cycle). The separation of the components is based on their different shapes within this period, assuming that the shape of each component stays approximately constant with time and that the amplitude of each component is a slowly changing function. The series is folded into matrix shape so that each cycle forms one column. The matrix is factorized by principal component analysis or by positive matrix factorization (non-negatively constrained factor analysis with individual weighting of data values), resulting in the shape and amplitude functions for the underlying components. Synthetic two-way demonstration examples are analyzed. As a real-life example, traffic-induced carbon monoxide concentrations in urban air are analyzed. The CO has a diurnal concentration cycle which changes shape on weekends. This behavior is explained by two factors, identified with work-related and other traffic. The CO data in fact contain another multiplicative cycle, the weekly workdays/weekend pattern. Arranging the data according to time of day, day of week, and week of the year creates a three-way array. The method is extended to the analysis of such arrays. Existing software for the well-known PARAFAC model is used for solving the three-way model. Two factors are again obtained. Their diurnal and weekly cycles correspond to the work-related and weekend-related traffic patterns. Analysis of cyclical multivariate data is discussed: such data are also governed by the three-way PARAFAC model. The advantage of the PARAFAC model relative to customary two-way methods is emphasized: there is usually no rotational ambiguity in PARAFAC results. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: factor analysis; principal component analysis; positive matrix factorization; cyclical time series; environmental time series; carbon monoxide

INTRODUCTION

In this work the term ‘time series’ is used in its basic meaning, denoting that some quantity has been observed repeatedly over time. An analysis is made of cyclical time series, i.e. of series in which similar values occur repeatedly in a fixed, previously known cycle. The amplitude of the cyclical signal varies with time. It is assumed that the cyclic behavior is clear and evident, so that there is no doubt about the existence or length of the cycle. In nature, many such cyclical series depend on the

* Correspondence to: P. Paatero, Department of Physics, University of Helsinki, Box 9, FIN-00014 Helsinki/University, Finland.
† E-mail: pentti.paatero@helsinki.fi

Contract/grant sponsor: Academy of Finland
Contract/grant sponsor: Vilho, Yrjö and Kalle Foundation
Contract/grant sponsor: University of Helsinki

Accepted 3 December 1999

diurnal rhythm or on the yearly rhythm. If a series depends on two rhythms which interact multiplicatively, then it is called a doubly cyclical series. The repetitive behavior is caused by the influence of external periodic circumstances, e.g. temperature, level of illumination, rainfall patterns, etc. In this work, univariate time series are mostly discussed. Important environmental multivariate applications also exist. These applications are mentioned later on.

Using the terminology of signal processing, the elementary signals to be studied may be described as amplitude-modulated signals where the carrier wave is not sinusoidal but has a specific constant periodic shape. The measured 'composite signal' consists of a superposition of several such elementary signals. All these signals have the same period of the carrier wave. However, both the periodic shape of the carrier wave and the shape of the modulating signal are unique for each of the elementary signals. The elementary signals are to be separated from each other, based on these unique properties. At the outset, these properties are unknown. The result comprises both the periodic shapes and the modulation shapes for each elementary signal. In customary signal processing, the elementary signals that should be separated from each other are usually of different frequencies. The present problem is more difficult because all the component signals have the same fundamental frequency.

It is not assumed that the observed quantity behaves according to the assumptions which are customary in traditional 'time series analysis'. Ecological, economic and also some environmental time series may exhibit dynamic or intrinsic periodicities, so that the period is not *a priori* known, and usually the length of the period is not fixed. Such oscillations are governed by the dynamic feedback properties of the process, and such time series are well handled by customary 'time series analysis' [1,2], perhaps with the autoregressive models. Such series are not considered further in this work.

The essential idea of the present work is that the cyclical series to be analyzed is folded so that it forms a matrix. The data for one cycle form one column of the matrix, the data for the next cycle form the next column, and so on. In this way the regularities of the series are transformed so that they appear as properties of the matrix. The existing mathematical tools for analyzing matrices are well advanced: they include e.g. principal component analysis (PCA), which is based on singular value decomposition (SVD), and the new positive matrix factorization (PMF), which is especially suitable for positively constrained data [3]. These established techniques are applied in the present work: no new computer programs are needed and the mathematical properties of the solutions are already known. These techniques allow one to determine the *shapes* of the components of cyclic phenomena instead of merely identifying that a periodicity is present.

A doubly cyclical series is analogously folded so that it forms a three-way array. The three-way part of the present work is based on the PARAFAC model, sometimes called CANDECOMP:

$$x_{ijk} \approx \sum_{h=1}^{p} a_{ih} b_{jh} c_{kh} \tag{1}$$

The PARAFAC model was first introduced by Harshman [4]. Later, Ross and Leurgans [5] added positivity constraints and weighting based on standard deviations of data values. An efficient algorithm 'PMF3′ for solving PARAFAC models was recently introduced by Paatero [6].

It has been suggested to the authors that no special techniques are needed for solving the present task. According to these suggestions, the usual tools of Fourier analysis would be sufficient. In the view of the authors, this is not true. This question will be discussed in a separate section, later on.

The multivariate case

Instead of a scalar variable x, the time series may also consist of observations of a vector-valued variable $\mathbf{x}$. The preceding discussion applies there too. In the singly cyclical case the multivariate data are arranged in the form of a three-way array which may be viewed as a matrix of vector values. Again, the PARAFAC model is needed for approximating the data array. In order to keep the exposition as simple as possible, the main part of this paper describes the univariate case. The multivariate equations are given later, in a dedicated section.

Terminology, notation

The word *cyclical* denotes here a function or sequence which repeats similar behaviour with a fixed known cycle length, without repeating itself exactly. Only values measured at *discrete times* are considered. In typical real-life applications the data points would typically be hourly, daily or monthly. The examples in this work are based on hourly data obeying a 24 h or diurnal cycle. In the doubly cyclical example the longer cycle is 1 week. The notation is based on these time units. For other applications one would have to translate the units accordingly. The *span* of the measurement denotes the time span from the very first to the very last measured value.

Equidistant step function denotes a function which is constant within the open intervals $(a + nr, a + (n + 1)r)$, $n = 0,1,\ldots$. Such a function may only change at the step instances $a + nr$, $n = 0,1,\ldots$. *Superscripts* denote the individual components (factors) in a multicomponent model.

The representations of the sequence in the different time frames are called *modes*. The first mode is the cyclical diurnal (hour-to-hour) shape. In doubly cyclical models there is also a second cyclical mode. The last mode is the non-cyclical *trend mode* which covers the whole span of the measurement. The trend mode is always denoted by T. The trend mode shows how the amplitude of the oscillation(s) evolves with time.

t	time; the index of the data point in the sequences $X(t)$ and $Y(t)$
$X(t)$	the tth value of the sequence of observed data values $\mathbf{X}$
$Y(t)$	the tth value of the model sequence $\mathbf{Y}$
h	hour; hour-of-day value of any instance t
d	day; day-of-week of any instance t
w	week; week of any instance t
(h,d)	equivalent representation of t in singly cyclical models (the cycle length is 24 h)
(h,d,w)	equivalent representation of t in doubly cyclical models (24 h and 1 week cycles)
$X(h,d)$	a value of the sequence $X(t)$ of data values, indexed with hour and day indices
x_{hd}	the values of the sequence $X(h,d)$, understood as a matrix $\mathbf{X}$
$X(h,d,w)$	a value of the sequence $X(t)$, indexed with three indices (hour, day and week)
x_{hdw}	the values of the sequence $X(h,d,w)$, understood as a three-way array $\underline{\mathrm{X}}$
σ_{hd}	standard deviation of x_{hd}
σ_{hdw}	standard deviation of x_{hdw}
$H(t) = H(h)$	a strictly cyclical hourly function with a diurnal cycle of length = 24 h
$D(t)$	a function with cyclic behavior; the cycle length is 1 week in this work, but could be e.g. 1 year in other applications
$T(t)$	a non-cyclic *trend* function whose change is slow or nil within the shorter time frame(s); in this work, $T(t) = T(d)$ or $T(t) = T(w)$
p	the number of basic components (factors) in the model
υ	an index enumerating the basic components (factors), $\upsilon = 1,2\ldots,p$
m, n	the number of data points in the cycle, the number of cycles in the span
m, n, o	numbers of data points in the shorter cycle, shorter cycles in the longer cycle, and longer cycles in the span of a doubly cyclical sequence

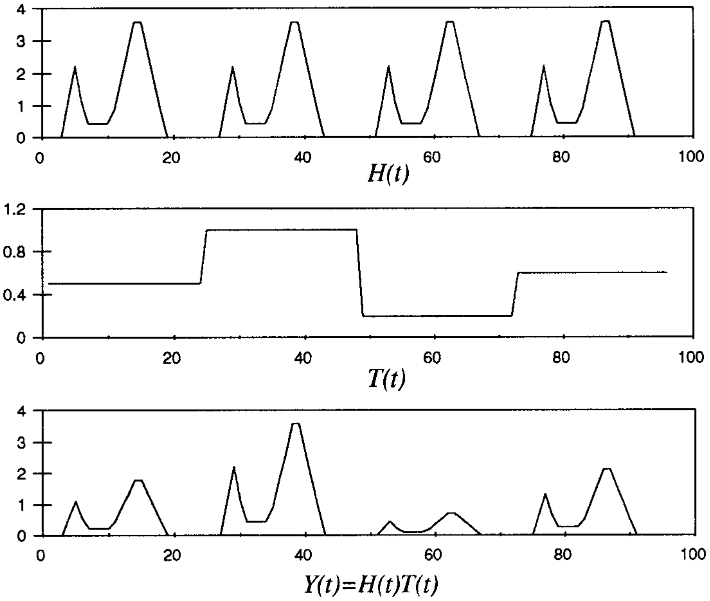

Figure 1. Synthetic example no. 1. $Y(t) = Y(h,d) = H(h,d)T(h,d)$ is a cyclic function with 24 h cycles. *H(h,d)* does not depend on *d*. *T(h,d)* is a step function which does not depend on *h*.

MATHEMATICAL MODELS FOR CYCLICAL AND DOUBLY CYCLICAL TIME SERIES

The basic component: the product of a cyclical function and an equidistant step function

The first basic building block is a sequence $Y(t) = Y(h,d)$ which has the representation

$$Y(t) = Y(h, d) = H(h)T(d) \tag{2}$$

Any value of the sequence $Y(h,d)$ is obtained as the product of an hourly value $H(h)$ and a daily value $T(d)$. The outer product of a column vector $H(h)$ ($h = 1,\ldots,m$) and a row vector $T(d)$ ($d = 1,\ldots,n$) also defines a matrix y_{hd} of dimensions (m,n). This matrix has rank = 1. The rank of the sequence is defined to be equal to the rank of the corresponding matrix. Thus the first basic building block is a sequence of rank = 1. It has the same shape on all days, but on different days it has a different amplitude or 'strength'.

The functions $H(h)$ and $T(d)$ of Equation (2) may be written as functions of t: $H(h) = H(h,d) = H(t)$ and $T(d) = T(h,d) = T(t)$. Here the function $H(h,d)$ does not depend on d, and similarly $T(h,d)$ does not depend on h. Thus $H(t)$ is a strictly cyclical function with cycle = 24 h. Similarly $T(t)$ is an equidistant step function. It has a constant value between the midnights when the steps occur.

The simplest time series model represents an observed sequence $X(h,d)$ as the sum of a rank = 1 sequence $Y(h,d)$ and random noise $E(h,d)$ according to the model

$$X(h, d) = Y(h, d) + E(h, d) = H(h)T(d) + E(h, d) \tag{3}$$

When $X(h,d)$ is given, solving the model means that the unknown vectors $H(h)$ and $T(d)$ are to be determined so as to minimize some norm of the residual sequence $E(h,d)$. Non-negativity may be required for $H(h)$ and/or for $T(d)$. Equation (3) defines a useful non-trivial problem, since some simple real series may well be represented by this model. The computational task is easy, with the solution being obtained using simple iterative techniques.

In this section, three synthetic examples demonstrate how the time series functions are composed of a cyclical shape function and a trend function. In later sections the same examples are used to show how to approximate the original functions by applying PMF to the time series. The first example, consisting of four cycles of a sequence with a cycle length of 24 h, is presented in Figure 1. The examples have been constructed so that they are unrealistically ‘easy’: the different features do not overlap each other. The shapes of the components may be discerned by carefully looking at them. In many real-life problems the overlap is severe and the shapes cannot be determined without using computational tools.

The product of a cyclical function and a slowly varying function

In the preceding subsection the trend function T was assumed to be a step function having change points at the beginning of each day (in other applications the changes could happen at each New Year, say). The sudden change is sometimes a realistic assumption, but in many situations the assumption is definitely not true. A more realistic assumption is often that the changes occur gradually. The model now becomes

$$Y(h,d) = H(t)T(t) = H(h)T(h,d) \tag{4}$$

Equation (4) represents an example of the second basic building block for modeling time series. Again $H(t)$ is a strictly cyclic function, but now the trend function $T(t) = T(h,d)$ depends on both h and d. Qualitatively it is assumed that $T(t)$ is a slowly varying function. In some cases the variation in $T(t)$ is so slow that one may safely approximate $T(h,d)$ with $T(d)$. Such approximation might be acceptable whenever random variation of the data is more significant than this artefact of the mathematical model.

A synthetic example consisting of 12 periods of a 24 h cycle is presented in Figure 2. There is an upward trend during the first 7·5 periods, after which the trend is downward.

A more complicated series: a superposition of several basic components and noise

The one-component Equations (3) and (4) are rarely adequate in real-world situations. Usually a multicomponent ‘complex’ model is needed. The complex model series $Y(h,d)$ is constructed as a sum of p basic building blocks:

$$\begin{aligned} Y(h,d) &= Y^1(h,d) + Y^2(h,d) + \ldots + Y^p(h,d) \\ &= H^1(h)T^1(h,d) + H^2(h)T^2(h,d) + \ldots + H^p(h)T^p(h,d) \end{aligned} \tag{5}$$

This series has rank = p if each $T^v(h,d)$ is replaced with $T^v(d)$. A real measured series $X(h,d)$ is approximated by the model of Equation (5) according to the key equation

$$\begin{aligned} X(h,d) &= Y^1(h,d) + Y^2(h,d) + \ldots + Y^p(h,d) + E(h,d) \\ &= H^1(h)T^1(h,d) + H^2(h)T^2(h,d) + \ldots + H^p(h)T^p(h,d) + E(h,d) \end{aligned} \tag{6}$$

where $E(h,d)$ represent the random or noise part of the measurement.

The practical computational problem, given the measurement $X(h,d)$, is to determine the unknown vectors $H^v(h)$ and $T^v(d)$ ($v = 1,\ldots,p$) (approximating $T^v(h,d)$) so that a suitable norm of the residual

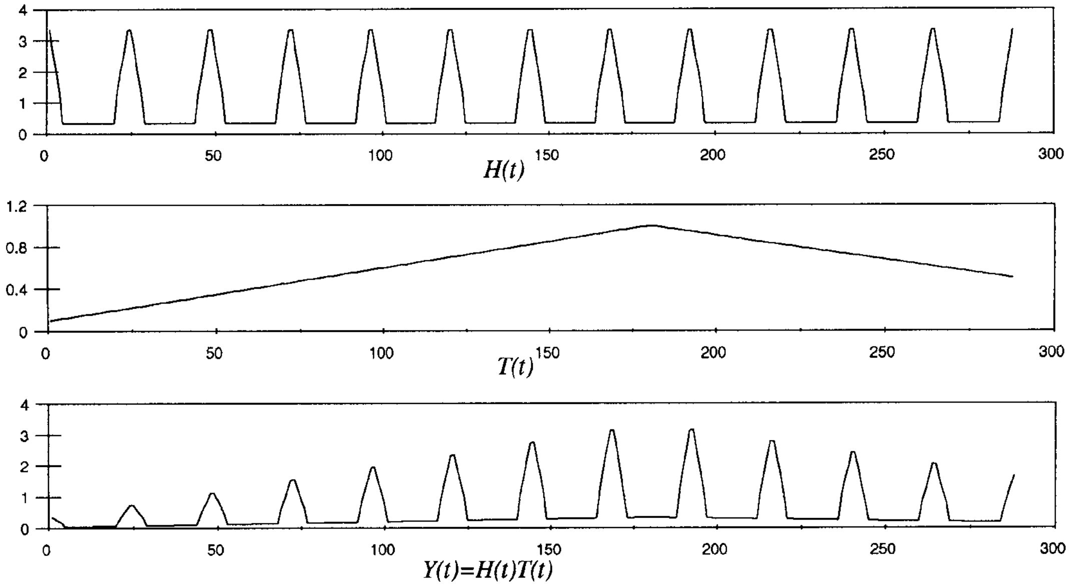

Figure 2. Synthetic example no. 2. $Y(t) = Y(h,d) = H(h,d)T(h,d) = H(h)T(h,d))$, where H is a cyclic function with 24 h cycles; H does not depend on d. The trend function $T(h,d)$ depends on both h and d.

series $E(h,d)$ is minimized. This problem is best solved with the existing techniques developed for factor analysis.

A cyclical time series with two components and added noise is shown in Figure 3d. The standard deviation of the normally distributed random errors is 30% of the mean of the series. There is always noise present in real time series owing to measurement errors, natural variability of phenomena, etc. In addition to these kinds of errors, the residuals of a model include all variation which cannot be explained by the model.

The doubly cyclical series

Two cycles may interact additively or multiplicatively. If there are two separate sources emitting the same compound so that one works with a diurnal rhythm and the other works with a weekly rhythm, then the two cycles are present in the time series in such a way that the contributions are added together. The present factor analytic model is not particularly suitable for analyzing such an additive interaction. On the other hand, the emission strength of a single source may be modulated by two periodic effects in such a way that the emission is proportional to the product of these two effects. The cycles of these effects then interact multiplicatively. In the following the factor analytic technique is extended to analyzing such doubly periodic sequences.

The general form of the one-component doubly cyclical time series model is defined as a product of three functions of time. The elements of the sequence $Y(t)$, consisting of mno elements, are represented by

$$Y(h,d,w) = H(t)D(t)T(t) = H(h)D(h,d)T(h,d,w) \qquad (7)$$

Here $H(h)$ is strictly cyclic, with a short cycle (24 h). The function $D(h,d)$ is also a strictly cyclical function whose cycle is a fixed multiple of the shorter cycle (1 week in this work). The situation is

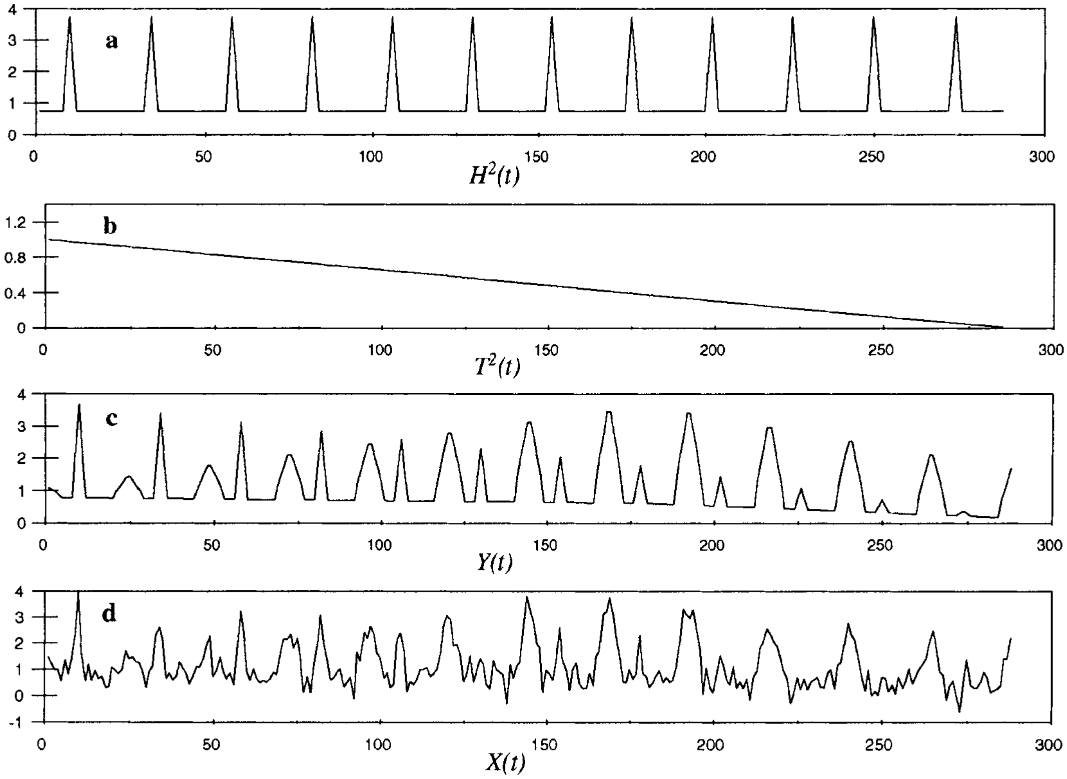

Figure 3. Synthetic example no. 3. The same series as in synthetic example no. 2 plus another cyclical function with 24 h cycles (a) and a linear downward trend (b). (c) The sum of the two functions. (d) The same with added white noise.

analogous to the singly cyclical case: both D and T are required either to be slowly varying or to be step functions. In the latter case, $D(h,d) = D(d)$ and $T(h,d,w) = T(w)$, giving

$$Y(h,d,w) = H(h)D(d)T(w) \tag{8}$$

This form is used as the basis for representing complex (multicomponent) doubly cyclical series. An observed doubly cyclical time series $X(h,d,w)$ is represented by a linear superposition of a number p of basic series and noise according to the key equation

$$\begin{aligned} X(h,d,w) &= Y^1(h,d,w) + Y^2(h,d,w) + \ldots + Y^p(h,d,w) + E(h,d,w) \\ &= H^1(h)D^1(d)T^1(w) + H^2(h)D^2(d)T^2(w) + \ldots + H^p(h)D^p(d)T^p(w) + E(h,d,w) \end{aligned} \tag{9}$$

The solution is again defined as a suitable variant of an LS fit. The sets of unknown vectors $H^v(h)$, $D^v(d)$ and $T^v(w)$ ($v = 1,\ldots,p$) are to be determined so that the chosen norm of $E(h,d,w)$ is minimized.

THE DECOMPOSITIONS SVD AND PMF OF THE MATRIX OF CYCLICAL TIME SERIES

Casting the series in matrix or array form

The key idea of the present work is that the cyclical series $X(h,d)$ be cast into matrix form x_{hd}. The values $X(h, 1)$ will form the first column of the matrix, $X(h, 2)$ the second column, and so on. The transposed arrangement is also possible. The doubly cyclical series $X(h,d,w)$ is similarly cast in the form of a three-way array, i.e. an array indexed with triple indices.

The sequences $H(h)$ and $T(d)$ from Equation (3) are now defined to be a column vector and a row vector respectively. The model series becomes a matrix y_{hd}. Similarly the residuals form a matrix e_{hd}. The problem is now a factor analytic task where the number of factors is $p = 1$.

In the multicomponent Equation (6), all the column vectors $H^{v}(h)$ are assembled into a matrix **H**. Similarly all the row vectors $T^{v}(d)$ are assembled into a matrix **T**. In matrix notation the equation takes the simple form

$$\mathbf{X} = \mathbf{HT} + \mathbf{E} \tag{10}$$

or in component notation

$$x_{hd} = \sum_{\nu=1}^{p} h_{h\nu} t_{\nu d} + e_{hd} \quad (h = 1, \ldots, m, \quad d = 1, \ldots, n) \tag{11}$$

Similarly Equation (9) for doubly cyclical series is written in component notation as

$$x_{hdw} = \sum_{\nu=1}^{p} h_{h\nu} d_{d\nu} t_{w\nu} + e_{hdw} \quad (h = 1, \ldots, m, \quad d = 1, \ldots, n, \quad w = 1, \ldots, o) \tag{12}$$

where the factor matrices **H**, **D** and **T** have been formulated so that the time sequences run along their columns (not rows). Equation (12) is an example of the PARAFAC model.

Solving Equation (11) means that when **X** is given, the unknown matrices **H** and **T** are determined so that a chosen norm of the matrix **E** is minimized. The most basic choice is to minimize the sum of the squares of elements of **E** or the ‘Frobenius norm’ of **E**. The minimum is found by principal component analysis (PCA). The standard solution is based on singular value decomposition (SVD): compute the SVD of **X** in the form $\mathbf{X} = \mathbf{USV}^{\mathrm{T}}$ and keep only the p most significant singular components of **U, S** and $\mathbf{V}^{\mathrm{T}}$ Depending on the desired normalization, the solution of the PCA problem (11) may be taken as either ($\mathbf{H} = \mathbf{US}$, $\mathbf{T} = \mathbf{V}^{\mathrm{T}}$) or ($\mathbf{H} = \mathbf{U}$, $\mathbf{T} = \mathbf{SV}^{\mathrm{T}}$).

The PCA solution usually contains negative values. These are not desirable if the quantities of the model are inherently non-negative (mass, number of individuals, energy, etc.). Also, in customary PCA, all data values have equal weight: PCA may thus only be optimal if all data values have equal or approximately equal errors. The new technique of ‘positive matrix factorization’ (PMF) corrects for these deficiencies. For comparisons of PCA and PMF and for more details of PMF, see References [3,6–8]. The individual standard deviations are also taken into account by the new maximum likelihood principal components approach of Reference [9].

According to PMF, the quantity to be minimized in the LS fit is

$$Q = \sum_{h=1}^{m} \sum_{d=1}^{n} \frac{e_{hd}^2}{\sigma_{hd}^2} = \sum_{h=1}^{m} \sum_{d=1}^{n} \frac{\left(x_{hd} - \sum_{\nu=1}^{p} h_{h\nu} t_{\nu d}\right)^2}{\sigma_{hd}^2} \tag{13}$$

Usually the minimization of this Q is constrained by non-negativity constraints for the unknowns h_{hv} ($h = 1,\ldots,m$, $v = 1,\ldots,p$) and t_{vd} ($v = 1,\ldots,p$, $d = 1,\ldots,n$). The values σ_{hd} are the known (or assumed) standard deviations for each element of the data matrix **X**.

For solving the three-dimensional PARAFAC model (12), one has to minimize the object function

$$Q = \sum_{h=1}^{m}\sum_{d=1}^{n}\sum_{w=1}^{o}\frac{e_{hdw}^2}{\sigma_{hdw}^2} = \sum_{h=1}^{m}\sum_{d=1}^{n}\sum_{w=1}^{o}\frac{\left(x_{hdw} - \sum_{\nu=1}^{p} h_{h\nu}d_{d\nu}t_{w\nu}\right)^2}{\sigma_{hdw}^2} \tag{14}$$

using iterative techniques. The solution matrices **H**, **D** and **T** may be required to be non-negative. Here the three-way array $\boldsymbol{\sigma}$ is analogous to the matrix $\boldsymbol{\sigma}$ of Equation (13). Several programs exist for solving this difficult task; for a comparison, see Reference [10]. In the present work the new program PMF3 was used.

Robust analysis

The distribution of environmental data (particularly concentrations) is typically skewed, with a small percentage of very large values. Mathematical transformations of the data (log or square root) are often used for controlling the influence of the largest values. However, non-linear transformations may distort the linear structure of the model [7]. In the present approach, standard deviations for residuals are specified proportional to data values. In this way, each large value gets weighted down because of the large standard deviation assigned to it, and no transformations are necessary.

The programs PMF2 and PMF3 may be set to work in a *robust mode* according to the Huber principle: the weights for outlying data points are dynamically decreased during the iteration so that a statistically robust factorization is obtained [8]. A data point is considered outlying it its residual exceeds the corresponding standard deviation by a user-specified factor, e.g. four. Such downweighting guarantees that a few outlying values may not totally ruin the result. This feature is extremely valuable when analyzing environmental data, which may contain non-representative or erroneous values. The examples in the present work were run in the robust mode.

Although some of the outlying values often are gross errors, other outlying values need not be in error at all. In order to determine the recurrent features of the data, the exceptional values need to be weighted down even if they are fully legitimate values. An example: one analyzes airborne dust and tries to attribute concentrations to components originating in different deserts. The objective might be e.g. to find out whether the dust emissions of some deserts are increasing with time. Dust emitted by a large volcanic eruption would cause a few outlying non-erroneous values in the measured time series. If these values are not somehow excluded from the analysis, the result could be utter nonsense. In some situations it may be necessary to report separately on the excluded values. Downweighting the outlying values in the periodic analysis does not justify ignoring them altogether, unless they are really considered to be gross errors.

SVD and PMF of the synthetic example matrices

Synthetic example no. 1. There is only one factor in the synthetic step function example no. 1, and no noise. The rank is then rank = 1 and the SVD of the data matrix y_{hd} indicates only one non-zero singular value (9·58). Both SVD and PMF find the only factor exactly (Figure 4a). It can be seen that the shape of the cycle is just the same as in Figure 1, and also the relative steps are the same.

Synthetic example no. 2. Although there is only one true factor in synthetic example no. 2, the three first singular values of the matrix y_{hd} were non-zero: 16·33, 1·20 and 0·05. Three factors are needed in the equation $\mathbf{Y} = \mathbf{USV}^{\mathrm{T}}$ to reproduce **Y** exactly. However, since the significance of the second and the third factor is small, the shapes of the 24 h cycle and the trend function are shown by the first factor of the SVD solution in Figure 4b. The shape of $(\mathbf{US})_{h1}$ is very similar to the shape of one cycle of *H(t)* in Figure 2, but not exactly the same. The points ν_{d1} correspond to the

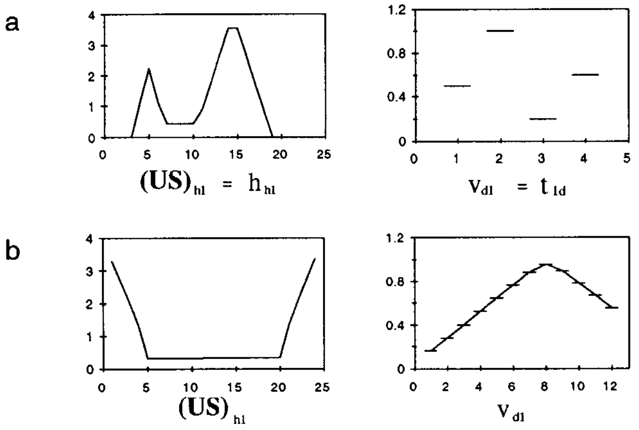

Figure 4. (a) The exact and only factor of synthetic example no. 1. The same values were found by both SVD and PMF. SVD: $y_{hd} = (\mathbf{US})_{h1}(\mathbf{V}^{\mathrm{T}})_{1d}$. PMF: $y_{hd} = h_{h1}t_{1d}$. (b) The first factor of synthetic example no. 2, as solved by SVD.

daily mean values of the function *T(t)* in Figure 2. The one-factor solution of PMF agreed with the first factor given by SVD within the graphical resolution.

Synthetic example no. 3. The noisy two-component example, shown in Figure 3d, was analyzed with PMF and with SVD. The shapes of the factors found by PMF (Figure 5) approximate well the original shapes used to compose the example (see Figures 2 and 3). The singular values were obtained as 22·87, 5·66, 2·51, 2·26, 2·10, 1·75,…, 0·54. However, the factor shapes produced by SVD are not meaningful without auxiliary rotation.

THE CARBON MONOXIDE EXAMPLE

As a real example, carbon monoxide concentrations measured hourly in the city of Helsinki during the year 1994 are analyzed. The measuring site is situated in the near vicinity of the crossing of several dense-traffic roads. Traffic is the predominant source of carbon monoxide in urban areas [11,12], and consequently the series is expected to be doubly cyclical with diurnal and weekly cycles. This example is called the *CO example*.

The mean of the 8718 hourly CO concentrations was 0·9 ppm, the median 0·7 ppm and the maximum 9·0 ppm. Of the hourly concentrations, 99% were below 3·4 ppm and 95% below 2·2 ppm.

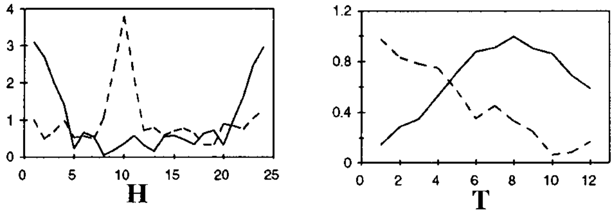

Figure 5. Unrotated PMF results of the noisy two-component synthetic example no. 3 (See Figure 3d).

The CO example is relatively simple and could probably be analyzed without sophisticated techniques. The 'true answer' to this example is fairly well known, and thus the result is without much scientific value. However, it is felt that such an 'easy' case is better suited for illustrating the technique than difficult cases where one could perhaps argue about the correct result. An easy case is also a better demonstration for scientists working in other fields of science.

Analyzing for the diurnal cycle only

The CO concentrations were analyzed by the two-way PMF by arranging the data in a matrix with 24 rows (hours) and 365 columns (days). The standard deviations of the data values were derived by assuming a 0·1 ppm absolute error and a 15% relative error in each data point. These values were suggested by expert opinion of the experimentalist. These standard deviations were used throughout the analysis. Sometimes the initially assumed values for standard deviation need to be refined during the analysis if the size of residuals is in conflict with the assumed standard deviation values. In this work the initially assumed values needed no refinement. No attempt was made at studying the distribution of the concentration values.

The solution of the two-factor PMF is presented in Figures 6a and 6c. Because both factors have lowest values during early morning hours, the PMF run was repeated by arranging the data matrix so that the first measurement started at 3 am (Figures 6b and 6d). It is beneficial to arrange the ends of cycles to be in the middle of a low-concentration period in order to minimize the risk of an artefact jump when connecting the ends of cycles. This arrangement of the data matrix in effect changes the

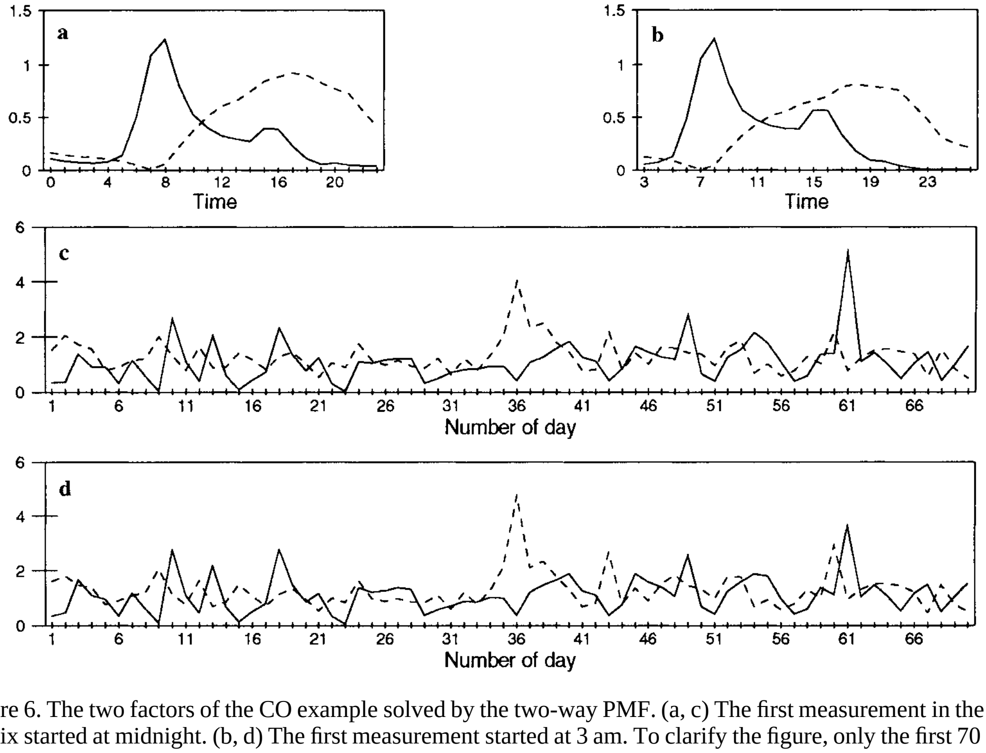

Figure 6. The two factors of the CO example solved by the two-way PMF. (a, c) The first measurement in the data matrix started at midnight. (b, d) The first measurement started at 3 am. To clarify the figure, only the first 70 days are shown. 1 January 1994 was a Saturday.

place of the steps in the model step function $T(h,d)$: the steps now occur at 3 am instead of at midnight. The steps may be interpreted as follows. At the time of the step the modeled traffic pattern switches over from the pattern of the previous day to the pattern of the next day. If the step were to occur at noon, say, it would mean that the model predicts that at noon the number of cars on the streets suddenly jumps. The jump would represent an increase or decrease, depending on whether the coefficient for the next day is higher or lower than for the previous day.

In the diurnal mode of the first factor (full line) a strong maximum can be seen during the rush hours in the morning and another maximum in the afternoon. In Finland the working time is usually from 7–8 to 16–17 five days a week. The diurnal mode of the second factor starts to rise at about 8 am, has a broad maximum during the afternoon and evening hours and decreases towards midnight. In the trend mode the first factor tends to have its lowest values at weekends, while the second factor behaves in the opposite way. The high value on the 36th day (5 February, a Saturday) was due to meteorological conditions: high pressure and an inversion situation with low temperature and a low wind speed, causing stagnant air. Day 61 was Wednesday 2 March, and at that time there was also high pressure and an inversion situation over Scandinavia.

The difference between the solutions when the time series was started at midnight (Figures 6a and 6c) and at 3 am (Figures 6b and 6d) is mostly due to rotational freedom in the solutions. Rotational ambiguity is always present in two-way factor analysis unless non-negativity (or other additional constraints) prevents rotations of the solution. The second trend factor (having the maximum concentration at noon) in Figures 6c and 6d is non-zero everywhere. Such rotations are allowed where a fraction of the first trend factor is subtracted from the second, while a similar fraction of the second diurnal shape is added to the first diurnal shape. Thus both sets of solutions should be considered as valid. In fact, the domain of rotationally possible solutions extends even further in the direction where the afternoon rush hour maximum of the first trend factor increases.

With midnight and 3 am starting times the values of Q in Equation (13) were 10 788 and 10 479 respectively. The better fit is probably mostly due to improved fitting of inversion situations where the night-time concentrations decrease much slower than in normal weather conditions. However, the difference between these Q values is rather small and may be called ‘insignificant’. In further two-way and three-way analyses the time series starting at 3 am will be used. It is noted in passing that ideally (i.e. if standardized residuals are independent and normally distributed) the quantity Q should have a χ^2 distribution with the number of degrees of freedom slightly smaller than the number of points in the time series. The values obtained for Q are too large, approximately by a factor of 1·25, suggesting that the assumed standard deviation values have been too optimistic. In principle, the analysis should be repeated with standard deviation values increased by a factor of $\sqrt{1{\cdot}25} = 1{\cdot}12$. However, in practice, such a small overall adjustment of the standard deviations did not change the results noticeably.

The two-way PMF analysis was also run with three factors, but the result could not be explained in a useful way and did not give any more information about the formation of the CO concentrations.

Analyzing for both the diurnal and 7 day cycles

In the trend mode of the two-way PMF (Figure 6), faint weekly periods could be seen. In order to find out the shapes of these cycles, the PARAFAC model was tried, by using the three-way program PMF3. The time series of hourly CO concentrations was arranged as a three-way array starting at 3 am on 1 January 1994 (Saturday).

The two-factor solution was also the most useful when running the three-way PMF (Figure 7). The high concentrations during the morning and afternoon rush hours can be seen clearly in the diurnal mode of the first factor (full line). The shape of the diurnal mode of the second factor is now quite

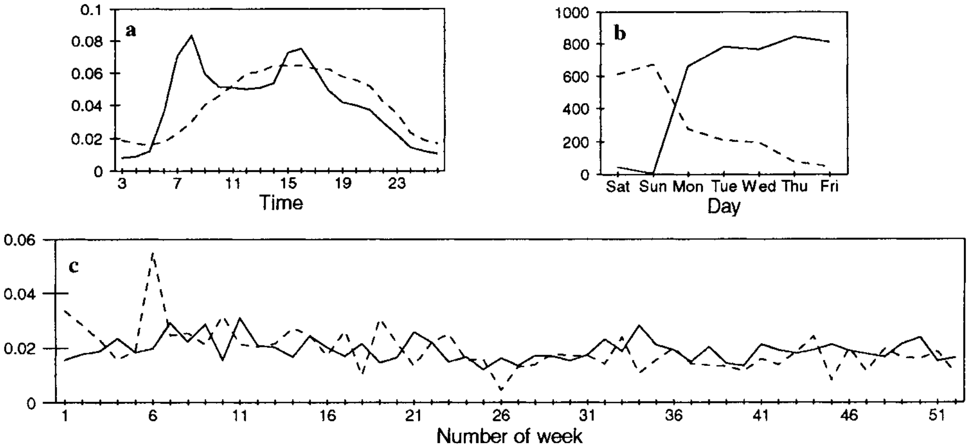

Figure 7. The two factors of the CO example solved by the three-way PMF. Starting time at 3 am on Saturday 1 January 1994.

symmetric, with a broad maximum from about noon to about 7 pm. The 7 day mode of the first factor has high values during weekdays and is almost zero at the weekend, while the second 7 day mode has its highest values at the weekend. The highest values of the trend mode of the second factor are seen in February and the lowest near midsummer. The variability of the first trend factor is much smaller and has no clear seasonality. The lack of seasonality is in accordance with Derwent *et al.* [13], who report that traffic flows observed on weekdays do not exhibit any seasonality.

As explained in the previous subsection, the location of the step in the step function is influenced by the arrangement of the data matrix. The three-way model contains steps in two directions. On the basis of the two-way PMF, it was decided to start the diurnal cycle at 3 am. In order to decide the starting day of the week, the three-way PMF was run repeatedly, starting on different days. The range of the *Q* values of the solutions was quite narrow, only 4%. The worst fits (the highest *Q* values) were connected with starting on Saturday or Sunday. The solutions obtained when starting on Tuesday, Wednesday, Thursday (Figure 8) or Friday were almost identical, so that Figure 8 is representative of them all. It can be seen that the diurnal cycles of both factors are similar in both Figure 7 (starting on Saturday) and Figure 8, but the weekly cycles differ. In Figure 7, Friday behaves just like the other weekdays, but in Figure 8 it differs from the other days by having high values for both factors. It is known that Friday differs from other working days. Especially in summer, people start to leave for their summer cottages earlier and the night traffic also continues longer than on the other weekdays. Thus the model in Figure 8 is acceptable and preferable to Figure 7, which has probably been distorted by the location of the step in the weekly cycle.

ANALYZING THE PROBLEM WITH THE TOOLS OF FOURIER ANALYSIS

Several sinusoidal signals of different frequencies are easily separated from each other and from noise by using Fourier analysis. Then one simply picks those frequency components which rise sufficiently high above the average noise level. The *shape* of one non-sinusoidal periodic signal is also efficiently analyzed with Fourier analysis: then one looks for an equidistant set of frequency components with above-noise amplitude. The pattern of amplitudes of these different harmonic (‘overtone’) frequencies is related to the periodic shape of the signal. Yet another simple case is an amplitude-modulated sinusoidal signal, i.e. a signal which is the product of a sine curve and a slowly varying

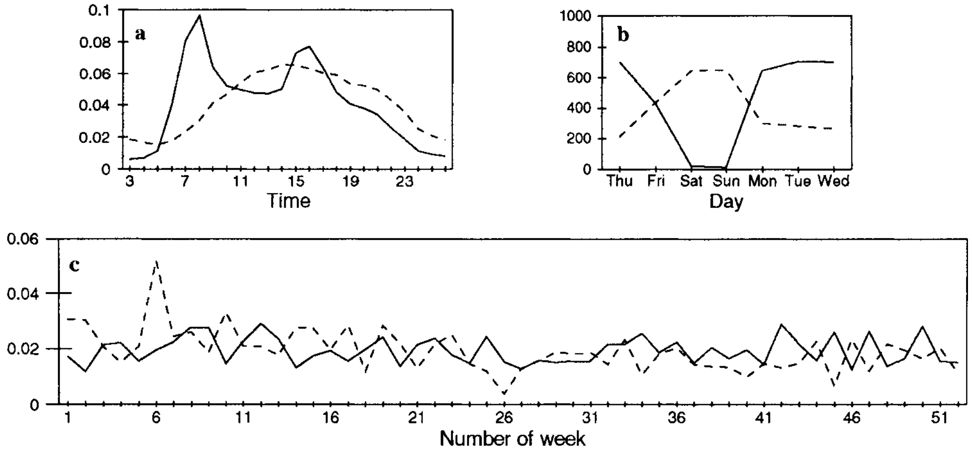

Figure 8. The two factors of the CO example solved by the three-way PMF. Starting time at 3 am on Thursday 30 December 1993.

trend curve. The Fourier transform (FT) of such a product signal is the convolution of the Fourier transforms of the two signals. The carrier frequency is spread into a narrow band whose narrowness results from the slowly varying nature of the trend signal.

The picture becomes less clear when the FT of a non-sinusoidal signal with varying amplitude is considered. As an example, the synthetic example from Figure 2 is presented. The computed FT, shown in Figure 9a, is also a convolution of the FT of the trend signal with the FT of the non-sinusoidal periodic signal. It is seen that both the basic frequency and all the harmonics are spread into bands of identical shape but different intensities. The properties of the original signal may still be recovered as follows. By integrating each band around the equidistant frequencies, one obtains the FT of the shape curve and hence the shape. By averaging all the bands, one obtains the FT of the trend curve (the modulating signal) and hence the trend curve. However, nothing special is gained by doing this analysis in the frequency domain. The operations corresponding to averaging and integrating could also be performed in the time domain when the cycle length is known.

Figure 9b shows the FT of the noise-free synthetic example from Figure 3c. Now there are two superposed component signals, each consisting of a non-sinusoidal periodic curve multiplied by a slowly varying trend curve. The same set of frequency bands is visible as in the previous case. However, now each band contains information from *two* components (and also some noise if real data are analyzed). There is no simple way of separating these components from each other, although the information is there, of course. Even if a technique could be devised for analyzing the information in the frequency bands, the non-negativity information would not be available; this useful auxiliary information is only present in the time domain. Similarly, weighting of individual data points and downweighting of outliers are not possible in the frequency domain. It is concluded that separating the components is not straightforward by using Fourier analysis, and even if the separation could be effected, there are several drawbacks in this approach. In practice, Fourier analysis is limited to finding the *average* periodic shape and the *average* trend behavior, averaged over all individual components present in the system.

Figure 9c shows the FT of the 8736 h (52 weeks) of the CO time series. The peak at frequency 0·0417 cycles per hour corresponds to the basic 24 h period. The harmonic frequencies are visible at 0·0833 and 0·125. The weekly periodicity is visible in the side peaks which are 0·006 units to the left and to the right of the basic peak and its harmonics. The overall trend may be visible in the spreading

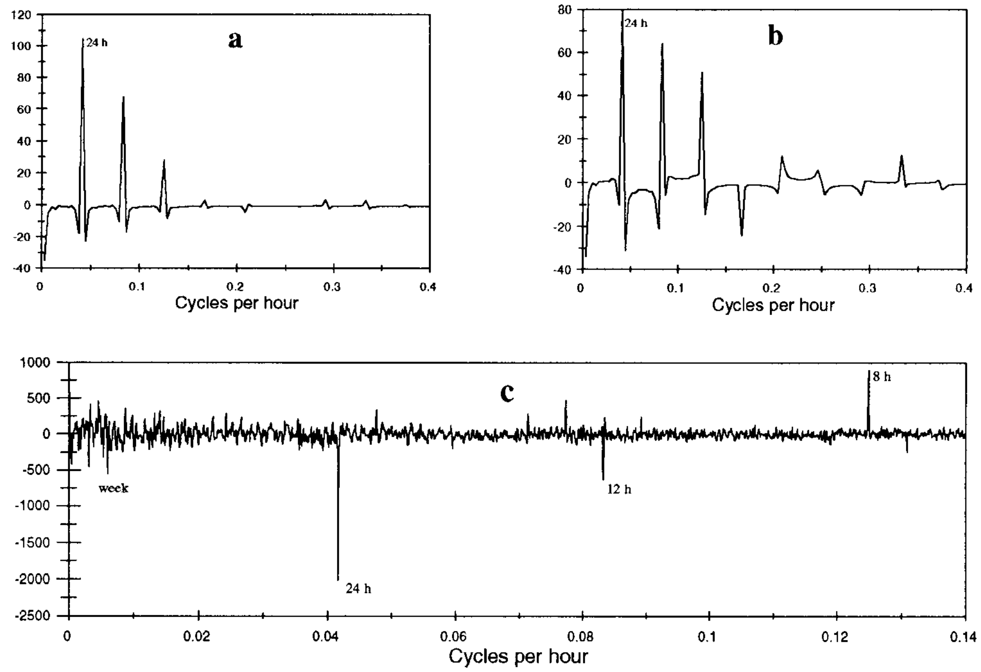

Figure 9. Real parts of the Fourier spectra (excluding the zero-frequency peak) of the examples. (a) Synthetic example no. 2 (See Figure 2). (b) Synthetic example no. 3 (See Figure 3c). (c) The CO example.

of all these peaks. It is seen that the doubly periodic situation creates an even more complicated Fourier transform. By using Fourier analysis, it would be possible to obtain the average weekly shape. However, there would be no way to derive the weekly shapes of the individual components.

ANALYZING CYCLICAL MULTIVARIATE TIME SERIES

Many environmental time series are in fact multivariate: either there are several parallel simultaneous measurements of one quantity, made at different locations, or several different quantities are measured in parallel from each sample. When atmospheric pollution is monitored, concentrations of several chemical elements or compounds are often determined from each sample.

The univariate techniques presented in this work are easily generalized for analyzing multivariate time series. The basic assumption is that each source has a constant profile among the sets of parallel observations: either the spatial distribution due to any individual source stays constant with time, or the chemical composition of the emission from each source stays constant. The multivariate analogy of Equation (11) is then the PARAFAC model

$$x_{chd} = \sum_{\nu=1}^{p} a_{c\nu} h_{h\nu} t_{d\nu} + e_{chd} \quad (c = 1, \ldots, C,\ h = 1, \ldots, m,\ d = 1, \ldots, n) \tag{15}$$

(In three-way equations the factors usually correspond to columns in all three factor matrices. This differs from the customary two-way notation.) The first index c enumerates the parallel observations. The second and third indices correspond to the first and second indices in Equation (11). The columns of the first factor matrix **A** represent the profiles of the individual sources: how strongly each source contributes to each of the parallel observations. The meaning of the matrices **H** and **T** is the same as in the univariate case.

The multivariate model has the same drawback as the univariate model: the trend functions $T^{\nu}(d)$ are equidistant step functions. In principle, this causes the modeled concentrations to contain sudden jumps at cycle end points (at midnight or at New Year). This may be objectionable in some applications.

A successful application of the multivariate technique has been published by Xie *et al.* [14]. Arctic aerosol composition data (weekly measurements during 11 years) were analyzed. These data exhibit a very clear yearly cycle. The pollutant concentrations due to different sources reach their maxima in the Arctic during different parts of the year. The year-to-year trend, if any, would reflect global changes in the atmosphere, either natural or anthropogenic. The aerosol data are arranged in a three-way array x_{cwy} so that the first index c enumerates the 24 different compounds whose concentrations have been measured, the second index w enumerates the weeks within a year, and the third index y enumerates the years. The columns of the factor matrices **A**, **H** and **T** are interpreted as follows: a_{ck} represents the composition profile for factor k, h_{wk} shows the seasonal shape of the concentrations due to factor k, and finally t_{yk} represents the trend behavior of factor k.

DISCUSSION

The important question of determining the number p of components has been extensively discussed in the factor analytic literature: for PCA, see References [15,16]; for PMF, see References [8,17]. For factor analytic treatment of time series problems, these references may be consulted.

No statistical criteria are currently available for estimating the confidence limits of results. In practice, one has to gain confidence in the results by repeating the analysis on several sets of similar data, collected e.g. during different years or from neighboring similar geographical locations. By comparing the results, one may reject those results which are caused by random variation of the data or by local peculiarities which invalidate the model for some data sets. Alternatively, one might be able to compare some of the computed results with previously known facts. If no comparisons are possible and the noise level in the data is high, the results should not be trusted because of the risk that they might just reflect the noise of the data.

The results of the two-way model suffer from rotational indeterminacy, familiar from factor analysis: different combinations of periodic and trend shapes produce identical fits to the data. Depending on the data, non-negativity constraints may eliminate some or all of this uncertainty. The doubly cyclical three-way model and the multivariate cyclical model are basically free from the rotational uncertainty.

Results of the examples

The simple synthetic examples with one or two components showed that positive matrix factorization (PMF) could find the factors well even in the presence of high-amplitude noise. The solutions obtained using traditional factor analytic methods (singular value decomposition or principal component analysis) were not equally useful. There was also a minor rotational uncertainty in the results given by PMF. Rotational uncertainty is always a reality in factor models, and PMF is of course no exception.

The solution of the CO example found by the two-way PMF showed the realistic diurnal variation of the concentrations. A large amount of rotational uncertainty was seen in these results. In the day-to-day trend mode, hints of some 7 day periodicity were vaguely visible. The shape of the 7 day cycle was only found by three-way factor analysis. Because the model forces the changes in both the weekly cycle and the trend function into discrete steps, the locations of these steps influence the results: different starting times in both directions (time of day, day of week) lead to slightly different factorizations. By repeating the analysis with different starting times, it was possible to pick stable representative solutions.

Similar three-way analysis was performed on other CO data sets, measured in the years 1990–1993. The diurnal shapes were found to be stable, similar in all results. More variation was encountered in the weekly shapes. For the work-related factor the ratio of the largest and smallest values among (Mon, Tue, Wed, Thu) is below three to two. There is no clear pattern in the variation. In all years the values obtained for Friday are intermediate between weekday and weekend values. The differences in weekly shapes of different years were similar to the differences between Figures 7b and 8b.

Standard deviations of data values

All factor analysis is based on implicit assumptions about the standard deviations of the data values, although this basis is not generally mentioned in textbooks [7]. In fact, most factor analysis assumes that all standard deviations are equal after the scaling done by standardizing the columns (or rows). It is essential that one communicates to the model the best information that is available about the data, including particularly information about the standard deviations. In addition to the PMF technique, individual standard deviations are taken into account by the maximum likelihood principal components approach of Reference [9].

If there is no better information than ‘all standard deviations are equal’, then one should use this knowledge and specify e.g. $\sigma_{hd} = 1$ for all h and d. Usually there is at least some information about the accuracy of experimental data, e.g. it is known that small concentrations (near the detection limit) have a larger relative laboratory error but a smaller absolute laboratory error than large concentrations. Such information should be expressed by means of the σ_{hd} values.

Comparisons with customary time series techniques

The autoregressive (AR) models predict each new data value based on a number of earlier values. The primary result is a set of coefficients describing this dependence. In AR models, additive trend may be included, but multiplicative trend, as needed for describing the varying amplitude of the periodic shape, is not available. There is no way of separating the composite signal into several components with the same frequency but unique periodic and trend shapes. Non-negativity is not included in autoregressive models. The AR techniques are especially useful when the period of the signal is not fixed. It is seen that AR techniques are so different that no numerical comparisons are meaningful.

As discussed in a preceding section, Fourier analysis is able to extract the average periodic shape and the average multiplicative trend from a cyclic time series. However, individual shapes and trends cannot be obtained for the individual components or ‘factors’. Numerical comparisons are not meaningful between the average result on one hand and individually separated results on the other hand.

Special properties of the doubly cyclical model and the multivariate model

The properties of the singly cyclical model derive from two-way factor analysis. These properties are well known. The general two-way solution contains rotational ambiguity. Requiring non-negativity eliminates some of the rotations; depending on the data, the result sometimes becomes well defined without any rotational uncertainty at all. On the other hand, requiring non-negativity and/or applying individual weighting of data values may generate local minima of the object function Q which is minimized in the LS fit. Thus in some cases the factorization problem does not have a unique solution. The scientist has to explore the different solutions. Sometimes they do not differ significantly from each other; sometimes some solutions may be non-physical so that they may be discarded; and finally in some cases it may be proper to report more than one possible interpretation of the data. These questions have been discussed by Paatero [8].

The properties of the PARAFAC solution are less well known, and many research problems still remain open. It has been shown by Kruskal [18] that under rather general conditions the PARAFAC model does not have rotational freedom. Simplified, this theorem states that if all three factor matrices are of full rank, then the factors cannot be rotated without making the fit worse. In clear-cut cases the uniqueness is a strong result: one may accept the factorization as it is without considering alternative rotational forms. There are borderline cases, however. It is intuitively clear that if the factor matrices are almost rank-deficient, then the Kruskal theorem is of little value: although a rotated solution would have a worse fit, the increase in the Q value would not be significant. As an example, assume that two factors have practically the same diurnal shape. Then there is rotational freedom between the 7 day shapes and the trend shapes of these two factors unless non-negativity prevents the rotations.

Another problem with the PARAFAC model may be caused by the existence of several local solutions. It is always prudent to assume that there are multiple solutions. Only if the same solution keeps reappearing when running the analysis with several pseudorandom starting points may one be satisfied that the solution is unique. One should note that the existence of competing solutions (i.e. local minima of Q) is a property of the model and not of the algorithm used for solving the model.

Future developments

The CO example demonstrated that the placement of cycle starting times influences the results. This was caused by the fact that a smooth trend function was approximated by a step function, causing a difference between the mathematical model and the real world. Depending on the placement of the step, the difference influences the result in different ways. The least distorted and most plausible result is obtained by placing the discontinuity to a moment where the periodic shape is at its minimum.

As long as standard factor analytic software is used, there is no way to avoid the discontinuity in the mathematical model. However, it is also possible to define a mathematical model where the trend function is defined as a truly smooth function without artificial discontinuities. Then the significance of cycle starting times will disappear entirely. Such models may be easily formulated and solved with the new program 'Multilinear Engine' [19]. Results of these experiments will be reported later.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Environmental Office of the Helsinki Metropolitan Area Council for the carbon monoxide data, and the Academy of Finland, the Vilho, Yrjö and Kalle Väisälä Foundation and the University of Helsinki for financial support.

REFERENCES

1. Andersson TW. *The Statistical Analysis of Time Series*. Wiley: Chichester, 1994.
2. Khalil MAK, Moraes FP. *J. Air Waste Mgmt. Assoc.* 1995; **45**: 62–63.
3. Paatero P, Tapper U. *Environmetrics* 1994; **5**: 111–126.
4. Harshman RA. Foundations of the PARAFAC procedure. *UCLA Working Papers Phonet*. 1970; **16**: 1–84.
5. Ross RT, Leurgans S. *Methods Enzymol.* 1995; **246**: 679–700.
6. Paatero P. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 223–242.
7. Paatero P, Tapper U. *Chemometrics Intell. Lab. Syst.* 1993; **18**: 183–194.
8. Paatero P. *Chemometrics Intell. Lab. Syst.* 1997; **37**: 23–35.

9. Wentzell PD, Andrews DT, Hamilton DC, Faber K, and Kowalski BR. *Chemometrics Intell. Lab. Syst.* 1997; **11**: 339–366.
10. Hopke PK, Paatero P, Jia H, Ross RT, Harshman RA. *Chemometrics Intell. Lab. Syst.* 1998; **43**: 25–42.
11. Zhang Y, Stedman DH, Bishop GA, Guenther PL, Beaton SP. *Environ. Sci. Technol.* 1995; **29**: 2286–2294.
12. Aarnio P, Hämekoski K, Koskentalo T, Virtanen T. In vol. 1, Tolvanen M, Anttila P, Kämäri J (eds). *Proceedings of the 10th Clean Air Congress*, The Finnish Air Pollution Prevention Society, 1995; paper 201.
13. Derwent RG, Middleton DR, Field RA, Goldstone ME, Lester JN, Perry R. *Atmos. Environ.* 1995; **29**: 923–946.
14. Xie Y-L, Hopke PK, Paatero P, Barrie LA, Li S-M. *J. Atmos. Sci.* 1999; **56**: 249–260.
15. Malinowski ER. *Factor Analysis in Chemistry*. Wiley: New York, 1991; 83–120.
16. Henry RC, Park ES, Spiegelman CH. *Chemometrics Intell. Lab. Syst.* 1999; **48**: 91–97.
17. Juvela M, Lehtinen K, Paatero P. *Mon. Not. R. Astronom. Soc.* 1996; **280**: 616–626.
18. Kruskal JB. *Linear Algebra Appl.* 1977; **18**: 95–138.
19. Paatero P. *J. Comput. Graph. Statist.* 1999; **8**: 854–888.