ENVIRONMETRICS
Environmetrics 2007; **18**: 823–840
Published online 8 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.842

Searching for structure in measurements of air pollutant concentration

Isabella Morlini*,†

Dipartimento di Scienze Cognitive, Sociali e Quantitative, Università di Modena e Reggio Emilia, Viale Allegri 9-42.100 Reggio Emilia, Italy

SUMMARY

When studying air pollution measurements at different sites in a spatial area, we may search for a typical pattern, common to all curves, describing the underlying air pollution process in a pre-specified period. Another area of interest to support local authorities in air quality management may be the classification of the different sites in homogeneous clusters and the group ranking that follows. Yet, there is variation in both amplitude and dynamics among the air pollutant concentrations measured at the different monitoring stations. Analyzing such measurements, where the basic unit of information is the entire observed process rather than a string of numbers, involves finding the time shifts or the warping functions among curves. The analysis is much more complicated if we consider a multivariate process, that is, vector-valued air pollutant measurements. Following our previous work where an improved dynamic time-warping algorithm has been developed, especially in the multivariate case, and used both for classifying functional data and estimating the structural mean of a sample of curves, we analyzed the measurements of some air pollutants in Emilia Romagna (northern Italy). In addition, for the univariate analyses, we applied the self-modeling warping function approach, which is also convenient for these data. Indeed, this method was found to be model-free and enough flexible to capture very complex and highly non-linear patterns.

KEY WORDS: cross-sectional mean; functional data; landmarks; warping functions

1. INTRODUCTION

The paper is concerned with measurements of air pollutant concentrations at selected monitoring sites throughout a spatial area. The measurements gathered from each station may be considered as functional data, that is, a sample of curves with some peculiar features. A typical structural pattern describing an underlying ecological process may be thought to be common to all curves. On the other hand, realizations of this typical shape at each site show different dynamics and intensity. In particular, peaks are shifted from site to site due to weather and environmental factors. Differences in timing of the peaks and other salient features (local minima, local maxima, inflection points, etc.) complicate the

*Correspondence to: I. Morlini, Dipartimento di Scienze Cognitive, Sociali e Quantitative, Università di Modena e Reggio Emilia, Viale Allegri 9-42.100 Reggio Emilia, Italy.
†E-mail: morlini.isabella@unimore.it

Received 9 November 2005
Accepted 11 January 2007

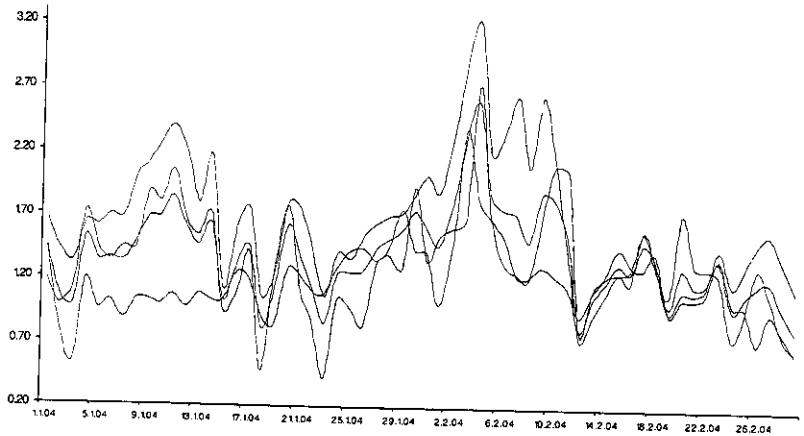

Figure 1. CO curves and cross-sectional mean (heavy dashed line)

analysis of these measurements. For example, in estimating the mean function, the naïve estimator, the cross-sectional mean, does not always reflect the average pattern. Due to a shift, also called time (or phase) variability, a structure is smeared or it may also disappear. The problem is illustrated in Figure 1, where carbon monoxide (CO) curves (as measured by three air quality monitoring stations in the city of Bologna [northern Italy]) are plotted together. The time axis varies from 1 January 2004 to 29 February 2004. From an ecological point of view, it is important to determine the daily maximum average concentration, since high concentrations of CO have negative effects on vegetation and human health. As Figure 1 shows, the cross-sectional mean (heavy dashed line) underestimates some of these peaks. For example, the maximum mean value around 4 February is underestimated, ignoring the time variation, since it occurs a few hours earlier for one of the three stations.

As to the classification problem, the goal is to provide information in order to support authorities for structural measures for emission reductions. If the basic unit of information is the entire observed process, with its characteristic shape, rather than a string of numbers in the data matrix, the shift or warping function from one curve to another to align two series of measurements must be estimated before computing the dissimilarity. Figure 2 shows the recordings of the CO concentrations at two close sites in the city of Bologna from 23 January 2001 to 9 February 2001. While the two curves overall have a similar shape, some peaks are not exactly aligned in the time axis. Cross-sectional dissimilarity, which assumes the ith point in one series to be aligned with the ith point in the other series, will produce a pessimistic measure. The nonlinear time warping allows a more intuitive longitudinal distance to be calculated. As an example, some of the salient points aligned by the dynamic time warping (dtw) algorithm (Sakoe and Chiba, 1978; Wang and Gasser, 1997, 1999) are connected with the dotted line in Figure 2.

The time variability problem for air pollutant data can be formalized as follows. At consecutive times (hours, days, months, ...) $j \in t = [1, n_i] \subset N$, measurements x_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ are gathered from a sample of monitoring stations of size m. The assumption that each realization of the underlying process generates a smooth curve then leads to a nonparametric regression model

$$x_{ij} = f_i(t_j) + e_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n_i \tag{1}$$

DOI: 10.1002/env

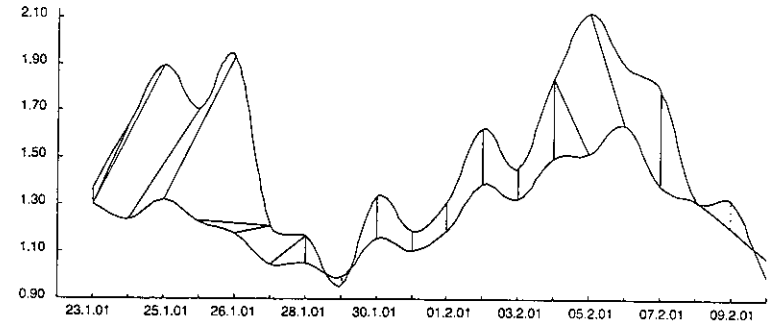

Figure 2. CO curves. The dotted lines show some of the points aligned by dtw

where e_{ij} denotes an unknown zero mean error term, (without loss of generality, since a positive or negative mean can be incorporated in the function f_i) and where f_i are unknown smooth functions. If a parametric model were available *a priori*, subsequent analyses would be simplified. However, at the beginning of data analysis, enough knowledge to build an appropriate parametric model does not usually exist. Moreover, existing models may be seriously deficient in the classical fields of application. For these reasons, nonparametric analyses have been used instead (Kneip and Gasser, 1992; Wang and Gasser, 1999; Gervini and Gasser, 2004). The need for nonparametric models is also more evident for ecological processes, which have great variability in both space and time. Kneip and Gasser (1992) applied a strictly monotone time transformation w_i from a physical to a 'biological' time scale (in a biological application) to eliminate nonlinear shifts between curves (the functions w_i are also called warping or alignment or registration functions). Analyzing subsequently, the $f_i(w_i(t))$ instead of the $f_i(t)$ led to more meaningful results, in particular, to a more meaningful structural average curve, reflecting the average dynamics and the average intensity. The determination of the transformation w_i in this approach is based on features common to the sample curves and is followed by monotone interpolation. The definition of features, also called landmarks, and their unequivocal identification in individual curves might pose problems in ecological data as well as in other applications. It may not also be feasible because the process becomes difficult and time consuming when sampled curves are in great number. As an example, Figure 3 shows daily recordings of CO by two monitoring stations in Bologna, from 1 January 2001 to 31 October 2004. Peaks around the end of December and beginning of January in each year are evident landmarks. However, their exact number and daily locations are not objectively identifiable.

For these reasons, as an alternative to landmark registration, we used an improved dynamic time-warping algorithm, both for computing the dissimilarity between each pair of sampled curves and for estimating the structural mean in a nonparametric way. The paper focuses on the application and shows how pollutant data may be treated and analyzed to achieve meaningful results. The main features of this work are the application of different statistical procedures to the same data set and the improvement of results through specification of particular strategies in data modeling. The improved dynamic time warping (dtw) presents two methodological contributions. The first one consists of a definition of an iterative procedure to estimate cross-sectional mean (Morlini and Zani, 2006). The original dtw only computes the dissimilarity between each pair of curves and cannot be applied to estimate the cross-sectional mean. Wang and Gasser (1999) have presented a method for aligning all curves to their average time scale. However, for this method, a variational problem in continuous time is formulated to obtain smooth shift functions instead of a warping path. In the improved dtw, we use a

DOI: 10.1002/env

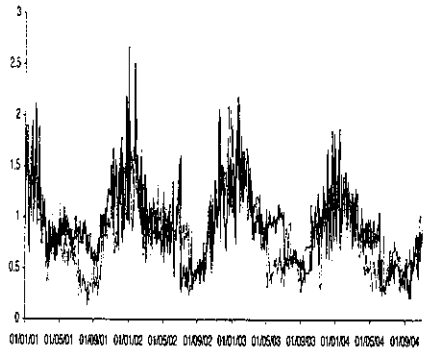

Figure 3. CO curves (daily values from 1.01.01 to 31.10.04)

discrete warping path instead. The second contribution (Morlini, 2004 and Corbellini and Morlini, 2004) is the process of spline smoothing with λ estimated by generalized cross validation, and thus 'optimally' chosen for each series, before applying dtw. Once again, the difference between this method and the other approaches (see, e.g., Wang and Gasser (1997, 1999) is that interpolating smooth functions are estimated to obtain less noisy and shape-dependent data, but the cost and warping functions remain discrete. Therefore, the original dtw is not denaturalized. It still relies on a minimization problem, which can be solved efficiently by using dynamic programming and can be applied in a straightforward manner to a vector-valued series. This is particularly convenient for pollutants because in the atmosphere, several pollutants are present at the same time and the effects on human health due to this simultaneous presence in the air must be considered.

For many other examples of functional data and statistical models for analyzing them, as well as for considerations about the philosophy underlying functional data, we refer to Ramsay and Dalzell (1991) and Ramsay and Silverman (1997). Examples of parametric warping functions are in Silverman (1995) and Ramsay and Li (1998).

The paper is organized as follows. Section 2 describes our way of performing dtw to make the paper self-contained. In this section, the self-modeling warping functions model, which is used as a benchmark, is also introduced and briefly described. Section 3 presents the air pollutant data, describes the pre-processing step, and shows the structural means obtained by improved dtw and by the self-modeling warping functions for each pollutant. Results gained by improved dtw with vector-valued series, that is, by simultaneously considering more than one pollutant, are reported and compared with those obtained with univariate series. Section 4 discusses the cluster analysis results. Section 5 concludes.

2. DYNAMIC TIME WARPING

2.1. *The algorithm*

Dynamic time warping, originally developed in engineering literature for speech analysis and speech recognition (Rabiner and Juang, 1993), has since then found applications in such different fields as data

DOI: 10.1002/env

mining, finance, manufacturing, biology, and medicine. In this section, we briefly describe how the method works.

To align two sequences x_{ij} with $i = 1, 2$ and $j = 1, \ldots, n_i$ and measure their dissimilarity, the dtw algorithm first implies the construction of a $n_1 \times n_2$ square lattice $\mathbf{M}$ in which the generic element (r, c) is the distance $d(x_{1r}, x_{2c})$ between the value of series 1 at time r and the value of series 2 at time c. Note that the series $x_i(t)$ may be vector valued, as would be the case, for example, if they indicate position in two- or three-dimensional space or measurements of different features regarding the same phenomenon. At this stage, any distance may be used in the construction of the square lattice $\mathbf{M}$. Each element (r, c) corresponds to the alignment between points x_{1r} and x_{2c}. The dissimilarity between $x_1(t)$ and $x_2(t)$, also called dynamic time warping cost (dtwc), is defined as follows:

$$\text{dtwc} = \min \sqrt{\frac{\sum_{k=1}^{K} d_k}{K}} \tag{2}$$

where $\max(n_1, n_2) \leq K \leq n_1 + n_2 - 1$, K is determined by the warping process, and the d_ks are elements of $\mathbf{M}$ subject to

- Boundary condition: $d_1 = d(x_{11}, x_{21})$ and $d_K = d(x_{1n1}, x_{2n2})$. This requires the first and the last addends in the sum to be diagonally opposite corner elements of $\mathbf{M}$.
- Contiguity constraint: given $d_k = d(x_{1r}, x_{2c})$ then $d_{k-1} = d(x_{1r'}, x_{2c'})$ where $r - r' \leq 1$ and $c - c' \leq 1$. This condition restricts two successive elements d_k in the sum to be adjacent (including diagonally) elements in $\mathbf{M}$.
- Monotonicity constraint: given $d_k = d(x_{1r}, x_{2c})$ then $d_{k-1} = d(x_{1r'}, x_{2c'})$ where $r - r' \geq 0$ and $c - c' \geq 0$. This forces the points for which distance is considered in the dtwc to be monotonically spaced in time.

While finding this measure of dissimilarity, dtw produces a relative shift between the two sampled curves. However, as shown in Figure 4a, the algorithm defines a warping path and, from this path, we cannot draw two strictly increasing warping functions to align $x_1(t)$ to $x_2(t)$ and to align $x_2(t)$ to $x_1(t)$,

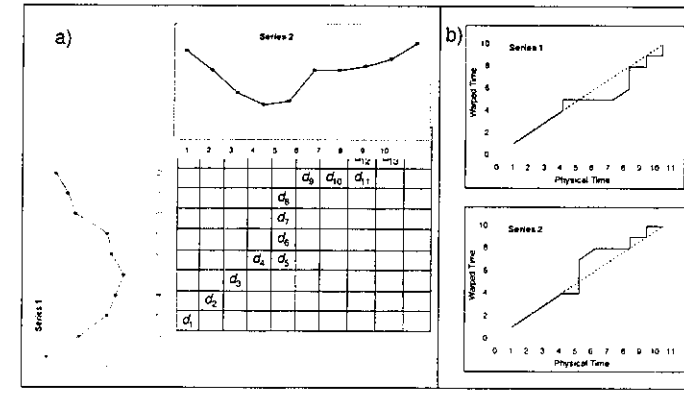

Figure 4. (a) An example of the distances included in the DTWC. (b) The warping path

DOI: 10.1002/env

since a single point on one time series may map onto a large subsection of the other series (Figure 4b). To find two monotonic, not strictly increasing, warping functions, one could eliminate the boundary condition $d_K = d(x_{1n1}, x_{2n2})$ and restrict the contiguity constraint such that $r - r' = 1$ for aligning $x_1(t)$ to $x_2(t)$ and such that $c - c' = 1$ for aligning $x_2(t)$ to $x_1(t)$. However, with this restriction, the dtwc turns asymmetric and cannot be a dissimilarity measure. To define a dissimilarity measure and a warping function at the same time, we used a modified parameterized path, which is not new in literature. In this path, the contiguity constraint is relaxed and $r - r' \leq 2$ and, $c - c' \leq 2$, with the exception of the case of both subtraction equal to 2. Other approaches formulate a variational problem in continuous time to obtain smooth shift functions or introduce new cost functions (e.g., Wang and Gasser, 1997, 1999; Roberts *et al.*, 1987). With these modifications, however, the method does not rely any more on a minimization problem, which can be solved efficiently by using dynamic programming and cannot be straightforwardly applied to multivariate sequences. Maintaining a discrete warping function instead allows the alignments of vector-valued series $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$, simply computing the distances $d(\mathbf{x}_{1r}, \mathbf{x}_{2c})$, for example, the Euclidean norm, in the square lattice $\mathbf{M}$.

2.2. *Some improvements*

As outlined in the Introduction, the values x_{ij} of sampled curves $i = 1, \ldots, m$ at time $j = 1, \ldots, n_i$ may differ because of two types of variability. The first is amplitude variability, seen when the values of two series $x_1(t)$ and $x_2(t)$ may simply differ at points of time at which they can be compared. The second is dynamic variability, which is exhibited when $x_1(t)$ and $x_2(t)$ should not be compared at fixed values of t but at times r and c at which the two values are essentially in a comparable state. Due to its flexibility, dtw may lead to overwarping when explaining amplitude variability between sampled curves in terms of dynamic variability. This is especially true for noisy data, which exhibit many extreme values. To reduce noise in the data and to obtain values that also depend on the overall shape of the series, we obtained smoothed estimates by interpolating a smoothing spline for each series. The process of spline smoothing employs a roughness penalty approach to establish a trade-off between regularity of the interpolating spline and goodness of fit, measured by the Euclidean distance between raw data and smoothed estimates. In brief, given the series $x_i(t)$, let $S(g_i)$ be the penalized sum of squares:

$$S(g_i) = \sum_{j=1}^{n_i} (x_{ij} - g_i(t_j))^2 + \lambda \int \{g_i''(k)\}^2 dk$$

with λ being the positive smoothing parameter controlling the trade-off between regularity and goodness of fit. The interpolating spline $\hat{g}_i$ is the minimizer of the functional $S(g_i)$ over the class of all twice differentiable functions g. The parameter λ can be set equal for all curves and fixed *a priori*. However, the degree of smoothing should be reasonably data-dependent and should be different for each curve. For these reasons, for estimating λ, we have implemented in Matlab the generalized cross validation (GCV) criterion, which automatically selects λ from the data (Corbellini and Morlini, 2004). Given the following score function,

$$\mathrm{GCV}(\lambda) = \frac{1}{n_i} \frac{\sum_{j=1}^{n_i} (x_{ij} - \hat{g}_i(t_j))^2}{\{1 - n_i^{-1} tr\mathbf{A}(\lambda)\}^2}$$

where n_i is the length of the series $x_i(t)$, $\mathbf{A}(\lambda) = (\mathbf{I} + \lambda \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T)^{-1}$, $\mathbf{Q}$ and $\mathbf{R}$ are two band matrices, the GCV choice of λ is carried out by minimizing the function $\mathrm{GCV}(\lambda)$ over λ. $\mathbf{Q}$ has size $n_i \times (n_i - 2)$ with

entries Q_{rc}, for $r = 1, \ldots, n_i$ and $c = 2, \ldots, (n_i - 1)$, where

$$Q_{c-1,c} = Q_{c+1,c} = (\Delta t)^{-1} \text{ and } Q_{c,c} = -2(\Delta t)^{-1}$$

with $Q_{rc} = 0$ for $|r - c| \geq 2$. The $(n_i - 2) \times (n_i - 2)$ matrix $\mathbf{R}$ is given by

$$\mathbf{R}_{rr} = \tfrac{2\Delta t}{3}, \text{ for } i = 2,\ n_i - 1$$
$$\mathbf{R}_{r,r+1} = \mathbf{R}_{r+1,r} = \tfrac{\Delta t}{6}, \text{ for } i = 2, \ldots, n_i - 2$$
$$\mathbf{R}_{rc} = 0 \text{ for } |r - c| \geq 2$$

The advantage of employing GCV over cross validation (see Green and Silverman, 1994, for a complete description of these criteria) is of computational reason, since it is possible to find the trace of $A(\lambda)$ without finding all its diagonal elements.

Another strategy to prevent time variability overfitting is to put a boundary on the time axis fluctuations to avoid excessive differences in timing and unrealistic warping. A windowing condition $|r - c| \leq u$, with u being a given positive integer, may be posted on the warping path. For example, in Figure 5, u is set equal to 5. Regarding the daily air pollutant measurements (see the next section), the positive integer u is the maximum number of days for which we assume the same weather factors influencing air pollution may be timed differently for the different spatial locations of the monitoring gauges. In general, this value should be kept according to the series' time scale and to the maximum temporal distance for which features in different subjects may be logically compared.

Regarding the estimation of the structural mean, this analysis implies the alignment of a sample of curves to some average time scale. Model (1) becomes

$$x_{ij} = a_i \mu(v_i(t_j) + e_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n_i \tag{3}$$

where μ is the structural mean, a_i is the amplitude variability, and $v_i(t)$ is the inverse of the warping function for the ith series. The estimation of μ clearly requires a global fitting criterion with respect to all the sampled curves. As outlined previously, dtw produces a relative shift function between two series instead and minimizes the dissimilarity measure (2) between each pair of sequences. However, if we define a reference vector, then all sequences can be aligned to this reference vector by dtw. Hence, the structural mean can be computed. The problem concerning the reference vector may be solved with

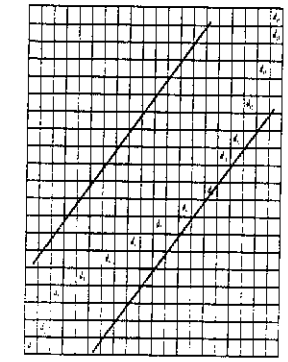

Figure 5. Pictorial representation of the windowing condition with $u = 5$

several possibilities (Wang and Gasser, 1997). In general, the reference vector should be close to the typical pattern of the sample curves and should have more or less the same features as most curves. In choosing this vector, a trade off between accuracy and computational effort should be considered. Here, we propose the iterative method that first considers the longitudinal mean as reference vector, warp every series to this vector, compute the total warping cost (that is, the sum of the dtwc between each series and the structural mean), and consider this structural mean as reference vector for the second step. The process is iterated until the total warping cost has negligible changes. Computation is not intensive since few iterations are usually enough. However, if the relative shifts among curves are large, then the cross-sectional mean might be too atypical to start with, since the structure gets lost. A more convenient method, in this case, is to take each vector $x_i(t)$ as reference, to warp every other series to this vector, compute the total warping cost, and choose the reference vector as the one corresponding to the maximum total cost.

2.3. *Self-modeling warping functions*

As a benchmark for the improved dtw algorithm described in subsections 2.1 and 2.2 and implemented in Matlab, for air pollutant data, we also applied the self-modeling warping function approach (Gervini and Gasser, 2004). We used the Matlab routines available on Web page http://www.uwm.edu./~gervini. This model is comparable with the improved dtw for computing the structural mean due to its flexibility and the peculiarities of being model-free and iterative. In this model, the warping functions are assumed to be linear combinations of q common components, which are estimated from the data (hence the term 'self-modeling'). Formally,

$$w_i(t) = t + \sum_{j=1}^{q} s_{ij}\phi_j(t)$$

where $w_i(t) = v_i^{-1}(t)$, ϕ_j are the component functions and s_{ij} are the component scores. Even small values of q provide remarkable model flexibility, comparable with nonparametric methods. At the same time, the approach is less prone to overfitting than parametric methods, like continuous monotone registration, because the common components are estimated by combining data across curves. Each component is a linear combination of p B-spline basis functions and accounts for time variability at different segments of the time axis. Hence, each component can be thought to be associated with a hidden landmark. Gervini and Gasser (2004) gave a rationale for this. This method is probably the best alternative to landmark registration when individual identification of landmarks is not feasible or cannot be accurate. With regard to model fitting, criteria minimized by the two algorithms are not comparable. Assuming the same length for all curves, that is $n_i = n$ for every $i = 1, \ldots, m$, self-modeling warping functions minimize the following average squared error:

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[x_{ij} - a_i\mu(v_i(t_j))\right]^2(t_j^* - t_{j-1}^*)$$

where $t_0^* = t_1$, $t_j^* = (t_j + t_{j-1})/2$ for $j = 2, \ldots, n-1$, $t_n^* = t_n$, which directly derives from model (3). Improved dtw minimizes the total dtwc, that is, the sum of dissimilarities (2) computed between each series and the estimated structural mean. The total dtwc can be written as follows:

$$\sum_{i=1}^{m}\inf_{w}\sum_{(r,c)\in w}(x_{ir} - \mu_c)^2$$

where w is the warping path connecting (1, 1) and (n, n) in the two-dimensional square lattice **M**. Since the two objective functions are not comparable, confronts should be better drawn based on the shapes of the estimated structural means, on the timing of salient features, and on the amplitudes of peaks and local minima.

3. STRUCTURAL MEANS OF AIR POLLUTANT DATA

3.1. *The data set and the pre-processing step*

We performed a structural analysis of air pollutant measurements gathered by a network of 67 fixed stations in the urban areas of the Emilia Romagna region. Data were provided by the Regional Agency for Environmental Prevention in Emilia Romagna (ARPA: www.arpa.emr.it).

Pollutants considered are CO, ozone (O_3), benzene, sulfur dioxide (SO_2), particulate matter (PM_{10}), and coarse dust particles (CDP). In the data set, we have 1400 daily measurements (from the 1 January 2001 to 31 October 2004) obtained by the 67 air quality stations for each pollutant. For PM_{10}, benzene, SO_2, and CDP, these measurements were simply 24-h average concentrations; for O_3, these were maximum hourly values; for CO, these were maximum 8-h average concentrations. A pre-processing step is done to replace missing values with a moving average of length 5. Also, unreliable values (extreme points, as shown by boxplots, inconsistent with measurements gathered in the two closest days) were replaced with moving averages of the same length. Previous analyses of these data are reported in Morlini & Zani (2006).

3.2. *Structural means*

Figure 6 shows structural means obtained with improved dtw (left panels), with a 12-day maximum time shift and with self-modeling registration (right panels), with $q = 4$ components of $p = 8$ B-spline basis functions of order 3 and equally spaced knots. The choice of the parameters, in both models, was somehow subjective. However, a windowing condition of about 10 day seemed reasonable and results were very much stable with respect to a different value of u ranging from 7 to 15. The same reasoning may be given for the parameter $q = 4$. Four hidden landmarks seemed acceptable for the series at hand, which, in general, have at least one major peak in each year. A value less than 4 seemed insufficient. With $q = 6$ and $p = 12$ or $p = 18$ results were similar. For more than six hidden landmarks, the parameters to be estimated were too numerous for the number of sampled curves. The peculiarity of these data, with respect to functional data sets usually used in literature, is that a great number of observations per curve are available, while the number of curves is quite limited, especially if the study area is small. For all the pollutants, the estimated trends are similar and do not show significant differences. However, structural means estimated by self-modeling warping functions were more wiggled and peaks and local minima were wider, probably due to the greater flexibility of the model (the windowing condition posed a significant constraint on the flexibility of the warping functions estimated by dtw). Extreme values, however, had exactly the same timing or at least a difference of 1 day. The Euclidean and the city block distances between the structural means estimated by dtw and the structural means estimated by self-modeling warping functions are reported in Table I. Among others, the following considerations are immediately evident from both graphs:

- CO clearly followed a cyclic pattern with peaks in December and January and minima during June and July. While local minima were stable in the 3 years, the amplitude of the winter peaks decreased from 2002 to 2004.

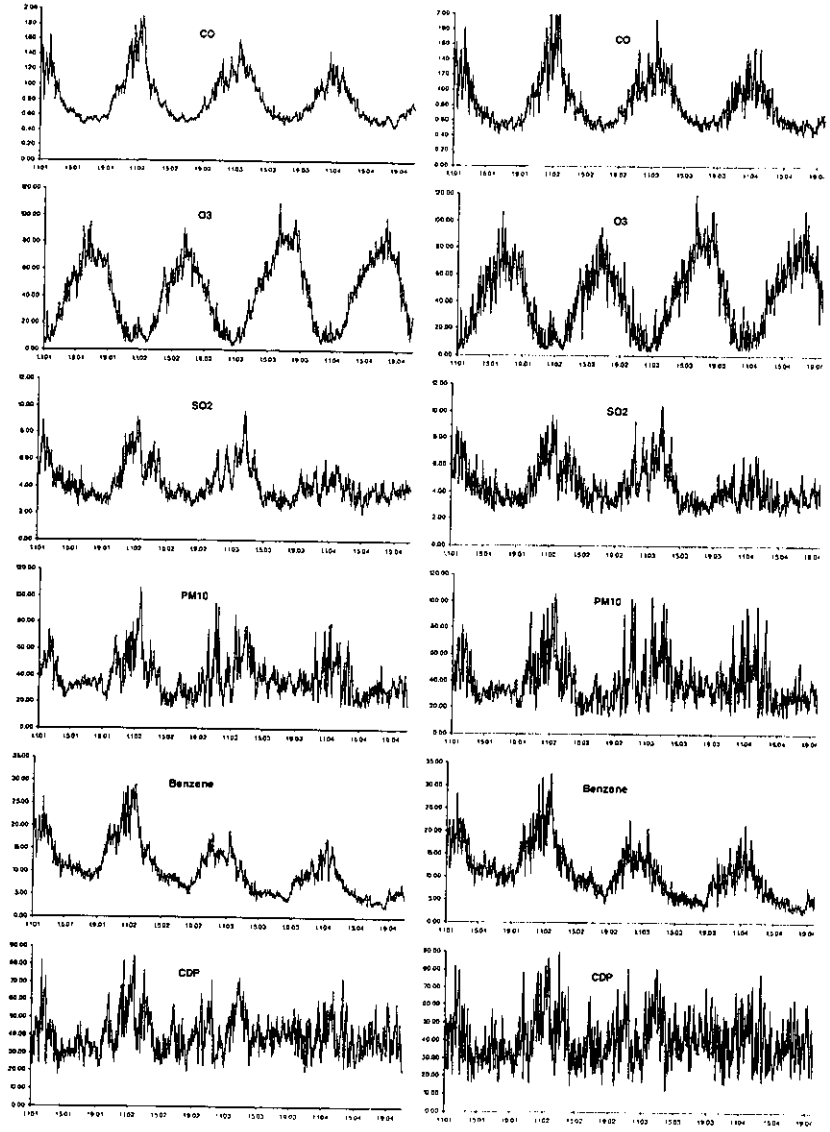

Figure 6. Structural means. On the left panels obtained with improved dtw, on the right panels with self modeling warping functions

Table 1. Distances between structural means estimated by improved dtw and self modeling warping functions (observations are 1400)

Pollutants	Euclidean distance	City block distance
CO (mg m^3)	3	79
O_3 (μg m^3)	179	4920
SO_2 (μg m^3)	22	643
PM_{10} (μg m^3)	220	5832
Benzene (μg m^3)	63	1707
CDP (μg m^3)	239	6685

- SO_2 and benzene followed the same cyclic pattern but more irregularly and with more fluctuations. For benzene, both peaks in winter and local minima in August were decreasing during the 3 years. For SO_2, the minima remained stable, while peaks were higher in 2003 with respect to 2002, but these were drastically reduced in 2004.
- O_3 had a cyclic pattern with peaks in June and minima in December and January. The amplitude of peaks increased during the study years.
- PM_{10} and CDP did not follow any cyclic pattern. Maxima values, however, were always reached in the winter months. Some peaks were timed exactly the same days as for benzene. Major annual peaks were decreasing from 2002 to 2004.

The self-modeling warping functions approach allowed ulterior knowledge about salient events. Indeed, representing the self-modeling components, we may individualize the landmarks they are associated with. As an example, Figure 7 shows the estimated components and the boxplots of the component scores for CO and O_3. From Figure 7a, we see that marker events for CO were minimum around 13 June 2001 (first component) and the peaks were around 21 December 2001 (second component), 16 December 2002 (third component), and 14 January 2004 (fourth component). The same amplitude as well as the same timing for these landmarks had also been identified by improved dtw. Figure 7e shows (by an arrow) each of these events on the graph of the structural mean estimated by dtw. Since the component scores are just deviations of individual landmarks from average landmarks, extreme values individualized by boxplots of Figure 7c show the monitoring stations for which landmark events occur earlier or later. For example, for stations identified by numbers 4 and 15, the minimum around middle of June 2001 (identified by the first component) occurred earlier (and much earlier for station 4), whereas for station 50, it occurred later. For the second marker event, the peak was around 21 December 2001, the average timing was near all individual timings, since the box of the second boxplot was quite narrow. Only for station 20 did it occur significantly later. Analogous considerations held for the fourth marker event, even if there were three sites for which this event occurred significantly earlier and one for which it occurred much later. For the third self-modeling landmark, the score distribution was not symmetrical and this reveals that the peak around 16 December 2002 occurred later for a majority of the stations.

For O_3, marker events were the peak around 23 June 2001 (first component), the local maximum around 5 August 2002 (second component), the minimum around 23 October 2003 (third component), and the local maximum around 5 June 2004 (Figure 7b and f). As for CO, these marker events were also present in the structural mean estimated by improved dtw. The sites for which they occurred significantly earlier or later were detected by boxplots of Figure 7d. The length of the box of each plot reveals the variability of the individual landmarks around the landmark average time. For the station identified by number 20, two peaks for CO and one local maximum for O_3 occurred later with respect to

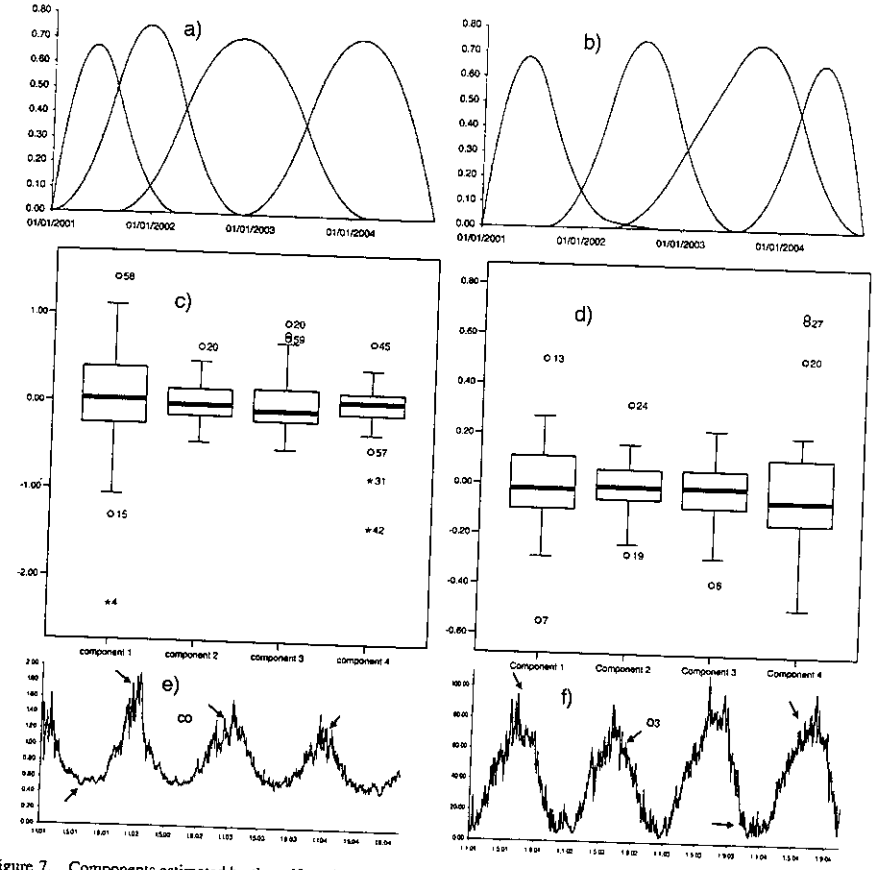

Figure 7. Components estimated by the self modeling warping functions for (a) CO and (b) O_3. Boxplots of the scores of each component for (c) CO and (d) O_3. Marker events associated with each components for (e) CO and (f) O_3

the estimated average time. Since this behavior was present also for some landmarks in other pollutants, this monitoring station may be characterized as a site for which the pollutant curves are shifted on the right side of the time axis (marker events occur later, in general, with respect to other stations). Note that landmarks are not all major peaks or valleys but may be also associated with local maxima or local minima. This is natural since the first and second derivatives for these points are the same. Furthermore, the component functions must be localized and account for time variability at different segments of t.

Figure 6 shows structural means with warping functions estimated separately for each pollutant. However, the pollutants are present simultaneously in the atmosphere and we may consider all of them to estimate a common warping function. This can be done by considering vector-valued series in the improved dtw. Figure 8, on the right-side panels, reveals the structural mean of the daily values of

DOI: 10.1002/env

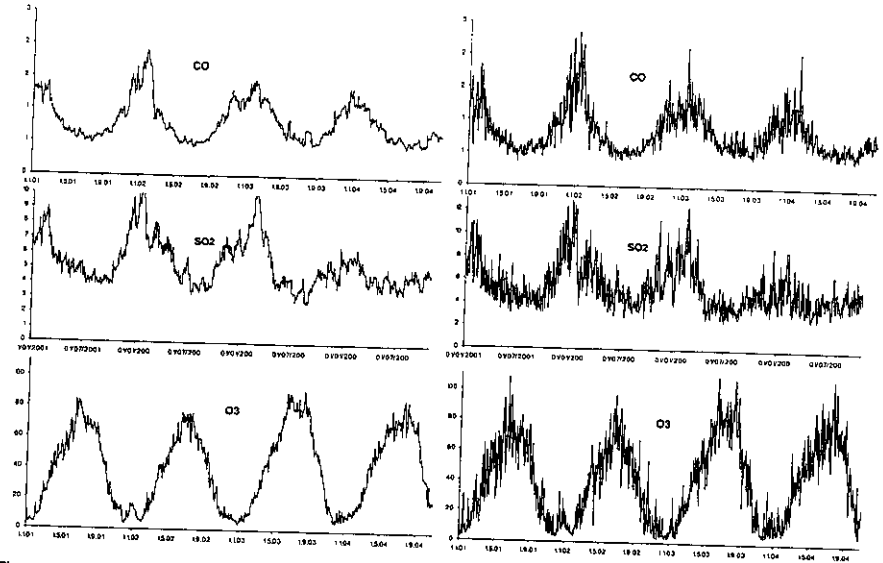

Figure 8. Estimated means of CO, SO_2, and O_3 (from the 1 January 2001 to 31 October 2004). Left panels report structural means estimated with vector-valued series. Right panels report longitudinal means

CO, SO_2, and O_3 obtained by considering multivariate series in the algorithm. The warping function of each series is equal for each pollutant and is estimated considering simultaneously the values of the three pollutants, which are thought as different aspects of the same phenomenon.

To reduce noise, the raw series were interpolated with tensor product splines. Here, the parameter λ was equal to 0.02 and was chosen *a priori*, since a cross-validated choice should have been too computationally intensive. Sequences were not standardized and hence the different ranges and units of the three pollutants had an influence on the results. The right-side panels of Figure 8 show longitudinal means, computed without warping. The amplitudes of peaks were similar but structural means showed less and better defined peaks. The maxima of SO_2 and CO (which are typically 'winter' pollutant) and the minima of O_3 (which is a 'summer' pollutant) were well aligned and the trend for each pollutant was also much more evident in the structural means. Comparing the graphs of structural means obtained with vector-valued series (Figure 8 left panels) with those obtained with univariate analyses (Figure 6) we see that in the first case, marker events were estimated more accurately and spurts were better defined. Although we are mainly concerned with the structural mean estimation, a potential use of improved dtw with vector-valued series may be the development of a daily pollution index, based on the values of all pollutants taken simultaneously. For pollutants for which the structural mean is computed with common warping functions, this index may be reached with a principal component analysis over the obtained structural means. For example, the first principal component of the structural mean of CO, SO_2, and O_3 had eigenvalues 2.56 and explained the 85.4% of total variability. The second and the third components had, respectively, eigenvalues 0.33 and 0.11 and explained, respectively, the 11% and the 3.6% of the total variability. Since the first principal component had significant correlations with all three structural means (Table 2), it may then be used as a quality index based on the

DOI: 10.1002/env

Table 2. Linear correlations between the first principal component and the structural means

Variable	Component 1
Structural mean CO	0.96
Structural mean SO_2	−0.91
Structural mean O_3	0.90

three pollutants CO, SO_2, and O_3. Breakpoints for this index may be defined, considering that scores have zero mean and one of the three structural means has a negative correlation with the principal component. As an example, Table 3 reports possible breakpoints for the pollution index based on the deciles of the standardized normal distribution, and frequency of days corresponding to each class. Note that breakpoints are considered with respect to the normal distribution and do not focus on health effects that one may experience within a few hours or days after breathing polluted air. As a matter of fact, a pollution index based on the first principal component of the structural means reached by improved dtw with vector-valued series may overcome the two main drawbacks of official air quality indexes (e.g., the AQI proposed by the Environmental Protection Agency (1999)). The first one is that, it is not realistic to assume the same air pollution index to be valid all over the world, since different areas are characterized by different climatic conditions, and both the construction of the index and the breakpoints should be data-dependent. The second drawback is that AQI refers to a single pollutant, whereas the level of pollution should be considered with respect to the different pollutants simultaneously present in the atmosphere. Indeed, the AQI of a site, where more than one pollutant is monitored, is simply the maximum AQI value among those obtained for each pollutant. Data reported in Tables 2 and 3 illustrate the potential use of a principal component analysis over the structural means obtained with multivariate series. However, the topic deserves further elaboration since the analysis should be performed over a wide range of pollutants.

4. CLUSTER ANALYSIS RESULTS

The dtwc computed for each pair of sequences may be used for classifying the monitoring stations. To reach clearer graphical representations, we have considered for this analysis only the 19 traffic area sites out of the 67 monitoring stations. Figure 9 reports dendrograms obtained for the CO series with the complete and the average linkage (see, e.g., Gordon, 1991 and 1996 for a complete description of hierarchical clustering methods). As a dissimilarity measure, we have considered both the Euclidean distance, which does not allow for time variation, and the dtwc, which allows time warping.

Table 3. Possible breakpoints for the pollution index proposed

Score of principal component I		Number of days
0–|0.25|	Good quality	169
|0.25|–|0.52|	Low pollution	214
|0.52|–|0.84|	Moderate pollution	299
|0-84|–|1.28|	High pollution	474
>|1.28|	Very high pollution	244

Breakpoints are based on the deciles of the standardized normal distribution.

DOI: 10.1002/env

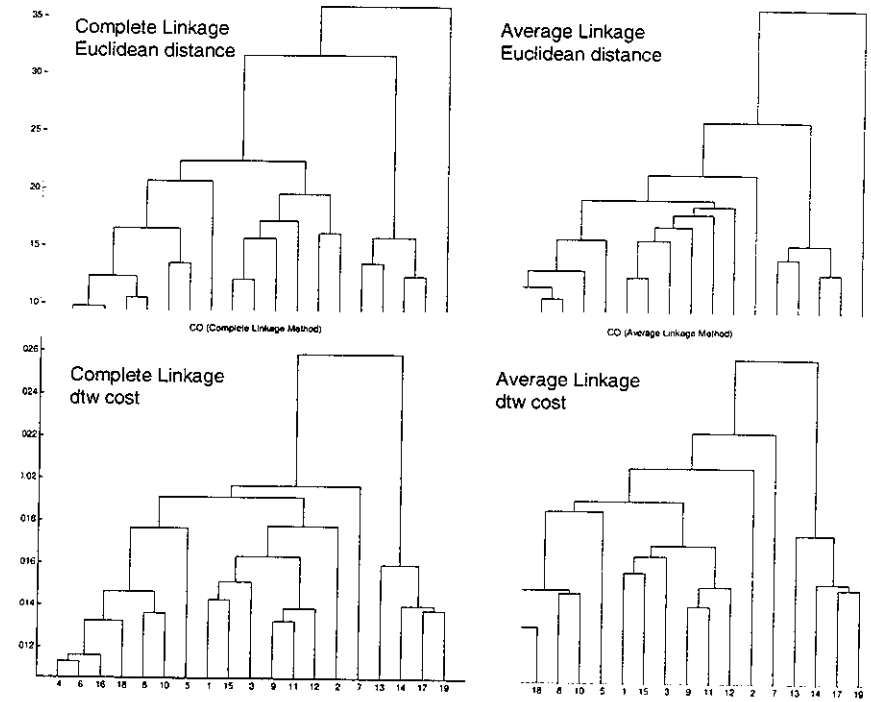

Figure 9. Dendrograms obtained with the CO series

Comparing dendrograms, we see that with dtwc, all possible partitions obtained with complete linkage were identical to those obtained with the average and the complete linkage. On the contrary, some partitions differed in the two dendrograms obtained with Euclidean distance. This gives an insight into the stability, with respect to the linkage used in the analysis, of the classification reached by dtwc.

Another observation was with respect to site number 7. Based on the Euclidean distance, we should consider this site as an outlier, very dissimilar from the other 18 monitoring stations. However, if we analyze the graphs of daily measurements, we see that this series has a similar trend as the other series, with relatively low peaks in winter months (especially in January 2001 and January 2002) but some spurts in summer months (especially in July 2003, September 2003, and September 2004). A totally different behavior was noted in sites 13, 14, 17, and 19, with high peaks in winter (especially in 2001 and 2002) and low concentrations in the summer months. Since the summer peaks of site 7 were not well aligned with those present, with less intensity, in other series with a similar trend, the distances between these series were amplified without time warping. On the contrary, with aligned peaks in the summer months, site 7 still remained isolated until the upper levels of the dendrograms, but it was first aggregated with sites having similar annual behavior (e.g., sites 9, 11, 12, and 2) and then with the cluster of monitoring stations without summer peaks and high winter concentrations. The dendrograms for the SO_2 and the O_3 series, obtained with the dtwc and the complete linkage agglomerative method,

DOI: 10.1002/env

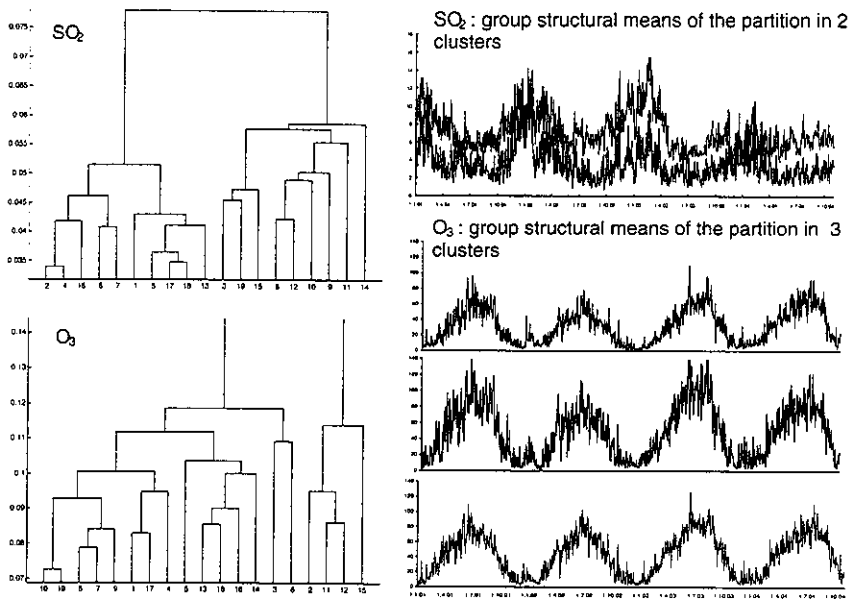

Figure 10. On the left panels: dendrograms obtained with the SO_2 series and with the O_3 series. On the right panels: structural means of groups for the partition in 2 clusters (SO_2) and for the partition in 3 clusters (O_3)

are reported in Figure 10. Dendrograms obtained with other linkages were indeed similar. Hence, the stability of partitions using dtwc was verified also for these data. For SO_2, the presence of two main clusters was evident. The structural means of these groups were reported in the right-side panel and showed that stations with low daily concentrations (sites 2, 4, 16, 6, 7, 1, 5, 17, 18, and 13) were separated from those with high daily concentrations (sites 3, 19, 15, 8, 12, 10, 9, 11, and 14). The pollution trend was similar for both groups, but the level was remarkably different.

For O_3, we represented the group structural means regarding the partition in three clusters. Here, the graphs are separated to reach a clearer representation. The first graph reports the structural means of sites 2, 11, 12, and 15. This group had lower level of O_3 pollution, while stations 3 and 8 (second graph) showed the highest levels of pollution. The other sites, grouped together, showed intermediate mean pollution values. While some sites showed relatively high or low concentrations both for SO_2 and O_3, the other sites belonged to the group of stations with high pollution for O_3 and to the cluster of sites with low pollution for SO_2 or vice versa. For example, sites 11, 12, and 15 were classified in the group with the best quality air, considering O_3, and in the high pollution group, considering SO_2. This different behavior of the pollutants in the different sites show the need for a multivariate analysis to discriminate between 'high' and 'low' air-polluted stations. This can be reached by using dtwc obtained with a vector-valued series. As an example, Figure 11 shows the dendrogram obtained with the complete linkage method and dtwc computed with the series of CO, SO_2, and O_3. The dendrogram was different from all three dendrograms obtained with univariate series (see Figures 9 and 10). Analyzing the

DOI: 10.1002/env

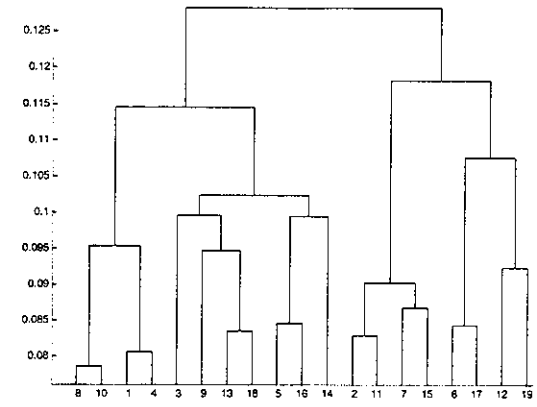

Figure 11. Dendrogram obtained with the complete linkage and the dtwc computed considering simultaneously the CO, SO_2, and O_3 series

aggregation steps, we may individualize sites in which the pollution follows a similar trend. For example, sites 8 and 10 showed a very similar pollution of CO, SO_2, and O_3 in the study period. The partition in four groups individualizes areas with relatively high pollution both in the summer and winter months (sites 8, 10, 1, and 4); areas with moderate pollution yearly (sites 3, 9, 13, 18, 5, 16, and 14); areas with moderate pollution in winter and low pollution in summer (sites 2, 11, 7, and 15); and areas with low pollution yearly (sites 6, 17, 12, and 19). Note that the level of air pollution for each group is defined in relative terms and does not consider the health effects one may experience within a few hours or days after breathing polluted air.

The study in this section is an example of how to synthesize measurements of pollutant concentration data. The approach may be applied to any study area and with more than three pollutants, if data are available. Information carried out by the methodology proposed regarding the level of dissimilarity between sites allows classification in homogeneous clusters. By analyzing the cluster means, the group may then be characterized and ranked, and public support warnings and emission reduction measures may then be handled differently for sites belonging to different groups.

5. CONCLUSIONS

Air quality monitoring data are more copious as more monitoring networks are installed and implemented in urban areas. Several pollutants are generally monitored by each station and hourly or daily average concentrations are calculated. Hence, a large mass of data are provided and their interpretation by public authorities and their synthesis become a hard task. Interpretation and synthesis may be reached by computing the average series of the pollutants. A decision support for public authorities may also be reached by computing dissimilarity distances between sites and by grouping monitoring stations of the study area in homogeneous clusters. This work shows how pollutant data may be treated and analyzed to obtain meaningful results. The methodologies illustrated considered

DOI: 10.1002/env

time variability beyond the well-known range variability and treated series of measurements of pollutants as functional data. The improved dtw algorithm proposed in this paper may also be applied to multivariate series to obtain structural means and partitions of the monitoring stations by considering simultaneously the values of several pollutants. The need for a multivariate analysis, both for clustering different sites and computing structural means, is also illustrated by means of a real data set. The potential use of improved dtw with a vector-valued series is shown for the development of a daily pollution index. The index provided is based on the values of several pollutants taken simultaneously, and this index considers comprehensive pollution emission. It may be useful for comparing, classifying, or ranking different cities or districts and for assessing air quality management. Finally, as boxplots of Figure 7 show, the results reached by the self-modeling warping function approach in the field of air pollution made a meaningful contribution to air quality modeling and in identifying the temporal structure of the pollutant emissions.

REFERENCES

Corbellini A, Morlini I. 2004. Searching for the optimal smoothing before applying Dynamic Time Warping. *Proceedings of the XLII Meeting of the Italian Statistical Society*, University of Bari: 47–50.

Environmental Protection Agency. 1999. Guideline for reporting of daily air quality – air quality index (AQI). EPA-454/R-99-010. Office of air quality planning and standards, Research triangle park, NC 27711.

Gervini D, Gasser T. 2004. Self-modelling warping functions. *Journal of the Royal Statistical Society, Series B* **66**: 959–971.

Gordon AD. 1991. *Classification 2nd edition*. Chapman and Hall: London.

Gordon AD. 1996. Hierarchical classification. In *Clustering and Classification*, Arabie P, Hubert LJ, De Soete G (eds). World Scientific Publ.: River Edge, NJ.

Green PJ, Silverman BW. 1994. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall: London.

Kneip A, Gasser T. 1992. Statistical Tools to analyze data representing a sample of curves. *Annals of Statistics* **20**(3): 1266–1305.

Morlini, I. 2004. On the dynamic time warping for computing the dissimilarities between curves. In *New Developments in Classification and Data Analysis*, Vichi M, *et al.* (ed.). Springer-Verlag: Berlin; 63–70.

Morlini I, Zani S. 2006. Estimation of the structural man of a sample of curves by dynamic time warping. In *Data Analysis, Classification and the Forward Search*, Zani S, *et al.* (ed.). Springer-Verlag: Berlin; 39–48.

Rabiner L, Juang B. 1993. *Fundamentals of Speech Recognition*. Prentice Hall: N.J.

Ramsay JO, Dalzell C. 1991. Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society* **53**: 539–572.

Ramsay JO, Li X. 1998. Curve Registration. *Journal of the Royal Statistical Society; Series B* **60**: 351–363.

Ramsay JO, Silverman B. 1997. *Functional Data Analysis*. Springer: New York.

Roberts K, Lawrence P, Eisen A, Hoirch M. 1987. Enhancement and dynamic time warping of somatosensory evoked potential components applied to patients with multiple sclerosis. *IEEE Transactions on Biomedical Engineering BME* **34**: 397–405.

Sakoe H, Chiba S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustic, Speech and Signal Processing* **26**: 43–49.

Silverman B. 1995. Incorporating parametric effects into functional principal components analysis. *Journal of Royal Statistical Society, Series B* **57**: 673–689.

Wang K, Gasser T. 1997. Alignment of curves by dynamic time warping. *The Annals of Statistics* **25**(3): 1251–1276.

Wang K, Gasser T. 1999. Synchronizing sample curves nonparametrically. *The Annals of Statistics* **27**(2): 439–460.