Methodology and Computing in Applied Probability (2020) 22:1457 –1479
https://doi.org/10.1007/s11009-018-9661-0

Dynamic Non-parametric Monitoring of Air-Pollution

Sotiris Bersimis[1] · **Kostas Triantafyllopoulos**[2]

Received: 9 October 2017 / Revised: 10 June 2018 /
Accepted: 12 August 2018 / Published online: 6 September 2018
© The Author(s) 2018

Abstract

Air pollution poses a major problem in modern cities, as it has a significant effect in poor quality of life of the general population. Many recent studies link excess levels of major air pollutants with health-related incidents, in particular respiratory-related diseases. This introduces the need for city pollution on-line monitoring to enable quick identification of deviations from “normal” pollution levels, and providing useful information to public authorities for public protection. This article considers dynamic monitoring of pollution data (output of multivariate processes) using Kalman filters and multivariate statistical process control techniques. A state space model is used to define the in-control process dynamics, involving trend and seasonality. Distribution-free monitoring of the residuals of that model is proposed, based on binomial-type and generalised binomial-type statistics as well as on rank statistics. We discuss the general problem of detecting a change in pollutant levels that affects either the entire city (globally) or specific sub-areas (locally). The proposed methodology is illustrated using data, consisting of ozone, nitrogen oxides and sulfur dioxide collected over the air-quality monitoring network of Athens.

Keywords Multivariate statistical process control · Time series monitoring ·
Air surveillance · Air pollution · Non-parametric control chart ·
Generalised binomial-type statistics · Markov chain embedded variables

Mathematics Subject Classification (2010) Primary: 62H12; Secondary: 62M10, 62P12

✉ Sotiris Bersimis
sbersim@unipi.gr

Kostas Triantafyllopoulos
kostas@sheffield.ac.uk

[1] Department of Statistics and Insurance Science, University of Piraeus, 80, Karaoli and Dimitriou Street, 185 34 Piraeus, Greece

[2] School of Mathematics and Statistics, Hicks Building, University of Sheffield, S3 7RH, Sheffield, UK

1 Introduction

In recent years, Statistical Process Control (SPC) has been proposed in environmental related monitoring problems (Pan and Chen 2008; Paroissin et al. 2016). Such problems typically involve processes in the presence of autocorrelation, and relevant SPC techniques used in conjunction with appropriate time-series models (Pan and Chen 2008; Triantafyllopoulos and Bersimis 2016). Over the recent years there has been a significant development of non-parametric monitoring methodology in SPC, as evidenced in Chakraborti et al. (2004), Qiu and Li (2011), Qiu (2018) and in references therein. Distribution-free statistical procedures for process-monitoring are favoured to parametric-based methods, as they relax the distributional assumption of the observed data. However, such non-parametric procedures usually focus on univariate i.i.d. processes. As it is well known in non-industrial processes several variables are often observed and exhibit significant autocorrelation, i.e. the observed process is a multivariate time series. Examples of such process include, but are not limited to, environmental and financial processes.

Air pollution consists of the introduction of chemicals, particulate matter and biological materials into the atmosphere, causing severe damage to the environment (Christodoulakis et al. 2017). The major air pollutants are the sulfur dioxide (SO_2), the nitric oxide (NO), also known as nitrogen monoxide, the nitrogen dioxide (NO_2), the carbon monoxide (CO) and the ozone (O_3). Recent research related to air pollution is focused on public health (Jiang et al. 2016; Raaschou-Nielsen et al. 2016). In the last twenty years, many epidemiological and medical studies, by showing the relation of air pollution to a series of serious and many times emergency health problems, have concluded that public health (especially in urban areas and large cities) is threatened (Mudway and Kelly 2000; Atkinson et al. 2001). As a result a continually monitoring mechanism is needed, to monitor jointly the levels of all major pollutants in a city, taking into account the contribution of climate related variables. This need is stressed by the United States Environmental Protection Agency, which launched in 2009 the Environmental Tracking Network in order to monitor air-quality in many locations in the United States. Furthermore, this mechanism should be able, to account for local (temporal) differences or even spatial variations of the levels of all major pollutants, realising in such a way an evidence-based monitoring system. The value of such a monitoring system lies in its capability of giving early alarms before high levels of pollution are realised and thus having a predictive nature. Multivariate statistical process (MSPC) methods is an ideal vehicle, in order to develop and implement such an automatic monitoring system. Over the last decade MSPC methods have become very popular for monitoring multivariate processes, see e.g. the review of Bersimis et al. (2007).

In this article we develop a statistical framework for automatic monitoring of air-pollution levels, which enable real-time detection of aberrant pollution and issuing warnings, in order to permit the Environmental Public Administration to activate safeguarding mechanisms early in time. We propose a dynamic linear model (Prado and West 2010; Petris et al. 2010) that defines the “normal” or “expected” evolution of the process and deviations from this is measured using residuals. An unusual positive residual indicates aberrant pollution levels beyond the expected, taking into account climate changes,as we are interested in pollution which exceeds the expected levels, extreme negative residuals do not count as aberrant. We then propose distribution-free monitoring of such residuals, based on generalised binomial-type statistics as well as on rank statistics. The novelty of the proposed methodology is the real-time (at daily frequency) automatic monitoring of pollution levels and the detection of sudden shifts, which are linked to respiratory conditions and deaths (Rosenlund et al. 2008; O'Neill and Ebi 2009; Jiang et al. 2016). Indeed, there is an established link between such sudden shifts of pollution, which are not captured by the recommended thresholds issued by the environmental agencies and are based on long-term pollution dynamics rather than short-term temporal variation, which is considered in this paper.

The rest of the paper is organised as follows. Section 2 describes the data and provides some related background. Section 3 discusses the proposed methodology, consisting of inference of a multivariate time series model as well as of monitoring procedures applied to the residuals of that model. The framework proposed in this paper is applied in the available real data in Section 4. Finally, in Section 5, some concluding comments are given, while technical arguments are included in the Appendix.

2 Data Description

Athens is a city of almost 4 million people (census of March 2001) located in an oblong basin of approximately 450 km^2 area. It is surrounded by mountains and the Saronic Gulf. The data consists of 2922 mean measurements recorded in daily frequency, from January 1, 2001 to December 31, 2008 in Athens, using the air-quality monitoring network (AQMN) of the city, consisting of 17 active stations (locations) placed in urban and peri-urban sites around the city. In this study we used 13 out of the 17 active stations of Athens AQMN (4 stations were not active during the study period, depicted in Fig. 1 by the unnumbered squares). Each station consists of a number of sensors, dedicated to the acquisition of the main pollutants: carbon monoxide CO, nitrogen oxides NO and NO_2, sulfur dioxide SO_2, and ozone O_3.

Figure 2 plots daily mean measurements of O_3, NO_2 and NO, for Stations 4 and 10 (this numbering refers to the map of Fig. 1). We observe that O_3 and NO exhibit annual seasonality with slight linear trend, while NO_2 has a more irregular variation, with negligible seasonality and no trend. There does not seem to be different effects of the dynamics of each of the three variables between the two stations. In addition to the above measurements, the stations record daily meteorological information, such as temperature, humidity and wind speed.

The original data are high frequency values, however, we use daily averages calculated over 24 hours in order to limit micro-spatial, micro-temporal sampling uncertainties as well as measurement error. In each day the collected data consist of hourly measurements from midday 12:00 until midday 12:00 hours of the next day; these measurements are averaged to produce a single measurement for each day. Daily averaging reinforces also the general hypothesis of symmetry of the data in hand (which will be assumed later). For the stations used in the analysis, missing data at a given site and time have been estimated by using standard imputation methods, basically obtained as weighted averages of neighbouring available data, so that to keep the annual seasonality in the original data; for a review of imputation methods in air quality data see Junninen et al. (2004).

3 Monitoring

Let $y_{j,t}^{(i)}$ denote the daily mean measurement of the jth climate variable ($j = 1, 2, \ldots, p_i$) at time $t = 1, 2, \ldots, N$ and location $i = 1, 2, \ldots, L$. Then, at a given location i, we work with

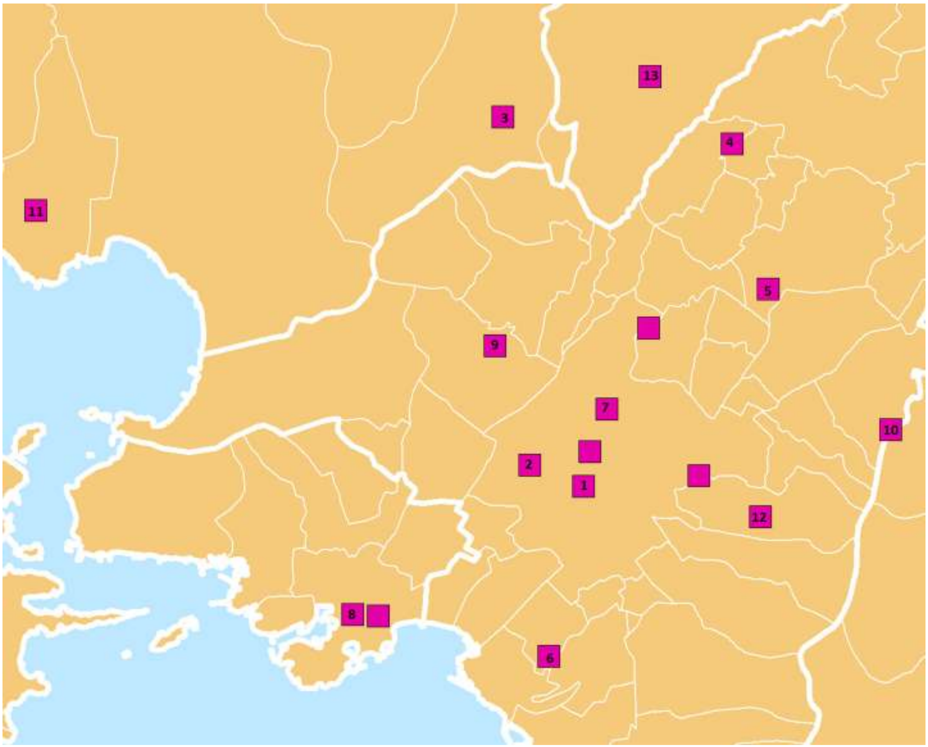

Fig. 1 Local map of Athens and its surrounding areas; highlighted are the active 17 stations (numbered stations are used in this study)

the p_i-dimensional time series $\{\mathbf{y}_t^{(i)}\}$, defined as the column vector of $y_{j,t}^{(i)}$ measurements, or $\mathbf{y}_t^{(i)} = (y_{1,t}^{(i)}, y_{2,t}^{(i)}, \ldots, y_{p_i,t}^{(i)})^\top$, where $\top$ denotes transposition, while for all L locations we work with a $r = \sum_{i=1}^{L} p_i$-dimensional time series $\{\mathbf{y}_t, \quad t = 1, 2, \ldots, N\}$, defined as the vector $\mathbf{y}_t = ((\mathbf{y}_t^{(1)})^\top, (\mathbf{y}_t^{(2)})^\top, \ldots, (\mathbf{y}_t^{(L)})^\top)^\top$, for $t = 1, 2, \ldots, N$. The r-dimensional vector $\mathbf{y}_t$, which contains the original measurements for all locations, forms a multivariate time series, which is autocorrelated and cross-correlated (contains a part of spatial effect for correlations between station variables). The proposed framework combines the use of time series forecasting, multivariate SPC and non-parametric methods for monitoring and early diagnosing beyond the expected air pollution levels. In particular, the following two analysis components are highlighted:

- **Process modelling.** We set-up a time series model which describes the "normal" long term behaviour of the time series $\mathbf{y}_t$. This is achieved at Phase I by using historical data as well as by allowing for some small temporal variation during Phase II (implementing a self-adaptive time series model).
- **Process monitoring.** We establish a real-time monitoring procedure on the residual vector $\mathbf{e}_t$ of the time series model by defining appropriate control procedures for identifying (i) an overall "aberrant" behaviour of the time series $\mathbf{y}_t$, (ii) an "aberrant" behaviour of the time series $\mathbf{y}_t$ in one or more neighbouring locations, and (iii) an

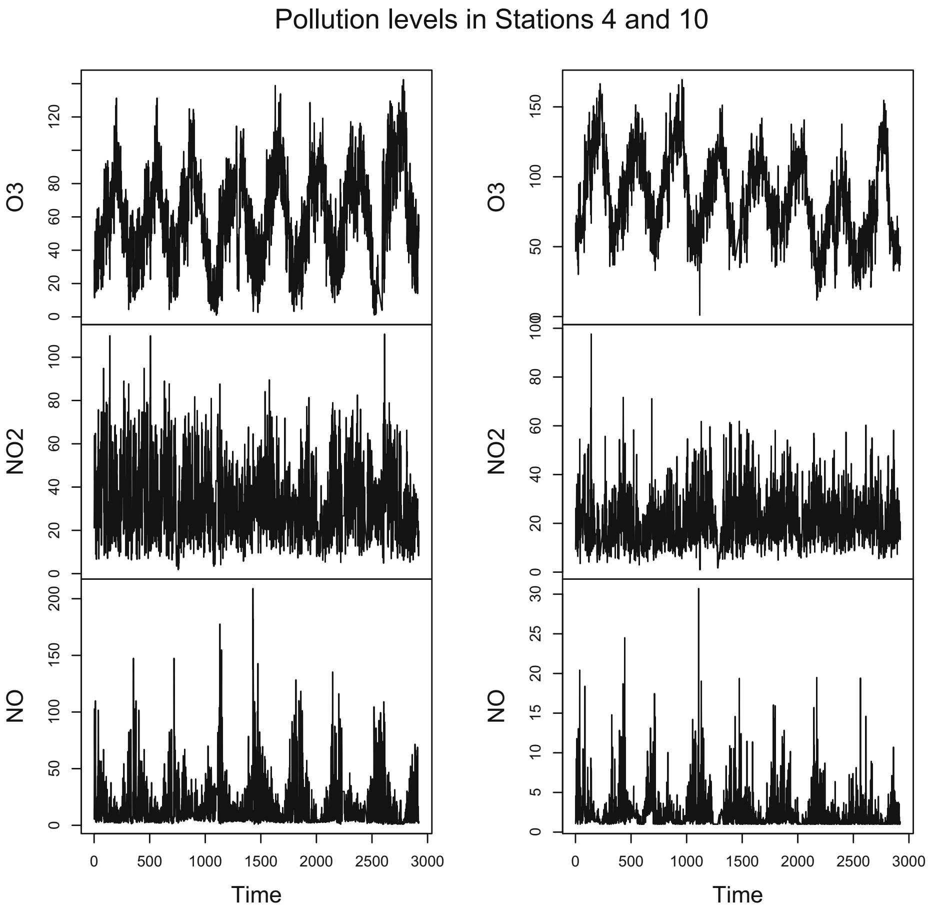

Fig. 2 Time series of O_3, NO_2 and NO for two stations (Station 4, left panel of plots and Station 10, right panel); see also Fig. 1 for the location of these stations in the map of Athens

"aberrant" behaviour of the time series $\mathbf{y}_t$ in specific components related to a certain air-quality variable.

3.1 Process Modelling

We consider a state space multivariate model for describing the temporal data variation in Phase I. In the sequel we discuss this model specification as well as inference. Let Θ_t be a time-varying $d \times r$ unobserved state matrix (for some positive integer d and $r = \sum_{i=1}^{L} p_i$ mentioned above), which drives the dynamics of $\mathbf{y}_t$ in the following state space model

$$\mathbf{y}_t = \Theta_t^\top \mathbf{F}_t + \epsilon_t. \tag{1}$$

The state matrix Θ_t is assumed to follow a Markov process, i.e. $\Theta_t = \mathbf{H}\Theta_{t-1} + \zeta_t$, where ζ_t is an innovation matrix with zero mean matrix and left covariance matrix $\mathbf{Z}_t$ and right covariance matrix Σ, written as $\zeta_t \sim (\mathbf{0}, \mathbf{Z}_t, \Sigma)$. The innovation term ϵ_t is assumed to have zero mean and covariance matrix Σ, written as $\epsilon_t \sim (0, \Sigma)$. It is noted that with vec$(\cdot)$ the column staking operator of a matrix, the covariance matrix of $\mathrm{vec}(\zeta_t)$ is $\Sigma \otimes \mathbf{Z}_t$, where $\otimes$ denotes the Kronecker product of two matrices. Hence the covariance matrix of $\mathrm{vec}(\zeta_t)$ is proportional to the observation covariance matrix Σ. This is commonly adopted in matrix-variate dynamic models when we wish to estimate Σ; see Triantafyllopoulos (2008) and Petris et al. (2010). According to Petris et al. (2010) there is no loss of generality specifying such a covariance structure for ζ_t, because the covariance matrix $\mathbf{Z}_t$ will compensate any discrepancies of Σ between ζ_t and ϵ_t. Furthermore, the innovation terms ϵ_t and ζ_t are assumed to be individually and mutually independent as well as independent of the initial state Θ_0. It is also assumed that the system is initialized at state Θ_0, with $\Theta_0 \sim (\mathbf{m}_0, \mathbf{P}_0, \Sigma)$, for some known matrix mean $\mathbf{m}_0$ and covariance matrix $\mathbf{P}_0$.

In the above model specification, the distributions of ϵ_t, ζ_t and Θ_0 are left unspecified; only means and variances of these quantities are specified. Based on this partial specification of ϵ_t and ζ_t, the likelihood function is not available, however, approximate Bayesian inference is achieved by Bayes linear methods, which approximate the posterior means and variances by minimising the expected posterior risk (Petris et al. 2010). The components of the vector $\mathbf{y}_t$ are likely to have the same distribution, albeit not a prespecified distribution such as the multivariate Gaussian distribution. Our study benefits by relaxing the distribution assumption and allowing a wider class of distributions, such as approximate Gaussian and Student t distributions; for a related discussion the reader is referred to Triantafyllopoulos and Harrison (2008).

Conditionally on the availability of model components $\mathbf{F}_t$, $\mathbf{H}$, Σ and $\mathbf{Z}_t$, we can use the celebrated Kalman filter, in order to obtain estimates of the state matrix and for forecasting. For the estimation of the covariance matrix Σ, we adopt the procedure of Triantafyllopoulos (2007), while the rest of the components are specified by the user (see Section 4.1). This procedure enables us to compute the standardized residuals $\mathbf{e}_t$, which are used in the monitoring stage of Section 3.2), and exhibit the following properties:

(a) they are serially independent, i.e. $\mathbf{e}_i$ is independent of $\mathbf{e}_j$, for $i \neq j$;

(b) for a fixed t, $\mathbf{e}_t = (e_{1,t}, e_{2,t}, \ldots, e_{r,t})^\top$ is a cross-independent random vector, i.e. $e_{j,t}$ is independent of $e_{k,t}$, for any $j \neq k$;

(c)] their distribution is approximately symmetric, since the limiting distribution of the residual vectors is Gaussian.

3.2 Process Monitoring

In this section the monitoring problem in Phase II is considered. The monitoring procedures, for the p_i air quality measurements as well as for the L distinct data recording stations, which are developed in this section, are based on the residual vectors $\mathbf{e}_t$, and on assumptions (a)–(c) of the previous section. The main idea is that if we assume that the sequence of the vectors $\mathbf{y}_t$, is fitted well in an appropriate time series model (like the one presented in the previous section), describing the usual in-control state of the process, the residual vectors $\mathbf{e}_t$, hold information about the deviation from "normal" pollution level (expected levels or in-control state of the process) to the actual values of the measurement vectors $\mathbf{y}_t$. In that way, large positive values for the components of $\mathbf{e}_t$, are implying deviation from the expected, or "normal", while small absolute values are implying a bond to the expected.

In the sequel three monitoring procedures are discussed, each of which aiming at the monitoring of the overall area, as well as for identifying a possible cluster of sub-areas or a specific air quality variable that exhibit aberrant behaviour. We propose binomial/generalised binomial type statistics and rank-based statistics in order to monitor the entire area under surveillance, a sub-section of that area (sub-area) of interest, or a particular set of variables of interest.

The binomial-type statistics deployed in each of the above three procedures use the vector $\mathbf{s}_t$, with elements $s_{i,t}$, defined as

$$s_{i,t} = \begin{cases} 1, & e_{i,t} \geq 0 \\ 0, & e_{i,t} < 0 \end{cases},$$

for $t = 1, 2, \ldots N$ and $i = 1, 2, \ldots, r$. Basically, this construction dichotomises the individual residuals $e_{i,t}$, into two classes: one for positive residuals, which correspond to potentially high levels of pollution, and the other with negative residuals, which correspond to low levels of pollution. In the sequel we propose particular binomial-based tests, exploiting the above dichotomy.

3.2.1 Overall Monitoring of the Area Under Surveillance

For the overall control of the area of interest we may use the statistic $T_{B,1}$ which represents the number of the components $e^{(i)}_{k,t}$, $k = 1, 2, \ldots, p_i$; $t = 1, 2, \ldots, N$; $i = 1, 2, \ldots, L$ of $\mathbf{e}_t$, $t = 1, 2, \ldots, N$ that are greater than zero. Specifically, statistic $T_{B,1}$ is defined as

$$T_{B,1} = \sum_{i,k} s_{i,k}, \quad k = 1, 2, \ldots, p; i = 1, 2, \ldots, L \quad \text{for some } t = 1, 2, \ldots, N$$

This statistic is based on the idea that if the process is in-control (which means that for all t, the vector $\mathbf{y}_t$ is fitted well in the appropriate time series model) the components $e^{(i)}_{k,t}$ of $\mathbf{e}_t$ must be randomly distributed above and below 0 (point of symmetry of the distribution). In contrast, if the vector $\mathbf{e}_t$ contains an extreme number of either positive or negative values we have evidence of a systematic out-of-control condition in most of the variables of interest and as well as in the whole area under inspection. This statistic under the null hypothesis that the process is in-control follows a binomial distribution with probability of success equal to 0.5 and a number of trials equal to the length of $\mathbf{s}_t$ $(r = \sum_{i=1}^{L} p_i)$.

Statistic $T_{B,1}$ is suitable for the overall control, since it is not sensitive to random or occasional outliers, as the summation in $T_{B,1}$ is responsible of issuing out-of-control signals only in the case that enough components of $\mathbf{e}_t$, $t = 1, 2, \ldots, N$ being systematically positive. We remark that since the procedure is based on the number of positive components of $\mathbf{e}_t$, $t = 1, 2, \ldots, N$ it is independent of the ordering of the components $e^{(i)}_{k,t}$ of $\mathbf{e}_t$.

For implementation purposes and effective application, we propose to use the normal approximation to the binomial distribution (since r is large enough, here 54), and thus we define the statistic $T'_{B,1} = (2T_{B,1} - r)/\sqrt{r}$, which follows the standard normal distribution $N(0, 1)$. Using this statistic, a classical rule is to issue an alarm when the observed value of $T'_{B,1}$ is larger than a one sided 3 sigma control limit (UCL = 3).

In general, we can implement additional control procedures, in order to improve the ability of $T'_{B,1}$ to be responsive and to adapt quickly to changes, but keeping the procedure as simple as possible. Such a procedure can be implemented by applying multiple limits, for the positive values of $T'_{B,1}$. We adopt three zones, each of which correspond to different statistical control decisions, i.e.

- Interval $I_1 = (-\infty, 1]$ of negative values of $T'_{B,1}$ or insignificant deviation, with probability $\Pr(T'_{B,1} \in I_1) = 0.841345$ (p_1).
- Interval $I_2 = (1, 3]$ of positive values of $T'_{B,1}$ with low deviation, with probability $\Pr(T'_{B,1} \in I_2) = 0.157305$ (p_2).

- Interval $I_3 = (3, +\infty)$ of positive values of $T'_{B,1}$ with high deviation), with probability $\Pr(T'_{B,1} \in I_3) = 0.001350$ (p_3).

Using these three zones, we can perform control at a weekly basis, or at another chosen period of time. Thus, for each week, we can track air pollution levels.

We propose that an alarm should be signalled, if either of the following rules apply:

- **Rule 1**: a plotted value in interval I_3 (the appearance of an extreme value, with probability 0.001350).
- **Rule 2**: 4 are plotted in interval I_2 in the last week (increasing trend for high deviation observations, with probability 0.000365). This is the probability to observe a particular event of the rule 4/7, e.g. I_2, I_2, I_1, I_1, I_2, I_1, I_2, or $0, 157305^4 \times 0, 841345^3 = 0, 000365$; there are 35 such events, with total probability 0.012763227.

Rule 2 can be modified, depending on the sampling frequency, e.g. if data are collected every 6 hours, then Rule 2 could be amended to 3 out of 4 observations of $T_{B,1}$. The study of such rule can be carried out in a similar way as that of 4 out of 7 procedure of Rule 2, which is studied in Appendix A, using the Markov chain embedding technique; for a related discussion of Markov chain embedding methodology the reader is referred to Balakrishnan and Koutras (2002). The mean and the standard deviation of the run length (RL) distribution are 147.22 and 143.29, respectively. This is in agreement with Frisen (2008), who states that, considering annual data, the usual ARL of the control procedures is set in the interval 120-240 containing the one third and the half of the year. We note that instead of using a Shewhart-type control chart supplemented with runs rules, a CUSUM or EWMA control chart may be proposed as an alternative. However, the choice of Shewhart type control chart and specifically the choice of a Shewhart control chart based only on the signs of the residuals, while supplemented with runs rules, is somehow straightforward. In particular, three key advantages are pointed out: (a) this procedure is easy to implement, which is necessary since the proposed framework is aimed at monitoring pollution and should be accessible to environmental scientists; (b) it is very robust to outliers (even conservative), which is critical since the nature of the application demands a very low level of false signals (just consider the case of a monitoring system providing a large number of false announcements of a hypothesised problem); and (c) this procedure is easily interpretable by the practitioner (see e.g. Koutras et al. 2007).

(ii) Monitoring Statistics Based on Ranks For the overall control of the area of interest we may use the statistic $T_{R,1}$ which represents the sum of the ranks of the positive values $e_{i,t}$, $R^+(e_{it})$, of the vector $\mathbf{e}_t$. In that way, we define a Wilcoxon-type monitoring statistic which is

$$T_{R,1} = \sum R^+(|e_{it}|), \quad t = 1, 2, \ldots; i = 1, 2, \ldots, r$$

where $R^+(\cdot)$ represents the rank of an element. This statistic is approximated by normal distribution for fairly large values of r (see Gibbons and Chakraborti 2010) using the following standardisation

$$T'_{R,1} = \frac{\sum R^+(|e_{it}|) - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}, \quad t = 1, 2, \ldots; i = 1, 2, \ldots, r$$

Under the hypothesis that the process is under control we may calculate the appropriate 3 sigma control limit (here $UCL = 3$, for $\alpha \approx 0.00135$).

3.2.2 Sub-area Control

(i) Binomial and Generalised Binomial Type Monitoring Statistics As already mentioned, the control of the whole area is not sufficient since many times the problem discussed appears in a small number of sub-areas (in general in a subset of sub-areas). Thus, the control chart proposed above, is inappropriate since it cannot detect clustering of sub-areas having extreme values. For example, this control procedure treats the same the case of a number of random components of $\mathbf{e}_t$ being positive and the case that any such number of positive components of $\mathbf{e}_t$ being observed at the same or nearby areas. Thus, we need an appropriate procedure that identifies clusters of extreme cases in the sequence of $e_{k,t}^{(i)}$, $k = 1, 2, \ldots, p_i$; $t = 1, 2, \ldots, N$; $i = 1, 2, \ldots, L$. For this reason the elements of $\mathbf{e}_t$ are ordered according to their spatial neighbouring. The ordering of the components of $\mathbf{e}_t$, can be roughly succeeded using the matrix of distances of the recording stations and one of the available techniques of multivariate ordering such as the minimal spanning tree. In other words, the components e_{it} are ordered according to their smallest distance (areas which are classified by the minimal spanning tree as "close" place their corresponding residuals one after the other).

In this case, we may use a runs-based statistic, say $T_{B,2}$, similar to the one introduced by Antzoulakos et al. (2003), defined as

$$T_{B,2} = \sum_{i=w_p}^{r} U_i,$$

where w_p is a constant positive integer depending on p (usually w_p is set to be slightly smaller than p) and

$$U_i = \begin{cases} w_p + j, & \text{if } s_{i-w_p-j+1} = s_{i-w_p-j+2} = \cdots = s_i = 1, s_{i-w_p-j} = s_{i+1} = 0 \\ 0, & \text{otherwise} \end{cases},$$

for the r components of $\mathbf{s}_t$ and $s_0 = s_{r+1} = 0$ by convention. This statistic is the sum of the run-lengths of length w_p and more. If this statistic takes large values, this means that there are some sub-areas that experience high levels of pollution to all or to the majority of their variables. On the contrary if $T_{B,2}$ takes small values then the 1s are randomly assigned to the vector, which means that no specific area has a major problem. The $T_{B,2}$ sums the lengths of runs each having at least length w_p or more w_p. We do not count the number of runs of a specific length or the number of runs of length above a threshold. Instead, we sum the lengths of runs of length w_p or more which enables the identification of clusters within a single sub-area or between two consecutive sub-areas. This is useful since, due to spatial correlation of neighbouring areas, pollution levels are expected to attain similar exposure.

According to Antzoulakos et al. (2003), the performance of this statistic as a randomness test is shown to be significantly more powerful than competitors when the type I error must be kept low, e.g. $\alpha = 0.01$, which is exactly our case. The control chart based on this statistic has also only an upper control limit since the clustering of areas with extreme values, naturally drives the $T_{B,2}$ statistic to high values. The control limit for the chart based on the distribution of $T_{B,2}$, is calculated using the formula

$$\begin{aligned} g_r(x) = {} & g_{r-1}(x) + \frac{1}{2} g_{r-1}(x-1) - \frac{1}{2} g_{r-2}(x-1) - \frac{1}{2^{w_p+1}} \{g_{r-w_p-1}(x) \\ & - g_{r-w_p-1}(x-w_p)\} - \frac{1}{2^{w_p+2}} \{g_{r-w_p-2}(x-w_p) \\ & - g_{r-w_p-2}(x-1)\}, \end{aligned} \tag{2}$$

for $x \geq 0, n \geq w_p + 2$, which is the result of appropriately adopting Theorem 3.2 provided by Antzoulakos et al. (2003). Table 1 shows upper control limits (UCL) for two values of α and for $k = 4, 5, 6, 7$.

For $r < w_p + 2$ the initial conditions of the same theorem can be explored appropriately. In this case the ARL of this control chart can be easily calculated using the geometric distribution with probability of success α. We note that the control chart based on $T_{B,2}$ may be used independently of the control chart based on $T_{B,1}$. This means that the control chart based on $T_{B,2}$ is not supplementary to the control chart based on $T_{B,1}$, since it can be used without using necessarily the $T_{B,1}$ control chart. However, it should be noted that $T_{B,1}$ and $T_{B,2}$ are not independent.

(ii) Control Statistics Based on Ranks As we already mentioned the control of the whole area is not sufficient since many times the problem appears in only one or two sub areas (in general in a subset of subareas). Using the residual vector $\mathbf{e}_t$ we may use the ranks of the components of the vector $\mathbf{e}_t$ and test if one or more areas have extreme values. For example, we may find that in one area the sum of the ranks is too high while in another area the sum of the ranks is too low. Using this idea, we can define a Kruskall-Wallis type control statistic (e.g. Gibbons and Chakraborti 2010) which is given by

$$T_{R,2} = \frac{12}{r(r+1)} \sum_{i=1}^{L} \frac{R(e_{i,t})^2}{p_i} - 3(r+1), \quad t = 1, 2, \ldots$$

where $R(\cdot)$ represents the sum of the ranks of the residuals related to the corresponding location. This statistic is approximated by a Chi-Square distribution with $L - 1$ degrees of freedom for fairly large values of r.

3.2.3 Variable Control

(i) Binomial and Generalised Binomial Type Statistics Reordering the components of $\mathbf{e}_t$ in such a way that the same air quality variable from all the recording stations are placed together in blocks using the following format for

$$\mathbf{e}_t = (\underbrace{e_{1,t}, e_{2,t}, e_{3,t}, e_{4,t}, e_{5,t}}_{\text{Air-Quality Variable 1}}, \underbrace{e_{6,t}, e_{7,t}, e_{8,t}, e_{9,t}, e_{10,t}}_{\text{Air-Quality Variable 2}}, \ldots, e_{r,t})^{\top},$$

we can immediately define a control procedure which aims in identifying if a specific air quality characteristic is out-of-control in all of the areas.

Specifically, we establish a similar procedure to that of sub-area monitoring, based on $T_{B,3}$, defined as

$$T_{B,3} = \sum_{i=w_L}^{r} U_i,$$

Table 1 Upper control limits for $T_{B,2}$ for significance levels α and $k = 4, 5, 6, 7$

k	UCL	α	k	UCL	α
4	23	0.01	4	28	0.001
5	19	0.01	5	25	0.001
6	16	0.01	6	22	0.001
7	15	0.01	7	20	0.001

where w_L is a constant (similar to w_p described in the previous section) and

$$U_i = \begin{cases} w_L + j, & \text{if } s_{i-w_L-j+1} = s_{i-w_L-j+2} = \cdots = s_i = 1, s_{i-w_L-j} = s_{i+1} = 0 \\ 0, & \text{otherwise} \end{cases},$$

for the r components of $\mathbf{s}_t$ and $s_0 = s_{r+1} = 0$ by convention. If the number of occurrences is high, then we assume that certain variables have an aberrant behaviour. The control chart based on this statistic has only an upper control limit since the same variable in many areas has extreme values, naturally drives the $T_{B,3}$ statistic to high values.

(ii) Control Statistics Based on Ranks Regrouping the components of $\mathbf{e}_t$ in such a way that the same air quality variable from all the recording stations are placed together and establishing a similar in nature to the rule of Section 3.2.2(ii), we develop immediately a control procedure which aims in identifying if a specific air quality characteristic is out of control.

3.2.4 Joint Effects of the Control Procedure Based on the Binomial-Type Statistics

In this section we study some of the joint effects of the performance of the control procedures described above. We start by looking at the probability that $T_{B,2}$ (sub-area control) or $T_{B,3}$ (variable control) give an out of control signal, provided that $T_{B,1}$ has signalled an out of control point. In Appendix B we show that the joint tail probability of $\{T_{B,2} > c_2\}$ and $\{T_{B,1} > c_1\}$ is

$$\begin{aligned} \Pr(T_{B,2} > c_2, T_{B,1} > c_1) = & \sum_{i=c_1+1}^{n} \sum_{j=c_2+1}^{n} \binom{n}{n-i}^{-1} \sum_{\ell=0}^{n-i+1} \sum_{i_1=0}^{\ell} \sum_{j_1=0}^{\ell-i_1} \\ & \times \sum_{j_2=0}^{i_1} (-1)^{\ell+i_1+j_1-j_2} \binom{n-i+1}{\ell} \binom{\ell}{i_1} \binom{\ell-i_1}{j_1} \\ & \times \binom{i_1}{j_2} \binom{\ell+a-1}{a} \binom{n-i+b}{b} \binom{n}{i} 0.5^n, \end{aligned} \tag{3}$$

where $a = j - i_1 + j_2 - w_p(j_1 + j_2)$ and $b = i - w_p\ell - i_1 - a$, for some constants c_1 and c_2. Therefore, the conditional probability of $\{T_{B,2} > c_2\}$ given $\{T_{B,1} > c_1\}$ is provided by the definition of the conditional distribution

$$\Pr(T_{B,2} > c_2 \mid T_{B,1} > c_1) = \frac{\Pr(T_{B,2} > c_2, T_{B,1} > c_1)}{\Pr(T_{B,1} > c_1)}.$$

Equation 3 provides the probability that both $T_{B,1}$ and $T_{B,2}$ or $T_{B,1}$ and $T_{B,3}$ signal an out-of-control point, while appropriate use of the conditional distribution can help us to control the overall probability of type I error (i.e. falsely declare the process as out-of-control while the process is actually in-control).

If the process is in-control and $T_{B,1}$ does not provide a signal it is hoped that the probability of $T_{B,2}$ or $T_{B,3}$ not signalling (since the process is assumed to be in control) should be high. For specified values of k and UCL, Table 2 shows the probability the statistic $T_{B,2}$ not to issue an out-of-control signal, provided that $T_{B,1}$ failed to issue a signal. We observe that under the null hypothesis (in-control state) $T_{B,2}$ has small probability to wrongly issue an out of control signal. Table 3 gives the probability that $T_{B,2}$ or $T_{B,3}$ signal an out-of-control point, given that $T_{B,1}$ has signalled an out-of-control point. We remark that as the observed value of $T_{B,1}$ increases, the probability that $T_{B,2}$ or $T_{B,3}$ issue and out-of-control signal is

Table 2 Probability of $T_{B,2}$ or $T_{B,3}$ not signalling, given the process is in control and $T_{B,1}$ did not signal

k	UCL	$\Pr(T_{B,2} < \text{UCL} \mid T_{B,1} < 39)$
4	28	0.800
5	25	0.898
6	22	0.876
7	20	0.933

increased too. As large probability of $T_{B,2}$, $T_{B,3}$ implies large observed value of $T_{B,2}$, $T_{B,3}$, it follows that $T_{B,1}$ and $T_{B,2}$, $T_{B,3}$ exhibit positive correlation.

Below is a summary of the proposed process modelling and monitoring procedures, which are put into practice in Section 4.

Summary of monitoring procedure

1. **Phase I (process fitting).** Set Phase I length and prior settings of Section 3.1 Fit the state space model (1) using the Kalman filter and the method of Section 3.1. For each time t, obtain the standardised residuals $\mathbf{e}_t$.
2. **Phase I (process testing).** Test whether assumptions (a)–(c) of Section 3.1 are satisfied. Make sure model (1) is a true representation of the process in Phase I. Optimise the hyperparameters of the model, such as the discount factor δ. If assumptions (a)–(c) not satisfied modify the model.
3. **Phase II (process monitoring).** Continue to fit model (1) using the optimised parameters from Phase I. Obtain sequentially standardised residuals $\mathbf{e}_t$, for each t in Phase II. Chart using binomial-type statistics or rank-based statistics, for the three specific cases:

   - Overall area control. Use the statistics $T_{B,1}$ or $R_{t,1}$ of Section 3.2.1.
   - Sub-area control. Use the statistics $T_{B,2}$ or $R_{t,2}$ of Section 3.2.2.
   - Variable control. Use the statistics $T_{B,3}$ or $R_{t,3}$ of Section 3.2.3.

Table 3 Probability of $T_{B,2}$ or $T_{B,3}$ signalling, given that $T_{B,1}$ has signalled an out of control point. Shown are the probabilities $\mathrm{P}(x) = \Pr(T_{B,2} > x \mid T_{B,1})$, for $k = 4, 5, 6, 7$

$T_{B,1}$	$T'_{B,1}$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
		P(28)	P(25)	P(22)	P(20)
38	2.99	0.200	0.102	0.124	0.067
39	3.27	0.352	0.190	0.206	0.117
40	3.54	0.541	0.320	0.322	0.192
41	3.81	0.730	0.487	0.466	0.298
42	4.08	0.877	0.667	0.625	0.433
43	4.35	0.960	0.825	0.776	0.588
44	4.63	0.992	0.931	0.893	0.743
45	4.90	0.999	0.982	0.963	0.871
46	5.17	1.000	0.997	0.992	0.954
47	5.44	1.000	1.000	0.999	0.990
48	5.72	1.000	1.000	1.000	0.999
≥49	5.99	1.000	1.000	1.000	1.000

4 Illustration: Air Pollution Data

4.1 Modelling

In this section we consider the pollution data described in Section 2. In the first stage we use a state space model for modelling the temporal data variation in the time period January 1, 2001 to 4 December 2008 (Phase I); this corresponds to 2895 observations in Phase I. The proposed model is a time-varying regression model which comprises trend and seasonal components in order to take into account temporal variation. Specifically, three covariates (humidity, temperature and wind speed), denoted by $x_{1,t}, x_{2,t}, x_{3,t}$, respectively, enter in the time series model of $\mathbf{y}_t$ as time-varying regressor variables, with corresponding regression parameters following a random walk evolution. The state space model adopted for $\mathbf{y}_t$ is

$$\mathbf{y}_t = \mathbf{R}_{1,t} + \mathbf{R}_{2,t} + \mathbf{R}_{3,t} + \mathbf{T}_t + \mathbf{S}_t + \epsilon_t = \Theta_t^\top \mathbf{F}_t + \epsilon_t, \tag{4}$$

so that $\mathbf{y}_t$ comprises three dynamic regression components $\mathbf{R}_{i,t} = \Theta_{i,t}^\top x_{i,t}$, a trend component $\mathbf{T}_t = \Theta_4^\top [1, 0]^\top$, a seasonal component $\mathbf{S}_t = \Theta_5^\top [1, 0, 1, 0, 1, 0, 1, 0, 1, 0]^\top$ and a random innovation term ϵ_t, where $\Theta_t^\top = [\Theta_{1,t}^\top, \Theta_{2,t}^\top, \Theta_{3,t}^\top, \Theta_{4,t}^\top, \Theta_{5,t}^\top]$. The state matrix and the rest of the state space model is as defined in Section 3.1. A weakly informative prior setting is adopted whereby $\mathbf{P}_0 = 1000\mathbf{I}$ (nearly zero precision $\mathbf{P}_0^{-1} \approx \mathbf{O}$) and $\mathbf{m}_0 = \mathbf{0}$. The observation covariance matrix Σ is estimated as discussed in Section 3.1, while $\mathbf{Z}_t$ is specified using a single discount factor δ (Petris et al. 2010).

It is noted that model (4) is implied by adopting the following setting

$$\begin{gathered}\mathbf{F}_t = \left[x_{1,t}, x_{2,t}, x_{3,t}, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0\right]^T, \\ \mathbf{H} = \text{block diag}\,[\mathbf{I}, \mathbf{J}_1, \mathbf{J}_2(\phi), \mathbf{J}_2(2\phi), \mathbf{J}_2(3\phi), \mathbf{J}_2(4\phi), \mathbf{J}_2(5\phi)],\end{gathered}$$

where “block diag” indicates a block diagonal matrix with components $\mathbf{I}$ for the identity matrix, $\mathbf{J}_1$ for a Jordan block with a unit eigenvalue, responsible for the trend variation, and $\mathbf{J}_2(\phi)$ for the harmonic component, responsible for the seasonal variation, i.e.

$$\mathbf{J}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{J}_2(\phi) = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}.$$

The above state space model is known as a reduced form trend-seasonal model (Petris et al. 2010), for the seasonal variation of which, a reduced form state space representation of the full Fourier expansion is used. For this data the cycle is $c = 365$ (daily data with annual seasonality) with frequency $\phi = 2\pi/c$; here for computational efficiency, we use only the first 5 harmonics out of a total of $(c-1)/2 = 182$, hence the reduced form.

Assumptions (a)-(c) of Section 3.1 were validated in Phase I, after the first 100 observations were considered as training data and excluded from the tests. Assumptions (a)-(b) were validated by performing standard white noise tests (Petris et al. 2010). The symmetry of the residuals (assumption (c)) was established by first grouping positive and negative residuals and then performing a Kolmogorov-Smirnov test to ensure the two groups have the same distribution. The p-value of the test was 0.1979 and so assumption (c) was validated. We used a standard Chi-square test applied on the residuals to assess the goodness of fit (Prado and West 2010; Petris et al. 2010). We concluded there was overwhelming evidence in favour of the model fit, with the corresponding p-value being almost equal to 1.

4.2 Monitoring

Considering Phase I analysis, in the period 11 April 2001 to 4 December 2008 (the first 100 observations are used as attaining set), we report on Rules 1 and 2, described above; Fig. 3 gives the Phase I control chart for the last 60 days of Phase I (6 October 2008 to 4 December 2008). Considering the entire Phase I period, there are 3 points beyond the 3σ control limit (Rule 1) while 3.9 were expected. Additionally, by applying the second rule (Rule 2), 15 more signals are given by the proposed control chart. In total 18 signals were found in Phase I, while 19 were expected, hence it is consistent with the in-control ARL = 147. One of the signals corresponds to clusters of time where the deviation from the model is consistent. As this analysis reveals, the period in which the deviation from the model is consistently large, corresponds to a period that the measurements are systematically quite large in relation to expected pollution levels, but still not beyond the limits set by environmental agencies. In this period, as the study revealed, an extreme heat wave hit Athens, Greece's capital. Figure 4, shows the original values of NO_2 (from a single station) together with estimated levels (averages) over the period of 10 October 2007 to 29 January 2008. The period of 4 days (time points 2546–2549), which is identified by the control chart based on T_1 statistic, corresponds to a period of very high values of NO_2 (the most extreme in a period of about 4 months); these 4 points are highlighted in Fig. 4.

In Fig. 5 the Phase II control chart based on $T_{B,1}$ statistic for overall control is given. As Phase II, we consider the time period 5 December 2008 to 31 December 2008 (27 days/observations). Employing the model from Phase I, with the optimal value of $\delta = 0.3$, we apply the adaptive estimation of the state space model, for the data in Phase II. That is,

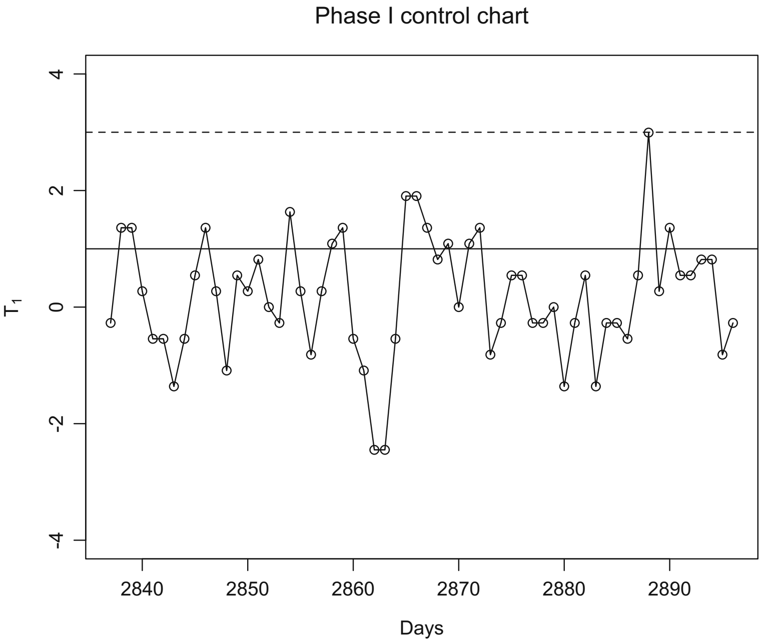

Fig. 3 Control chart for overall monitoring in Phase I. Shown are values of $T_{B,1}$ statistic in Phase I for the last 2 months of Phase I (6 October 2008 to 4 December 2008)

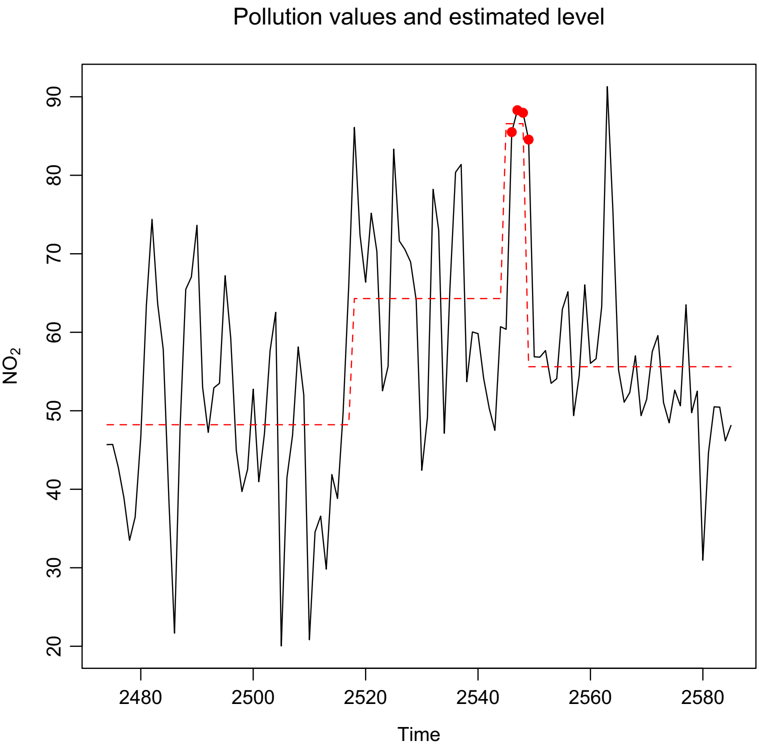

Fig. 4 Original data of NO_2 together with estimated levels in the period 10 October 2007 to 29 January 2008

at each time t we apply the Kalman filter and the related time series algorithm described above, in order to obtain sequentially residuals, which measure deviations from the normal process.

We observe that no value of $T_{B,1}$ exceeds the 3 control limit and so there is no signal issued under Rule 1. Considering Rule 2 (4 or more values of $T_{B,1}$ exceeding 1 in last week) we observe that within the first 10 days in Phase II (5–14 December 2008) there are four days in a window of less than or equal to 7 with $T_{B,1}$ exceeding 1 (observations of 9, 10, 11 and 14 December 2008). Thus, following Rule 2, on 14 December there is a signal issued (indicated by red point in Fig. 5), which highlights a sequence of large pollution levels, in this time interval. This suggests that in this week, large deviations from normal (average) pollution were present. This conclusion is supported by the data itself, since in that period a high pollution peak is present to most of the variables.

Proceeding to sub-area control, we need to rearrange the residual vector $\mathbf{e}_t$ according to station proximity. This is achieved by considering the ordering of a minimal spanning tree (MST), implemented in R using the package vegan. Exploring the MST in relation to the map (Fig. 1) we note that stations 3, 4, 13 are close enough in the map and this is depicted in the MST; stations 5, 9 are also close enough and the MST agrees, and so forth. Thus, we conclude that the output of the MST is consistent with the map. Comparisons (not shown here) of MST with hierarchical agglomerative clustering resulted in the same sensor grouping.

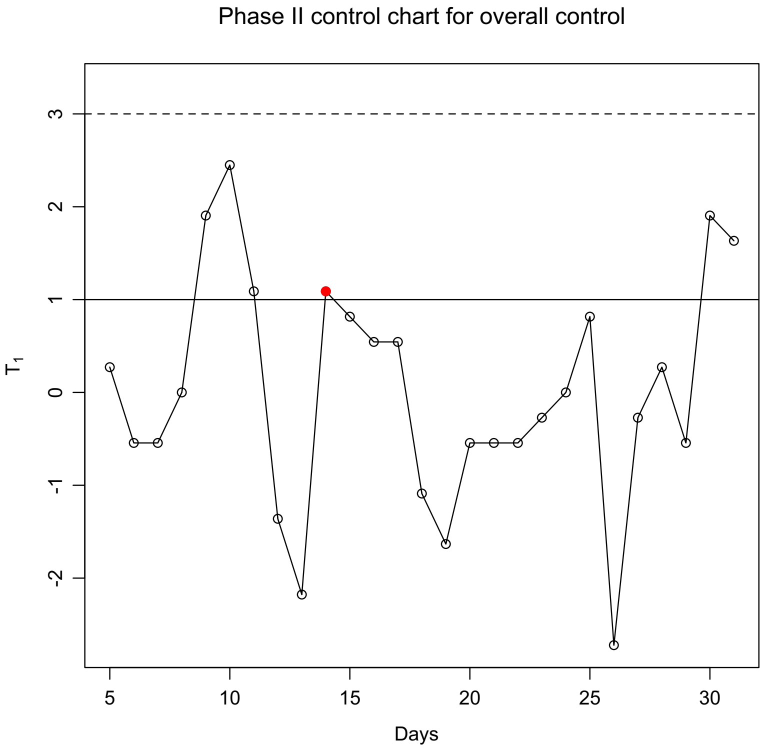

Fig. 5 Control chart for overall monitoring. Shown are values of $T_{B,1}$ statistic in Phase II (5 December to 31 December 2008)

For $t = 2901$ (corresponding to the date 10 December 2008) the residuals are ordered as

$$\begin{aligned}\mathbf{e}_{2901} = {} & [1.40, -0.47, 0.77, 0.19, 1.06, 0.04, 0.26, -0.82, 0.35, 1.22, 0.75, -0.53,\\ & -0.13, 0.66, 0.05, 0.93, 0.22, -1.39, -0.41, -0.32, 0.56, 0.51, 0.38, -2.13,\\ & -2.35, -0.21, 1.27, 0.74, 0.15, 0.59, 0.15, 1.99, -1.09, 0.47, 1.64, 0.92,\\ & 0.69, -0.67, 0.17, 0.51, 0.11, 0.24, -0.80, 0.56, 1.32, -1.20, -0.61, 0.33,\\ & 0.77, -0.19, 0.87, -0.61, -0.06, 0.01]^\top.\end{aligned}$$

Thus according to the definition of $s_{i,t}$ in the previous section, the following sequence of vector with binary variables is obtained:

$$\begin{aligned}\mathbf{s}_{2901} = {} & [1, 0, \underbrace{1, 1, 1, 1, 1}_{5}, 0, 1, 1, 1, 0, 0, \underbrace{1, 1, 1, 1}_{4}, 0, 0, 0, 1, 1, 1, 0, 0, 0, \underbrace{1, 1, 1, 1, 1, 1}_{6},\\ & 0, \underbrace{1, 1, 1, 1}_{4}, 0, \underbrace{1, 1, 1, 1}_{4}, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1]^\top.\end{aligned}$$

For example we observe that since $e_{1,t} = 1.40 \geq 0$, we have $s_{1,t} = 1$; likewise as $e_{2,t} = -0.47 < 0$, we have $s_{2,t} = 0$ and so forth. $T_{B,2}$ gives a signal if a large number of runs of length at least 4 are observed; in the above $\mathbf{s}_t$ such consecutive units are indicated. Thus, for this point of time (10 December 2008), there is an out of control signal issued, since $T_{B,2} = 23$. The control chart, applies this control procedure for each time t in Phase II and it is shown in Fig. 6. We observe that on 10 December a signal is issued, as having high levels of pollution when sub-area ordering is taken into account. A careful study of the data in this period reveals that stations in the centre of Athens recorded high values across the pollutants.

In order to explore the performance of $T_{B,2}$ as opposed that of $T_{B,1}$, we observe that the points $t = 2900$ and $t = 2921$ (corresponding to 9 December and 30 December 2008, respectively) yield the same value of $T_{B,1}$, i.e. $T_{B,1} = 34$ ($T'_{B,1} = 2$), while $T_{B,2}$ give respective values of 18 and 8. Thus, while the number of extreme values is the same for both points (same value of $T_{B,1}$, with same number of units in $\mathbf{s}_t$ vector), the values of $T_{B,2}$ have a notable difference (more than 100%), which could possibly translate to different control decisions (out-of-control and in-control). This simple example shows that the related correlation of $T_{B,1}$ and $T_{B,2}$ (both depend on the number of variables p), is not dominant, with regards to respective control decisions. Furthermore, $T_{B,2}$ describes a local control procedure (defined by the station proximity), while $T_{B,1}$ describes a global control procedure.

Finally, we obtain a new ordering of $\mathbf{e}_t$, according to the common air-variables, for each station. For example, the first 13 elements of $\mathbf{e}_t$ correspond to the values at time t of O_3, for

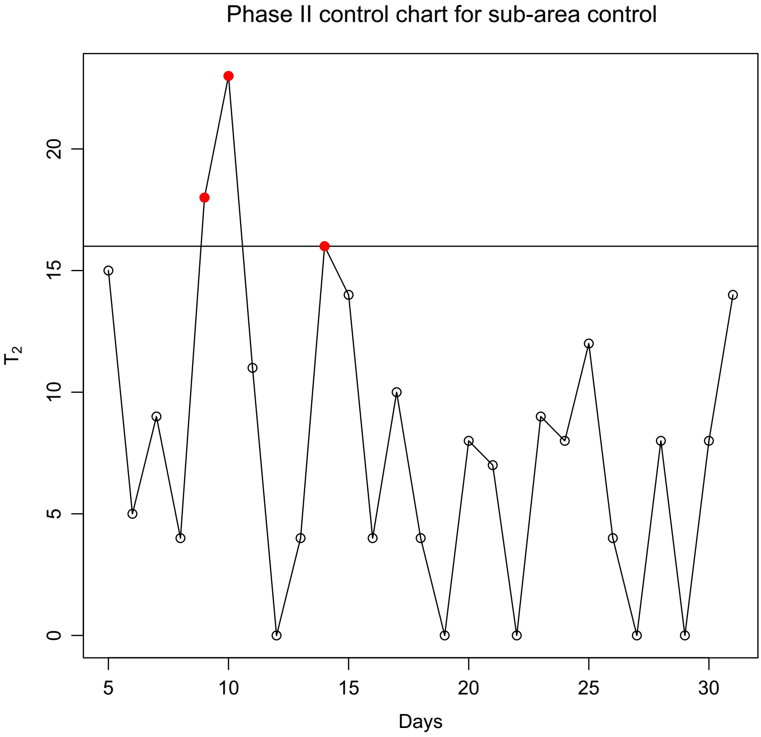

Fig. 6 Control chart for sub-area monitoring. Shown are values of $T_{B,2}$ statistic in Phase II (5 December to 31 December 2008)

each of the 13 stations. Based on this construction, a similar vector of binary variables $\mathbf{s}_t$ is obtained, i.e.

$$\begin{aligned}\mathbf{s}_{2901} = {} & [1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0,\\ & \underbrace{1, 1, 1, 1, 1}_{5}, 0, 1, 0, \underbrace{1, 1, 1, 1, 1}_{5}, 0, 0, 0, \underbrace{1, 1, 1, 1, 1, 1}_{6}]^\top.\end{aligned}$$

Here we can see the difference between $\mathbf{s}_t$ of sub-area control and variable control. The control is similar as in $T_{B,2}$, but here we use $w_L = 5$, i.e. 5 or more consecutive units account for a signal at time t. The control chart using procedure $T_{B,3}$ is depicted in Fig. 7. We observe that dates 10, 14 and 30-31 December give out of control signals. We note that using the $T_{B,3}$ statistic we may sum runs of excess length that belong to consecutive blocks of variables.

The choice of w_p and w_L in the statistics $T_{B,2}$ and $T_{B,3}$, respectively, is indicative. As suggested in Antzoulakos et al. (2003), it is usual practice to use lower values of these parameters (3 or 4), since we do not expect all variables to be positive in an area with high pollution levels.

Having discussed the binomial-type monitoring statistics it is noted that the application of rank statistics had a similar performance in both phases. However, their performance appear to be slightly deteriorated in the case of Athens considered in this paper. In particular, this is observed when comparing the performance of $T_{B,2}$ based control chart with the one that is

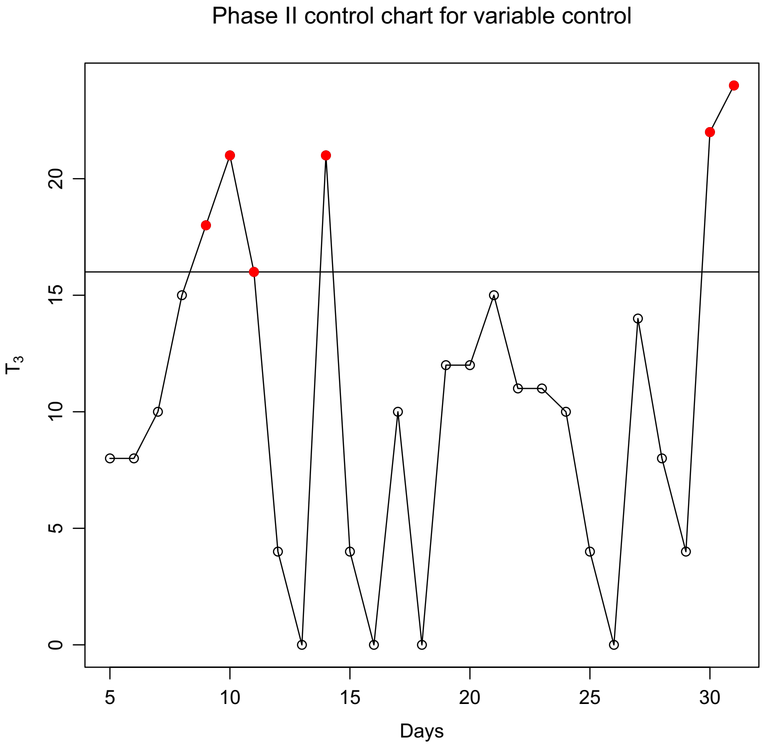

Fig. 7 Control chart for variable monitoring. Shown are values of $T_{B,3}$ statistic in Phase II (5 December to 31 December 2008)

based on Kruskall-Wallis test statistic ($T_{R,2}$). This is due to the deterioration of the Kruskal-Wallis type statistic as the number of the different areas increases. The rules based on $T_{B,1}$ supplemented with runs appeared to be better even from standard CUSUM-type procedure. For this reason we propose that for large number of variables and sub-areas the binomial and the generalised binomial control procedures are preferred than the rules based on ranks.

4.3 Discussion

By comparing the three Phase II control charts (overall control, sub-area control and variable control) we see that on the period 10-14 December all charts indicate high levels of pollution, although the overall control occurs a delay by only detecting this abnormal behaviour on the 14 December, while $T_{B,3}$ chart is faster. The overall control does not signal any other aberrant behaviour (the same is true for $T_{B,2}$), while the variable control issues a further signal on 30 December. It appears that when $T_{B,1}$ signals an out-of-control point, this is the result either or both of out-of-control signals in $T_{B,2}$ or/and $T_{B,3}$. To expand on this point, $T_{B,1}$ fails to issue a signal on 10 December, while $T_{B,2}$ and $T_{B,3}$ captures this on 10 December. Thus, $T_{B,2}$ and $T_{B,3}$ signal locally the out-of-control signal, which then becomes global and it is captured by $T_{B,1}$ later on 14 December. On 30 December $T_{B,3}$ issues a clear out-of-control signal, while $T_{B,1}$ shows some tendency towards a signal, but it is not clearly issued. It seems that this implies that a local concentration of high pollution levels tends to increase the global pollution levels. Since the values of $T_{B,2}$ are far from the limits for 30 December, it appears that the problem is captured well and on-time by $T_{B,3}$ which means that the problem is associated with clustering of high values to specific pollutants.

In conclusion in the period 10-14 December there is a certain deviation from the expected pollution levels. This deviation starts by a local problem (which is clearly captured with an out-of-control signals by $T_{B,2}$) and becomes a global problem as some pollution levels are systematically high ($T_{B,3}$ gives two out-of-control signals in these 5 consecutive dates). In the last two time points (30 and 31 December) $T_{B,3}$ gives a clear out-of-control signal. As the analysis revealed, in Phase I we identified periods of time with extreme deviations from expected while in Phase II the proposed monitoring scheme captured extreme pollution deviation, both locally and globally.

5 Conclusions

This paper develops a unified framework for monitoring the mean effects of several air-pollution variables, over a network of stations. This framework combines (a) the use of time series modelling, for the temporal description of mean pollution levels and estimation of cross-correlation of pollution variables, and (b) multivariate control chart methods, for the detection of deviations from the expected pollution levels. We develop a multivariate statistical process control approach, which allows the monitoring of the pollution levels in the overall area and in sub-areas, determined by rules of neighbouring similarity. Air-pollution data from the city of Athens are used to put in practice the proposed monitoring approach.

The application to Athens data reveals that the proposed methodology can provide remarkable evidence about the source of the pollutant temporal variation. In particular, it is able to identify whether temporal exceedance come from a specific station (location) or it is attributed to a specific pollutant. It is able to identify short periods of large variations that are related to excess numbers of hospital admissions even when the formal thresholds (issued by the environmental protection agencies) are not exceeded.

This paper makes a contribution on multivariate SPC for non-industrial processes, which is currently in demand (see relevant discussion in the introduction), and is an area which is likely to receive even more attention in the near future.

Acknowledgements This work is supported by the General Secretariat for Research and Technology (GSRT, Ministry of Education, Greece) research funding action ARISTEIA II.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix A: Derivation of the Run Length Distribution of the Control Chart Using Rules 1 and 2 Based on $T_{B,1}$

The ARL calculation will be performed using the Markov embedding technique. According to this technique a discrete random variable, say W, defined on a sequence of multi-state trials may be described by a Markov chain $\{V_t, t = 0, 1, 2, \ldots\}$ defined on a finite state space $\Omega = \{\alpha_1, \alpha_2, \ldots, \alpha_s\}$. If we let α_s be the absorbing state, the cumulative distribution of a random variable W is given by

$$\Pr(W \leq n) = \Pr(V_n = \alpha_s) = \boldsymbol{\pi}_0^\top \boldsymbol{\Lambda}^n \mathbf{e}_s,$$

where $\boldsymbol{\pi}_0^\top = \{\Pr(V_0 = \alpha_1), \Pr(V_0 = \alpha_2), \ldots, \Pr(V_0 = \alpha_s)\}$ is the vector of initial probabilities of the Markov chain, $\boldsymbol{\Lambda} = [\Pr(V_t = \alpha_j \mid V_{t-1} = \alpha_i)]_{s\times s}$ is the transition probability matrix and finally, $\mathbf{e}_s^\top = (0, 0, \ldots, 0, 1)_{1\times s} \in \mathbb{R}^s$. Using the above procedure, the probability distribution of W may be computed (Balakrishnan et al. 2009).

A suitable state space for V_t is $\Omega = \Omega_1 \cap \{\alpha_s\}$, where

$$\begin{aligned}\Omega_1 = \{&(i_0, i_1, i_2, i_3) : i_0 = 0, 1, 2, 3, \quad i_1 = 2, 3, 4, 5, \quad i_2 = 1, 2, 3, 4,\\ &i_3 = 0, 1, 2, 3 \quad \text{and} \quad i_1 > i_2 > i_3\}.\end{aligned}$$

The number of states in Ω are $\binom{7}{4} + 1$, while the four coordinates may be interpreted as follows:

- i_0 records the number of points fall into interval I_2 in a window of length 7 (at most) as the process evolves in time, and
- i_1, i_2 and i_3 specify the positions of the last three points fall into interval I_2 in the window of length 7 (at most) as the process evolves in time.

Finally, the non-vanishing transition probabilities associated with the transition probability matrix are

- $\Pr(V_t = (i_0 + 1, i_1 + 1, i_2 + 1, i_3 + 1) \mid V_{t-1} = (i_0, i_1, i_2, i_3)) = p_2$, $i_0 = 0, 1, 2$, $i_1 = 1, 2, 3$, $i_2 = 0, 1, 2$, $i_3 = 0, 1$;
- $\Pr(V_t = \alpha_s \mid V_{t-1} = (i_0, i_1, i_2, i_3)) = p_2 + p_3$, $i_0 = 3$, $i_1 = 1, 2, 3$, $i_2 = 0, 1, 2$, $i_3 = 0, 1$;
- $\Pr(V_t = \alpha_s \mid V_{t-1} = (i_0, i_1, i_2, i_3)) = p_3$, $i_0 = 0, 1, 2$, $i_1 = 1, 2, 3$, $i_2 = 0, 1, 2$, $i_3 = 0, 1$; and

- $\Pr(V_t = (i_1 - 2, i_2 + 1, i_3 + 1, 0) \mid V_{t-1} = (i_0, i_1, i_2, i_3)) = p_1$, $i_0 = 0, 1, 2, 3$, $i_1 = 2, 3, 4, 5$, $i_2 = 1, 2, 3, 4$, $i_3 = 0, 1, 2, 3$,

respectively. Taking into account that α_s is the absorbing state, the transition matrix $\mathbf{\Lambda}$ may be written in the following block-form

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}^* & \mathbf{h} \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{5}$$

where $\mathbf{h} = \mathbf{1} - \mathbf{\Lambda}^*\mathbf{1}$ and $\mathbf{\Lambda}^*$ is the matrix $\mathbf{\Lambda}$ after removing the last column and the last row while $\mathbf{0}$, $\mathbf{1}$, are $(s-1) \times 1$ column vectors of zeros and ones, respectively.

Using the partition of $\mathbf{\Lambda}$ (5), we deduce the cumulative probability function of W as follows:

$$\Pr(W = n) = \boldsymbol{\pi}_1^\top (\mathbf{\Lambda}^*)^{n-1}\mathbf{h}, \tag{6}$$

where $\boldsymbol{\pi}_1$ is a $(s-1) \times 1$ column vector that contains all entries of the initial probability vector $\boldsymbol{\pi}_0$ except the last one. Using Eq. 6 we may obtain the following expression for the probability generating function of the random variable W

$$G(z) = z\boldsymbol{\pi}_1(\mathbf{I} - z\mathbf{\Lambda}^*)^{-1}\mathbf{h}.$$

More details on the development of these formulae may be found in Balakrishnan et al. (2009). Replacing appropriately in the last formulae and performing the necessary calculations we take the following recursive scheme for the probability function

$$\begin{aligned} f(n) = {} & p_1 f(n-1) + p_1^2 p_2 f(n-3) + p_1^3 p_2^2 f(n-5) + p_1^4 p_2^2 f(n-6) \\ & +5p_1^4 p_2^3 f(n-7) + p_1^5 p_2^3 f(n-8) - 3p_1^6 p_2^4 f(n-10) - p_1^8 p_2^5 f(n-13) \\ & -10p_1^8 p_2^6 f(n-14) - 5p_1^9 p_2^6 f(n-15) - p_1^{10} p_2^6 f(n-16) \\ & +3p_1^{10} p_2^7 f(n-17) - p_1^{11} p_2^7 f(n-18) + 6p_1^{11} p_2^8 f(n-19) \\ & +10p_1^{12} p_2^9 f(n-21) + 3p_1^{13} p_2^9 f(n-22) - p_1^{14} p_2^{10} f(n-24) \\ & -4p_1^{15} p_2^{11} f(n-26) - 5p_1^{16} p_2^{12} f(n-28) + p_1^{19} p_2^{14} f(n-33) \\ & +p_1^{20} p_2^{15} f(n-35), \end{aligned}$$

where $n > 35$. This scheme is by far faster than Eq. 6, since there is no need to compute high powers of $\mathbf{\Lambda}^*$. For $n \leq 35$, the corresponding probabilities may be calculated by exploiting appropriately $G(z)$, or using Eq. 6.

Appendix B: Derivation of the Joint Probability of $T_{B,1}$ and $T_{B,2}/T_{B,3}$

We have

$$\begin{aligned}
\Pr(T_{B,2} > c_2, T_{B,1} > c_1) &= \Pr(T_{B,2} > c_2 \cap \{T_{B,1} = c_1 + 1 \cup T_{B,1} = c_1 + 1 \cdots \\
&\quad \cup T_{B,1} = c_1 + n\}) \\
&= \sum_{i=c_1+1}^{n} \Pr(T_{B,2} > c_2, T_{B,1} = c_1 + i) \\
&= \sum_{i=c_1+1}^{n} \Pr(T_{B,2} = c_2 + 1 \cup \{T_{B,2} = c_2 + 2 \cdots \\
&\quad \cup T_{B,2} = c_2 + n - k, T_{B,1} = c_1 + i\}) \\
&= \sum_{i=c_1+1}^{n} \sum_{j=c_2+1}^{n} \Pr(T_{B,2} = j, T_{B,1} = i) \\
&= \sum_{i=c_1+1}^{n} \sum_{j=c_2+1}^{n} \Pr(T_{B,2} = j \mid T_{B,1} = i)\Pr(T_{B,1} = i). \qquad (7)
\end{aligned}$$

From Theorem 4.2 from Antzoulakos et al. (2003) we have

$$\begin{aligned}
\Pr(T_{B,2} = j \mid T_{B,1} = i) &= \binom{n}{n-i}^{-1} \sum_{\ell=0}^{n-i+1} \sum_{i_1=0}^{\ell} \sum_{j_1=0}^{\ell-i_1} \sum_{j_2=0}^{i_1} (-1)^{\ell+i_1+j_1-j_2} \\
&\quad \times \binom{n-i+1}{\ell}\binom{\ell}{i_1}\binom{\ell-i_1}{j_1}\binom{i_1}{j_2}\binom{\ell+a-1}{a} \\
&\quad \times \binom{n-i+b}{b},
\end{aligned}$$

where $a = j - i_1 + j_2 - w_p(j_1 + j_2)$ and $b = i - w_p\ell - i_1 - a$, while $\Pr(T_{B,1} = i)$ is provided by the binomial distribution with probability of success 0.5. Substituting these into Eq. 7 provides the required formula (3).

References

Antzoulakos DL, Bersimis S, Koutras MV (2003) On the distribution of the total number of run lengths. Ann Inst Stat Math 55:865–884

Atkinson RW, Anderson HR, Sunyer J, Ayres J, Baccini M, Vonk JM, Boumghar A, Forastiere F, Forsberg B, Touloumi G, Schwartz J, Katsouyanni K (2001) Acute effects of particulate air pollution on respiratory admissions results from APHEA 2 Project. Am J Respir Crit Care Med 164:1860–1866

Balakrishnan N, Koutras MV (2002) Runs and scans with applications. Wiley, New-York

Balakrishnan N, Bersimis S, Koutras MV (2009) Run and frequency quota rules in process monitoring and acceptance sampling. J Qual Technol 41:66–81

Bersimis S, Psarakis S, Panaretos J (2007) Multivariate statistical process control charts: an overview. Qual Reliab Eng Int 23:517–543

Chakraborti S, van der Laan P, van de Wiel MA (2004) A class of distribution-free control charts. J R Stat Soc Ser C 53:443–462

Christodoulakis J, Tzanis CG, Varotsos CA, Ferm M, Tidblad J (2017) Impacts of air pollution and climate on materials in Athens, Greece. Atmos Chem Phys 17:439–448

Frisen M (2008) Financial surveillance. Chichester

Gibbons JD, Chakraborti S (2010) Nonparametric statistical inference, 5th edn. Chapman and Hall, New-York

Jiang X-Q, Mei X-D, Feng D (2016) Air pollution and chronic airway diseases: what should people know and do? J Thoracic Dis 8:E31–E40

Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907

Koutras MV, Bersimis S, Maravelakis PE (2007) Statistical process control using Shewhart control charts with supplementary runs rules. Methodol Comput Appl Probab 9:207–224

Mudway IS, Kelly FJ (2000) Ozone and the lung: a sensitive issue. Mol Asp Med 21:1–48

O'Neill MS, Ebi KL (2009) Temperature extremes and health: impacts of climate variability and change in the United States. J Occup Environ Med 51:13–25

Pan JN, Chen ST (2008) Monitoring long-memory air quality data using ARFIMA model. Environmetrics 19:209–219

Paroissin C, Penalva L, Pétrau A, Verdier G (2016) New control chart for monitoring and classification of environmental data. Environmetrics 27:182–193

Petris G, Petrone S, Campagnoli P (2010) Dynamic linear models with R. Springer, New York

Prado R, West M (2010) Time series: modelling, computation and inference. Chapman and Hall, New York

Qiu P (2018) Some perspectives on nonparametric statistical process control. J Qual Technol 50:49–65

Qiu P, Li Z (2011) On nonparametric statistical process control of univariate processes. Technometrics 53:390–405

Raaschou-Nielsen O, Beelen R, Wang M, Hoek M, Andersen ZJ, Hoffmann B, Stafoggia M, Samoli E, Weinmayr G, Dimakopoulou K, Nieuwenhuijsen M, Xun MM, Fischer P, Eriksen KT, Sørensen M, Tjønneland A, Ricceri F, de Hoogh K, Vineis P (2016) Particulate matter air pollution components and risk for lung cancer. Environ Int 87:66–77

Rosenlund M, Picciotto S, Forastiere F, Stafoggia M, Perucci CA (2008) Traffic-related air pollution in relation to incidence and prognosis of coronary heart disease. Epidemiology 19:121–128

Triantafyllopoulos K (2007) Covariance estimation for multivariate conditionally Gaussian dynamic linear models. J Forecast 26:551–569

Triantafyllopoulos K (2008) Missing observation analysis for matrix-variate time series data. Statist Probab Lett 78:2647–2653

Triantafyllopoulos K, Bersimis S (2016) Phase II control charts for autocorrelated processes. Qual Technol Quant Manag 13:88–108

Triantafyllopoulos K, Harrison PJ (2008) Posterior mean and variance approximation for regression and time series problems. Statistics 42:329–350