Developing an Efficient Surveillance Scheme for Assessing Compliance with Air Quality Standards

Ronit Nirel

NRCSE

Technical Report Series

NRCSE-TRS No. 044

March 11, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

DEVELOPING AN EFFICIENT SURVEILLANCE SCHEME FOR ASSESSING COMPLIANCE WITH AIR QUALITY STANDARDS

Ronit Nirel *
The Hebrew University of Jerusalem, Jerusalem, Israel.

March 11, 2000

Abstract

One of the principle objectives of air quality monitoring is to determine compliance with air quality standards. Since many countries maintain online surveillance of air pollution, sequential detection schemes can be used interactively to control violations of these standards. The Shiryayev-Roberts (SR) scheme is an efficient and flexible procedure for detecting a change in distribution. This work proposes a three-step procedure for implementing the SR scheme to air pollution data. The first step analyzes the distributional properties of the process; the second step reformulates the standard in terms of the respective distribution; and the last step constructs the SR statistic and calibrates the scheme to attain a prespecified rate of false alarms.

The procedure is demonstrated for daily sulfur dioxide (SO_2) data from an air quality monitoring station in Israel. A three-parameter lognormal distribution is fitted to the detrended series, and serial correlations are incorporated through the respective likelihood function. The scheme's parameters are determined by a Monte Carlo study. The results indicate that online surveillance in the considered area would have detected a significant increase in SO_2 levels eleven months earlier than the common once-a-year inspection scheme.

key words: Change-point detection, Shiryayev-Roberts, sulfur dioxide, average run length.

**Corresponding author address: Dr. Ronit Nirel, Department of Statistics, The Hebrew University of Jerusalem, Mt. Scopus, Jerusalem 91905, Israel. E-mail: msronit@mscc.huji.ac.il*

1 Introduction

Ambient air quality standards are promulgated as part of air quality control policy. The standards specify the level of pollutant concentration, the averaging time and the form of the statistic used to evaluate compliance. For example, the US federal 24hr sulfur dioxide (SO_2) standard is set at a daily concentration of 0.14ppm ($365\,\mu\mathrm{g\,m}^{-3}$) not to be exceeded more than once per year.

The level of a pollutant at a given time and location is a random quantity, with associated probability distribution. Despite some awareness of the role of uncertainty and variation in the formulation and surveillance of air quality standards (Schwartz and Siegel, 1970; Vaughan and Russell, 1983), current surveillance of air quality incorporates almost no statistical considerations, but rather treats the level of pollution as a deterministic quantity (Barnett and O'Hagan, 1997). Compliance with air quality standards is generally assessed in retrospect, that is, after a prespecified period has elapsed (e.g., one year, for the above SO_2 standard). Recent works by Carbonez et al. (1999) and Thompson et el. (2000) deal with statistical aspects of the 'retrospective' implementation of standards. Specifically, Thompson et al. (2000) suggest a hypothesis testing framework for a retrospective inspection of compliance with standards. There are situations, however, when a corrective action may or should be taken before, say, one year would have passed. Examples may include failure of a component of a major pollution sources, like a power plant or oil refineries; or an increase in the level of pollution along certain traffic routes due to changes in land usage (e.g., new industrial area, new shopping center). In such cases, compliance with standards can be assessed interactively using statistical process control (SPC) procedures.

SPC concerns monitoring stochastic processes over time and determining sequentially whether the process is stable or whether a change in distribution has occurred. (for introductory to SPC see Wetherill, 1991). Since most countries maintain online monitoring of air pollution, a surveillance scheme can be designed to detect interactively violations of air quality standards. Examples of application of SPC, or change-point, procedures in environmental processes include detection of a change in smoke level due to the switch from coal to oil in England at approximately 1965 (Bennet et al., 1976); change in respiratory conditions (Malachowski et el., 1994); changes in ozone level in Los Angeles due to the opening of a freeway and changes in gasoline (Box and Tiao, 1975); and changes in minimal monthly temperatures (Jandhyala et al., 1999a). Bordignon and Scagliarini (1999) applied SPC procedures to on-line quality control of hourly ozone data. Other applications are described by MacNeill et al. (1991) and Abdullah and Husain (1999). These applications deal with detection of change-points in environmental data in general and not with violations of air quality standards.

The objective of this study is to formulate an interactive surveillance scheme for detection of violations of air quality standards, while taking into account the complex stochastic properties of air pollution data. The complexity of air pollution data stems from the non-stationarity, non-normality and serial correlation inherent in these processes. Our approach accounts for the non-stationarity by detrending the data. The non-normality may be dealt with either directly, by fitting a non-Gaussian distribution to the data, or by appropriate transformation. As for serial correlation, much of the published work on tests for change-point assume that the data are independent and normally distributed. When 'routine' methods are used for structured data, and their underlying assumptions are ignored, the protection against false alarms, for example, is reduced (Yashchin, 1993; McGilchrist and Fraser, 1996). Common remedies include modifications of standard test statistics (Tang and MacNeill, 1993), monitoring residuals from a time-series model (Berthouex et al., 1978; Malachowski et al., 1994) and application of nonparametric resampling methods (Liu and Tang, 1996). We take account of serial correlation through the likelihood function.

The report is organized as follows. Section 2 briefly reviews the Shiryayev-Roberts (SR) scheme and its properties. Section 3 outlines the steps leading from the statutory definition of a standard, through its formulation in terms of distribution functions and the operationalization of the SR test. The proposed method is illustrated for SO_2 data from the Ashdod area, located in the southern coastal plain of Israel. The distributional properties of the series are studied in Section 4 and the SR procedure is then applied to the data in Section 5. The report concludes (Section 6) with a discussion of possible expansions and extensions of the proposed procedure.

2 The Shiryayev-Roberts (SR) Procedure

For the purpose of this report, we formulate the change-point problem as follows. Let $X_1, X_2, \ldots$ be a sequence of variables such that $X_1, \ldots, X_{\nu-1}$ are identically distributed (i.d.) with density f^{μ_0} and $X_\nu, X_{\nu+1}, \ldots$ are i.d. with density f^{μ_s}. Note that independence is not assumed. The time of change, ν, is unknown but we assume that f^{μ_0} and f^{μ_s} are known. Denote by $f_\nu^{\mu_s}(x_1, \cdots, x_n)$ the joint density of $X_1, \ldots, X_n$ when a change occurrs at ν and the pre- and post-change distributions are as above. Let $\mathrm{E}_\nu^{\mu_s}$ be the expectation with respect to $f_\nu^{\mu_s}$. The objective is to raise an alarm as soon as possible if a change has occurred, subject to a restriction on the rate of false alarms. Hence, the detection scheme is a stopping time on the sequence of observations $x_1, x_2, \ldots$. For a stopping time N, the restriction on false alarm may be expressed in terms of the average run length (ARL) when there is no change, $\mathrm{E}_\infty N$. The expected delay may be expressed by the largest ARL to detection, $\sup_{1 \le \nu < \infty} \mathrm{E}_\nu(N - \nu | N > \nu)$. Often, this is equal to the ARL to detection given that a change has occurred at the first observation, $\mathrm{E}_1 N$ (cf. Pollak, 1985).

For any $1 \le k < n$ the likelihood ratio statistic based on $x_1, \cdots, x_n$ for testing $H_0 : \nu = \infty,\ \mu = \mu_0$ against $H_1 : \nu = k,\ \mu = \mu_s$ is

$$\Lambda_k^n(\mu_s) = \frac{f_k^{\mu_s}(x_1, \ldots, x_n)}{f_\infty(x_1, \ldots, x_n)}. \tag{1}$$

Let the sum of likelihood ratios for all runs ending at n be

$$R_n(\mu_s) = \Sigma_{k=1}^n \Lambda_k^n(\mu_s).$$

The Shiryayev-Roberts (SR) detection scheme (Shiryayev, 1963; Roberts, 1966) is defined by the stopping time $N = N(A, \mu_s)$ satisfying

$$N = \min\{n | R_n(\mu_s) \ge A\}. \tag{2}$$

The threshold A is determined so that the average run length to false alarm is at least a prespecified constant B, that is,

$$\mathrm{E}_\infty N \geq B.$$

Results concerning the optimality properties of the SR procedure for homogeneous processes, in terms of speed of detection of an actual change, were obtained by Pollak (1985), Pollak and Siegmund (1985), Srivastava and Wu (1993), Tartakovsky (1995) and Yakir (1997). These works show that the asymptotic properties of the CUSUM and the SR procedures are similar.

To operationalize the SR procedure, a relationship between the specified ARL to false alarm, B, and the threshold A in (2) should be determined. The following results (Pollak, 1987; Pollak and Siegmund, 1991) are useful:

1. For $\nu = \infty$ and N the stopping time of (2), $R_N - N$ is a zero expectation martingale, that is, $\mathrm{E}_\infty(R_N - N) = 0$. We use this property to verify the Monte-Carlo results presented in Table 3.

2. $\frac{\mathrm{E}_\infty N(A,\mu_s)}{A} = C(\mu_s)(1 + o(1))$ where $o(1) \to 0$ as $A \to \infty$.
   Hence, if the ARL to false alarm is not smaller than B, then the value of A is given approximately by $A = B/C(\mu_s)$. The constant $C(\mu_s)$ is determined either through theoretical formulae or by simulation. For a normal distribution, Pollak and Siegmund (1991) suggested the approximation $C(\mu_s) = \exp(0.583\Delta) + o(\Delta^2)$, where $\Delta = \mu_s - \mu_0$. Pollak's results (1987) indicated that for a change in the mean of normal i.i.d. variables, the asymptotic approximations are good, even for low values of A. This result is used in Section 5 to extract the threshold A.

3 Air Quality Standards and the SR Scheme

In this section we outline the steps leading from the statutory specification of air quality standards to a specification of a standard in terms of the parameters of a distribution function, and formulate the respective SR statistic. Air quality standards specify the level of pollutant concentration, the averaging time units and the statistic used to evaluate compliance. In many cases, the standards are formulated, or can be re-formulated, in terms of a percentile of the pollutant distribution which should not exceed a threshold. For example, the Israeli 0.5hr SO_2 standard is attained when the level of 500 μg m^{-3} is not exceeded more than 0.25% of the time of the year. Namely, when the 99.75th percentile of the annual half-hourly SO_2 distribution does not exceed 500 μg m^{-3}.

For a Gaussian pollution process $X_1, X_2, \ldots$ with $X_i \sim N(\mu, \sigma^2)$, the p-th percentile of the marginal distribution x_p can be expressed by

$$x_p = \sigma z_p + \mu$$

where $z_p = \Phi^{-1}(p)$ and Φ is the standard normal cdf. If the respective standard is formulated in terms of a percentile exceeding c, then the standard is violated if $x_p > c$ or $\mu > c - \sigma z_p$. Thus, if σ^2 is known, testing for violation of a standard amounts to testing $H_0 : \mu \leq c^*$

against $H_1 : \mu > c*$, where $c^* = c - \sigma z_p$. Often, two points in the parameter space are chosen and the simple $H_0 : \mu = \mu_0$ (representing some realized state of compliance) against $H_1 : \mu = \mu_s$ is tested, where μ_s is a threshold of interest (e.g., a local standard) satisfying $\mu_0 < \mu_s \leq c^*$. In this formulation, the retrospective approach tests a violation of a standard by studying the observations of, say, the previous year and inferring on a change in the mean. The change-point approach deals with a similar problem by using a sequential test. That is, rather than wait a full year, we test interactively (e.g., every day) whether a change in μ has occurred, at an unknown time ν during the year.

To illustrate the construction of a SR scheme for such a problem, we consider a Gaussian process $\{X_i\}$ which forms a first-order Markov chain. Assume that the process up to time $\nu - 1$ is stationary with mean μ_0 and from time ν it is stationary with mean μ_s. Let the variance σ^2 and the first-order serial autocorrelation ρ be constant over time. For $\nu = \infty$, $X_i \sim N(\mu_0, \sigma^2)$ and $X_i \,|\, X_{i-1} = x_{i-1} \sim N(\,\rho x_{i-1} + \mu_0(1-\rho), \sigma^2(1-\rho^2)\,)$, $i = 2, 3, \ldots$. Omitting the subscript ∞ from f_∞, the joint likelihood of the first n observations is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_n|x_{n-1}, \ldots, x_1) f(x_{n-1}, \ldots, x_1) \\ &= f(x_n|x_{n-1}) f(x_{n-1}, \ldots, x_1) \\ &\ \ \vdots \\ &= f(x_1)\ \Pi_{i=2}^n f(x_i|x_{i-1}) \\ &= (2\pi)^{-n/2} \xi^{-(n-1)} \sigma^{-1} \exp\{-(x_1-\mu_0)^2/2\sigma^2 - \textstyle\sum_{i=2}^n (x_i - \eta_{0,i})^2/2\xi^2\}. \end{aligned}$$

where $\eta_{0,i} = \rho x_{i-1} + \mu_0(1-\rho)$, $i = 2, \cdots, n$; $\xi^2 = \sigma^2(1-\rho^2)$. To write the joint likelihood of the first n observations under a change at time $\nu = k$, $f_k^{\mu_s}(x_1, \cdots, x_n)$, we assume that $f_k^{\mu_s}(x_k|x_{k-1}) = f^{\mu_s}(x_k)$ (that is, X_{k-1} and X_k are not correlated). Denoting $g(x_i) = f(x_i|x_{i-1})$, $\Lambda_k^n(\mu_s)$ of (1) is given by

$$\begin{aligned} \Lambda_k^n(\mu_s) &= \frac{f_k^{\mu_s}(x_1, x_2, \ldots, x_n)}{f_\infty(x_1, x_2, \ldots, x_n)} \\ &= \begin{cases} \frac{f^{\mu_s}(x_1)}{f^{\mu_0}(x_1)} & k = 1,\ n = 1 \\ \frac{f^{\mu_s}(x_k)}{g^{\mu_0}(x_k)}\ \Pi_{i=k+1}^n \frac{g^{\mu_s}(x_i)}{g^{\mu_0}(x_i)} & k = 1, \ldots, n-1,\ n > 1 \\ \frac{f^{\mu_s}(x_n)}{g^{\mu_0}(x_n)} & k = n,\ n > 1. \end{cases} \end{aligned}$$

The statistic $R_n(\mu_s)$ can be extracted recursively by

$$\begin{aligned} R_n &= \textstyle\sum_{k=1}^n \Lambda_k^n \\ &= \frac{g^{\mu_s}(x_n)}{g^{\mu_0}(x_n)} \textstyle\sum_{k=1}^{n-1} \Lambda_k^{n-1} \ + \ \Lambda_n^n \\ &= \frac{g^{\mu_s}(x_n)}{g^{\mu_0}(x_n)}\ R_{n-1} + \frac{f^{\mu_s}(x_n)}{g^{\mu_0}(x_n)}. \end{aligned}$$

For the above Gaussian process we get

$$R_n = \exp\left[\{-(x_n - \eta_{s,n})^2 + (x_n - \eta_{0,n})^2\}/2\xi^2\right]\ R_{n-1} + \frac{\xi}{\sigma} \exp\ \left\{-(x_n - \mu_s)^2/2\sigma^2 + (x_n - \eta_{0,n})^2/2\xi^2\right\}$$

where $\eta_{s,i} = \rho x_{i-1} + \mu_s(1-\rho)$.

So far we have used a simple scenario to illustrate (a) how to express a test for standard violation in terms of a change-point problem, and (b) how to construct the fundamental statistic of the SR scheme. Frequently, however, the pollution process is non-stationary, the measurements are highly skewed and the serial dependence may be of higher order than one. Therefore, statistical testing of standard violation may require preliminary data manipulation (e.g. detrending, transformation). Evidently, such manipulation may be required for both the retrospective and sequential approaches. Any data manipulation requires, in turn, an appropriate reformulation of the standard. It is beyond the scope of this report to deal with this issue in depth. Instead, the following examples illustrate typical difficulties and possible solutions.

- *Seasonal effects.* Most pollutants have seasonal variation. Suppose that this variation can be modeled by a discrete function with k values, namely, that the ith observation of the jth sub-period, x_{ji} $j = 1, \ldots, k$, has mean μ_j. The overall mean is $\mu = \sum_{j=1}^{k} w_j \mu_j$, where $\sum_{j=1}^{k} w_j = 1$. The goal is to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_0 + \Delta$. This translates for the seasonally-adjusted series $\{x^*_{ji}\} = \{x_{ji} - \mu_j\}$ to a test of $H_0 : \mu^*_j = 0,\ j = 1, \cdots, k$ against $H_1 : \mu^*_j = \Delta_j,\ j = 1, \cdots, k$ where μ^*_j is the mean of x^*_{ji} and $\sum w_j \Delta_j = \Delta$. For example, we may choose Δ_j to be proportional to μ_j, that is, $\Delta_j = \frac{\Delta}{\mu}\mu_j$. It is straightforward to obtain the expressions for the corresponding likelihood ratios.

- *Transformation.* In many cases, normality can be attained by a suitable transformation of the data. Consider a lognormal variable X, such that $Y = \log(X - \theta)$ is $N(\mu, \sigma^2)$ (θ is a shift parameter). Suppose that a standard is formulated in terms of x_p, the pth percentile of the distribution of X. Then, $x_p = \theta + \exp\{\sigma z_p + \mu\}$ and a change of $\Delta_x = x^s_p - x^0_p$ in the pth percentile in the original scale corresponds to a change of

$$\Delta_y = \mu_s - \mu_0 = \log \frac{x^0_p + \Delta_x - \theta}{x^0_p - \theta} \tag{3}$$

  in the log-scale.

Once the parametric setup is resolved and the standard is expressed in a correct scale, the final step in the construction of the SR scheme is to determine A. For this we need to extract $C(\mu_s)$. For the normal i.i.d. case, theoretical results are available. For more complex cases, Monte Carlo studies may be useful, as illustrated in Section 5.

4 Analysis of SO_2 Data

The proposed procedure is illustrated for SO_2 data from the Ashdod air quality monitoring station, located in the southern coastal plain of Israel. Daily half-hour measurements were averaged to produce a daily half-hour mean. The 1990-1997 series is displayed in Fig. 1 (n=2854) and the annual distribution in Fig. 2. An increase in level is evident in 1997.

Our next step is to identify sources of non-stationarity in the series. Typically, non-stationarity in air pollution data is attributed to long term trend, seasonality and possibly day-of-the-week variability. To eliminate the trend we choose the period 1994-1996 to represent the 'in-control' status. The post-change period in our analysis is 1997.

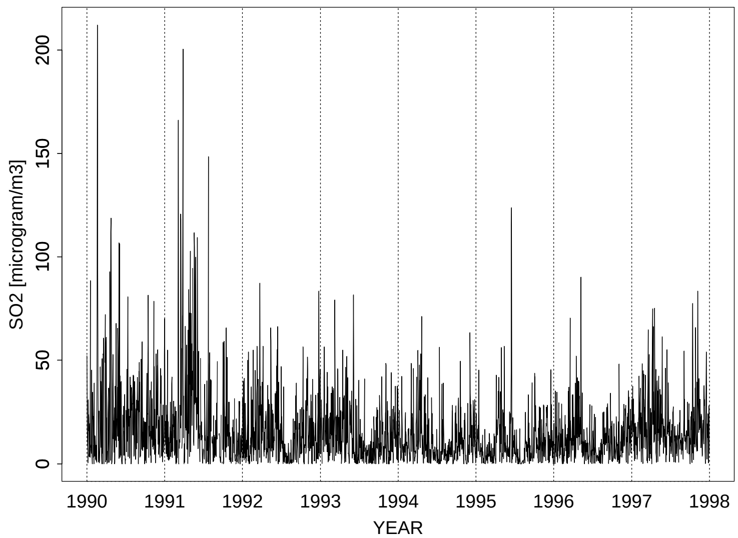

Figure 1: SO_2 series for daily 0.5hr mean at Ashdod, Israel, 1990-1997 (n=2854).

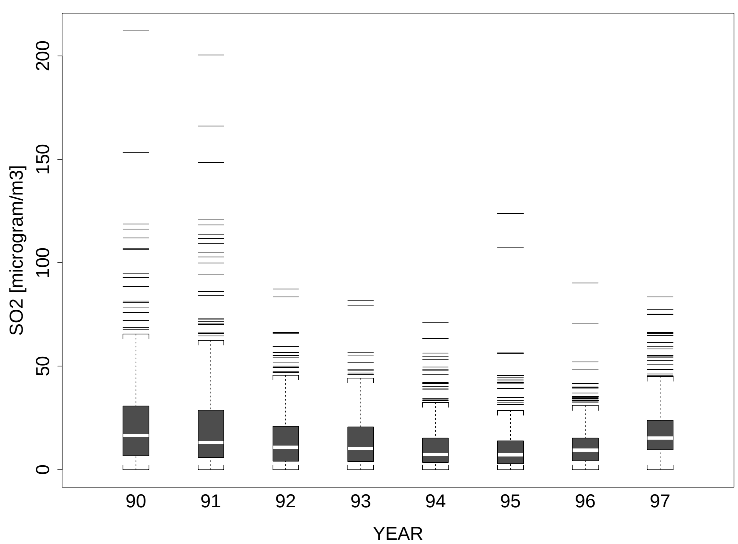

Figure 2: Annual distribution of daily SO_2 averages.

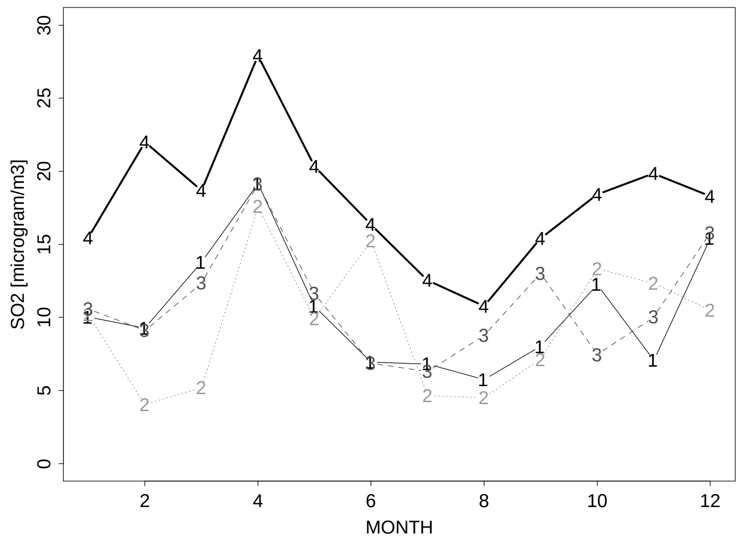

Figure 3: Trimmed mean of daily 0.5hr SO_2 measurements at Ashdod by year and month. The numbers 1-4 on the lines correspond to the years 1994 to 1997. Four percent of the highest and lowest values are truncated.

Some insight into the seasonal variation in the 1994-1997 series is gained from Fig. 3. statistics by year. The figure indicates seasonal variation with high levels during the transitional seasons (April, October-November). This pattern reflects the meteorological conditions and the fuel-management policy in the area.

The rest of the section discusses adjustment for seasonality, estimation of the marginal distribution of the seasonally-adjusted series and the stochastic structure of the series.

Adjustment for Non-stationarity

A moving-average filter of 31 days is used to estimate the seasonality effect. The final estimate is computed as the average of the estimates for the three pre-change years (1994-1996). The upper plot in Fig. 4 shows the estimated effects for the years 1994-1996 and their mean. The lower plot shows the estimate and the original data. The seasonally-adjusted series is obtained by subtracting the seasonality estimate from the raw series.

Marginal Distribution

The distribution of the seasonally-adjusted series is highly skewed (Fig. 5). Either a gamma or lognormal distribution can be fitted to the data. Since estimation under the assumption of a lognormal distribution is simpler, we choose a three-parameter lognormal distribution. A shift parameter is required since the seasonally-adjusted series is centered around zero.

The problem of estimating the shift parameter of the log-normal distribution is well known. Hill (1963) showed that the global maximum likelihood estimate (MLE) of θ is $\min(y_i)$ and that the MLE of the other parameters in this environment are infinite. However, in many situations a local maximum likelihood estimate (LMLE) exists, and maintains the

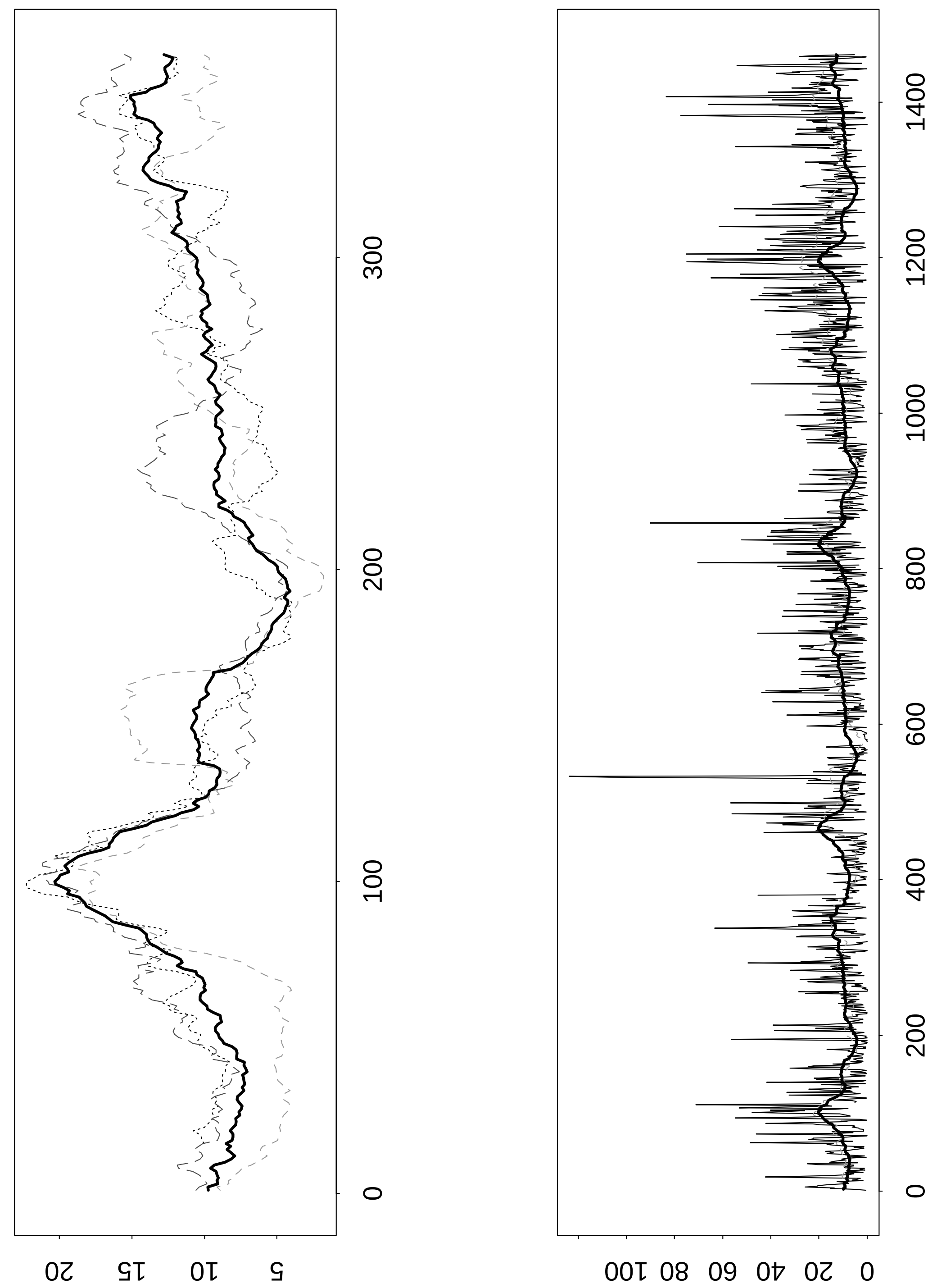

Figure 4: Estimates of the seasonal effect. The upper figure shows the annual estimates for 1994-1996 and their mean (bold line). The lower figure shows the estimate and the raw data for 1994-1997.

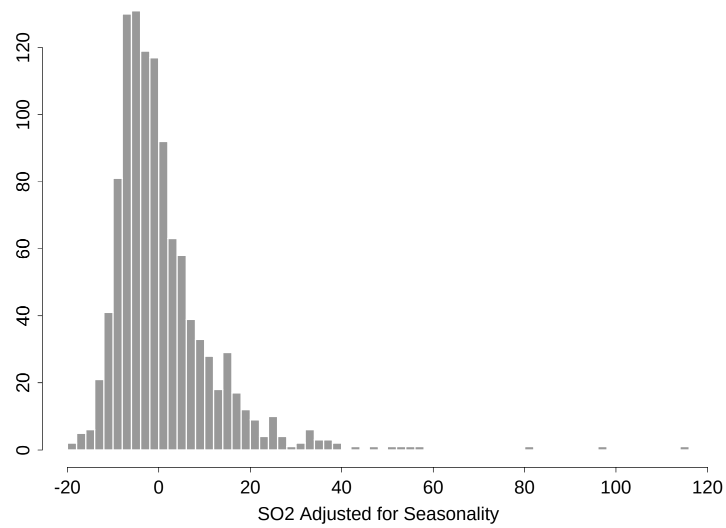

Figure 5: Distribution of the seasonally-adjusted SO_2 series, 1994-1996.

optimal properties of efficiency and consistency of the MLE (Cohen 1951; Harter and Moore 1966; Calitz 1973).

The existence of a LMLE of the shift parameter for the Ashdod series is examined graphically by plotting the profile log-likelihood $L^*(\theta)$ against θ. The profile log-likelihood is extracted by substituting the MLEs of $\hat{\mu}$, $\hat{\sigma}$ in the log-likelihood function:

$$L^*(\theta) = -n\,\{\log(\hat{\sigma}_\theta) + \hat{\mu}_\theta\}, \tag{4}$$

where $\hat{\mu}_\theta = \sum_{i=1}^n \log(x_i^* - \theta)\,/\,n$, $\hat{\sigma}_\theta = \sum_{i=1}^n \{\log(x_i^* - \theta) - \hat{\mu}_\theta\}^2/n$ and $x_1^*, \ldots, x_n^*$ is the seasonally-adjusted series. Once the existence of the LMLE is verified, the estimated value may be found graphically by focusing on a neighborhood of the local maxima. The accuracy of this estimate is usually sufficient in practice. The graphical method of estimation also enables a simple construction of an approximate confidence interval for θ, given by

$$L^*(\hat{\theta}) - L^*(\theta) \leq \chi^2_{(\alpha,1)},$$

where $\chi^2_{(\alpha,1)}$ is the α percentile of the χ^2 distribution with one degree of freedom (Box and Cox, 1964). Graphical estimation of the shift parameter is illustrated in Fig. 6. We find that $\hat{\theta} = -22.25$ ($L^* = -2367.8$), $\chi^2_{(0.95,1)} = 3.84$ and the 95% confidence interval is given by [-24.80, -20.60].

Fig. 7 and Table 1 indicate that the lognormal fit for the pre-change seasonally-adjusted data is fair ($p = 0.033$ for the Shapiro-Wilk test). The estimated shift parameter for 1997 is $\hat{\theta} = -29.50$. Fig. 8 illustrates the fit for 1997: Figs. 8(a) and (c) illustrate the fit for $\hat{\theta} = -29.50$ and Figs. 8(b) and (d) for $\hat{\theta} = -22.25$ (the 1994-1996 estimate).

Table 1 summarizes the estimated parameters for the two periods and two shift estimates. It is seen that (a) the fit of the normal distribution is better for the post-change period, even for $\hat{\theta} = -22.25$, and (b) that for $\hat{\theta} = -22.25$, $\hat{\sigma} = 0.42$ and 0.44 for the pre- and post-change

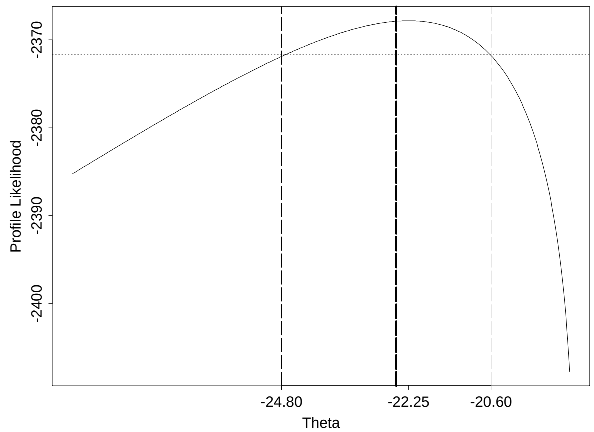

Figure 6: Profile log-likelihood of θ for the seasonally-adjusted Ashdod series and a 95% confidence interval, 1994-1996.

Perod	$\hat{\theta}$	$\hat{\mu}$	$\hat{\sigma}$	Skewness	Kurtosis	Pr<W
1994-1996	-22.25	3.04	0.42	0.21	1.46	0.033
(n=1095)	-29.50	3.36	0.31	0.71	1.74	0.0001
1997	-22.25	3.32	0.44	-0.39	1.43	0.289
(n=365)	-29.50	3.57	0.34	0.07	0.61	0.638

Table 1: Parameters of the pre- and post-change distributions and tests for normality for different values of $\hat{\theta}$. Pr<W is the P-value for the Shapiro-Wilk test.

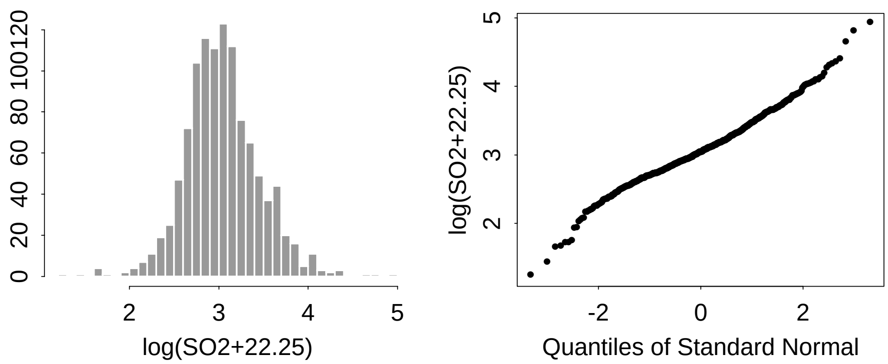

Figure 7: Normal fit for the seasonally-adjusted Ashdod series with shift $\hat{\theta} = -22.25$, 1994-1996.

periods, respectively. We conclude that we may formulate the change-point problem as a change in the location parameter μ only, and assume that the shift and variance parameters are constant over time.

Remarks:

a) The 'retrospective' difference between the 1994-6 and 1997 means is highly significant ($t = 3.89$).

b) Equation (4) represents the profile-likelihood of i.i.d. observations. Since our data are serially correlated, the effect of dependence on the shift estimate was studied. Four subsamples were created by drawing every fourth observation. The resulting set of shift estimates for the four subsamples is {-18.8, -21.0, -24.0, -23.0}, with mean -21.7. This estimate is very close to the estimate obtained for the full dataset (-22.5).

Stochastic Structure

Estimates of the serial correlations for the seasonally-adjusted series are presented in Table 2. The estimated first-order autocorrelations are $r_1 = 0.42$ and 0.28 for the pre- and post-change periods, respectively. For the pre-change period, there is indication for a higher order dependence. In particular, $r_7 = 0.15$ and $r_{14} = 0.11$, suggesting a day-of-the-week effect. The partial correlations for the pre-change period also support a higher order model. For the post-change period, all partial correlations are not significant, except for lag eleven.

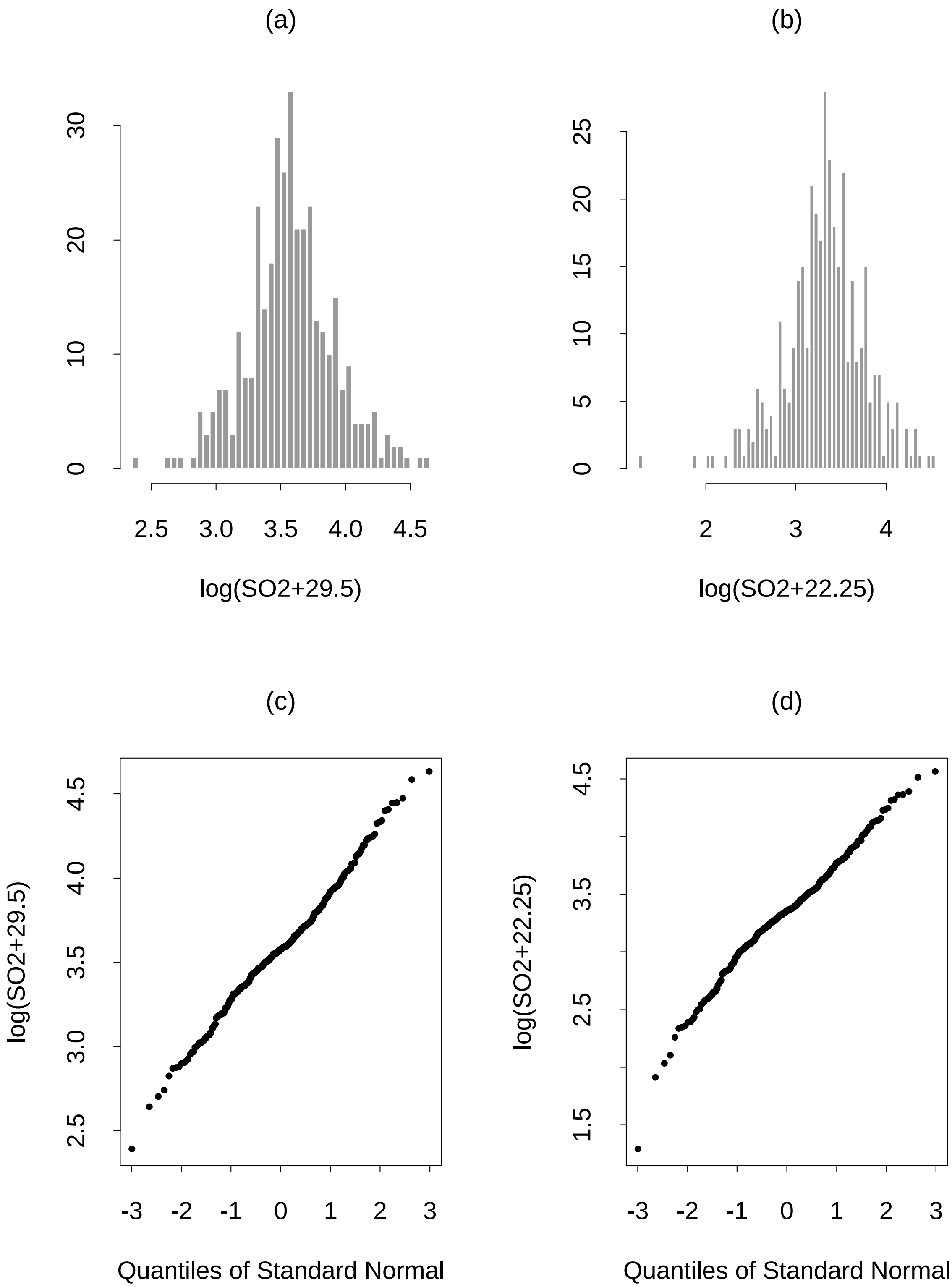

Figure 8: Normal fit for the seasonally-adjusted Ashdod series for 1997 with histograms for (a) $\hat{\theta} = -29.5$, and (b) $\hat{\theta} = -22.25$; (c) and (d) are the respective Q-Q plots.

Lag	Pre-change 1994-1996, n=1094		Post-change 1997, n=364	
	Serial	Partial	Serial	Partial
0	1.0000	* 0.4176	1.0000	*0.2785
1	* 0.4176	-0.0304	*0.2785	-0.0353
2	* 0.1492	* 0.0842	0.0449	-0.0415
3	* 0.1209	0.0341	-0.0352	-0.0766
4	* 0.1055	-0.0600	-0.0932	0.0161
5	0.0130	* 0.1133	-0.0330	0.0451
6	* 0.0834	* 0.0904	0.0356	0.0427
7	* 0.1537	* -0.0865	0.0671	-0.0290
8	0.0355	-0.0150	0.0128	-0.0963
9	-0.0189	0.0482	-0.0905	0.0680
10	0.0436	-0.0436	0.0006	0.0074
11	0.0079	0.0209	0.0103	*-0.1249
12	-0.0080	0.0338	-0.1003	0.0109
13	0.0438	* 0.0706	-0.0309	0.1040
14	* 0.1109	-0.0435	0.0800	0.0126

Table 2: Serial and partial autocorrelations for the pre- and post-change periods. * indicate significance at an $\alpha = 0.05$ level

5 Application of the SR Procedure

The Israeli 24hr SO_2 standard sets a maximal level of 280 μg m^{-3} not to be exceeded at any time. Evidently, this type of standard does not allow for any uncertainty or variability in the pollution process. In addition, the data for Ashdod since 1990 suggest that this standard is quite liberal. Hence, we do not attempt to test for compliance with this standard but rather illustrate the SR procedure for two 'reasonable' values of μ. Assume that the 'pre-change' period (1994-1996) represents a state of compliance with the standard and that we declare the process to be out of compliance whenever a change of 0.5σ in the mean is identified. We further assume that the pre-change seasonally-adjusted Ashdod series is lognormal with $\mu_0 = 3.04$, $\sigma = 0.42$ and $\theta = -22.5$, and that it is stationary first-order Markov chain with $\rho = 0.42$. The post-change distribution is assumed to be lognormal with the same σ, θ and ρ. The problem is to detect a change of 0.5σ in the mean, where the time of change is unknown. Note that for $p = 0.9975$, $z_p = 2.8070$, $y_p = 4.22$ and $x_p = 45.71$, a change of $\Delta_y = 0.21$ in the log-scale corresponds to a change of $\Delta_x = 15.9\,\mu$g m^{-3} in the original scale (see (3)).

Estimates of the threshold A of (2) are obtained by a small Monte Carlo study. The results in Table 3 are based on simulations from a null $N(3.04, 0.42^2)$ and $\mu_s = \mu_0 + 0.5\sigma = 3.25$. The i.i.d. ($\rho = 0$) and first order Markov chain ($\rho = 0.42$) cases are studied for $A = 10, 20, 30, 50, 100$. The length of the simulated series is chosen so that it provides 200 stopping events for each entry in the table. Each time R_n exceeded the threshold A, the

ρ		A 10	20	30	50	100
0.00	$\mathrm{E}_\infty N_A$ (s.e.)	13.52(0.64)	28.48(1.52)	42.42(2.22)	63.22(3.56)	144.28(9.44)
	$\mathrm{E}_\infty\{R(N_A)-N_A\}$(s.e.)	0.53(0.73)	-1.29(1.57)	-1.72(2.30)	4.07(3.86)	-10.19(9.69)
	$C(\mu_s)$	1.35	1.42	1.41	1.26	1.44
0.42	$\mathrm{E}_\infty N_A$(s.e.)	12.61(0.49)	24.54(0.92)	38.93(1.85)	56.57(3.22)	119.96(7.55)
	$\mathrm{E}_\infty\{R(N_A)-N_A\}$(s.e.)	0.50(0.54)	-0.16(0.94)	-2.78(1.87)	4.50(3.34)	-1.12(7.66)
	$C(\mu_s)$	1.26	1.28	1.30	1.13	1.20

Table 3: Monte-Carlo estimates of the parameters of the Shiryayev-Roberts procedure for $\sigma = 0.42$, $\mu_0 = 3.04$, $\mu_s = \mu_0 + 0.5\sigma = 3.25$. Each estimate is based on 200 stopping events.

surveillance is renewed.

The table is divided into two parts, for $\rho = 0$ and $\rho = 0.42$. The first row in each part indicates the estimated ARL to false alarm, $\mathrm{E}_\infty N$, and the standard error of N. For the i.i.d. case, the Monte Carlo estimates are very similar to those given by Pollak (1987, Table 1). For the non-i.i.d. case, the estimated rates are lower than those for the i.i.d. case. The values of $\mathrm{E}_\infty\{R(N)-N\}$ and the standard error of the difference $R(N)-N$ are presented in the second row. We find that for both the i.i.d and correlated cases the Monte Carlo estimates are not significantly different from zero ($\alpha = 0.05$). Estimates of $C(\mu_s)$ are presented in the third row. For the i.i.d. case the estimate is 1.44, and for the correlated case it is 1.20.

Using the results in Table 3, the estimates in Table 1 and theoretical results alluded to in Section 2, the required parameters of the SR for a change in the Ashdod series can be calculated. We estimate that $C(3.25) = 1.20$ and require that a false alarm is not raised earlier than once in ten months, that is $B = 304$ observations. Hence, $A = B/C(3.25) = 304/1.20 = 253$.

Application of the resulting SR scheme to the Ashdod series is shown in Fig. 9. Each of the four pairs of figures (a)-(d) corresponds to one year, with some overlapping. Each time R exceeded 253, the surveillance was re-started. The first alarm was raised around day 500 (around mid-May 1995, Fig. 9(b)) and then about a month later (mid-June 1995). The next alarms (Fig. 9(c)) occurred around mid-May 1996 and then about three months later. The most persistent sequence of alarms was raised a week after the beginning of 1997 (Fig. 9(d)). The mean lag between alarms is 32 days, in accordance with the expected ARL to detection. Note that the change in mean is detected at the beginning of January 1997, while a retrospective approach would have declared that the process is out of compliance only at the beginning of 1998 (assuming that compliance is assessed at the end of each calendar year). In this case the elevated concentrations were caused by malfunction of the oil refineries in the study area. The refineries were closed for repairs only by mid-1999.

6 Discussion

The purpose of this study is to show that it is feasible to use the Shiryayev-Roberts scheme for an interactive surveillance of air quality standards. Once a parametric model has been

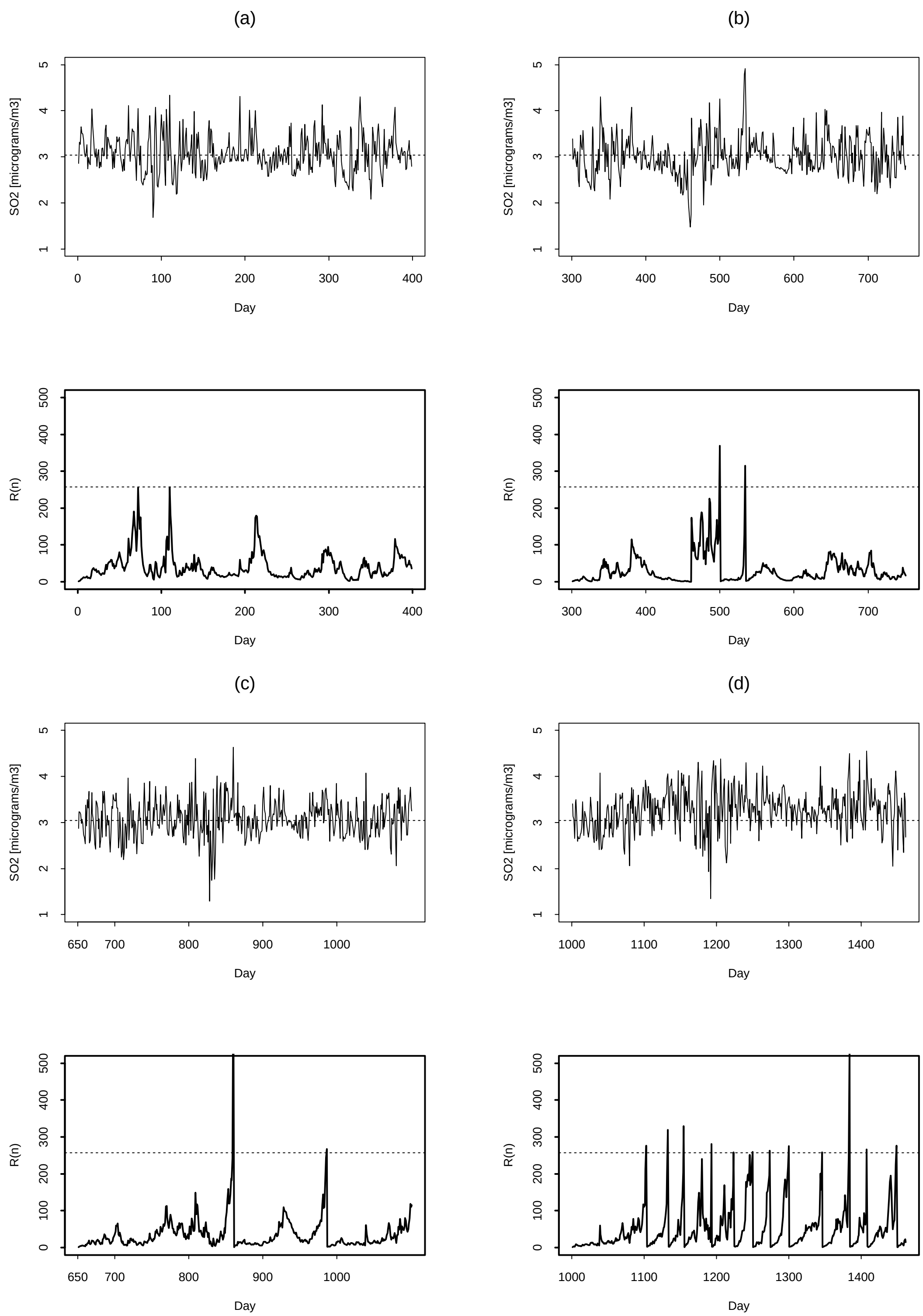

Figure 9: SR scheme applied to the Ashdod series, for the detection of a change of 0.5σ in the mean. Four pairs of figures (a)-(d) correspond to four overlapping segments of the series, starting on 2/1/94 (Day=1) and ending on 12/31/97 (Day=1461). For each pair, the upper figure displays the logarithm of the seasonly-adjusted series plus shift; the horizontal line is drawn at $\mu_0 = 3.04$. The lower figure presents the statistic R; the horizontal line is drawn at $A = 253$.

developed, the application of the scheme is relatively simple even for complex parametric setups. Extension of the procedure to detection of a change in two (or more) parameters is straightforward. A recent review by Jandhyala et al. (1999b) describes applications of change point analysis in linear regression problems, with or without continuity (see also Kim and Siegmund, 1989). Such procedures may be used to detect change in non-stationary series. Future work may compare the parametric scheme to a nonparametric procedure, suggested by Gordon and Pollak (1995). Other extensions may consider variants of the SR scheme when the post-change distribution is unknown (Pollak, 1985) or when the initial level is unknown (Pollak and Siegmund, 1991).

A major issue to be further addressed is the formulation of the statutory standard in terms of parameters of the pollution process. Such formulation is required for any statistical procedure which tests for compliance with regulatory standards, whether interactive or retrospective.

Acknowledgements

The author is grateful to Mary Lou Thompson for her valuable suggestions and comments, and to Moshe Pollak, Paul Sampson and Peter Guttorp for most helpful discussions. Thanks are also due to Doron Lahav from the Ashdod-Hevel Yavne Association of Towns for the Protection of Environment Quality for providing the data.

Most of this work was carried out while the author was a visitor at the National Research Center for Statistics and the Environment, University of Washington.

References

[1] Abdullah, M. and Husain, N. (1999). On the application of change-point detection to environmental data. SESS-TIES conference 23-27 August, 1999. Athens, Greece. `http://www.stat-athens.aueb.gr/esess/Session13.htm`.

[2] Barnett, V. and O'Hagan, T. (1997). *Setting Environmental Standards: The Statistical Approach to Handling Uncertainty and Variation.* Chapman and Hall, London.

[3] Bennet, R. J., Campbell, W. J. and Maugham, R. A. (1976). Changes in atmospheric pollution concentrations. In Brebbia, C. A. (ed): *Mathematical Models for Environmental Problems.* Pentech Press, 221–235.

[4] Berthouex, P. M., Hunter, W. G. and Pallesen, L. (1978). Monitoring sewage treatment plants: Some quality control aspects. *J. Quality Tech.*, **10**, 139–149.

[5] Bordignon, S. and Scagliarini, M. (1999). Monitoring algorithms for detecting changes in ozone concentrations. Preprint.

[6] Box, G. E. P. and Cox, D. R. (1964). Analysis of transformations (with discussion), *J. Roy. Statist. Soc.*, **B26**, 211–246.

[7] Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with application to economic and environmental problems. *J. Amer. Statist. Assoc.*, **70**, 70–79.

[8] Calitz, F. (1973). Maximum likelihood estimation of the parameters of the three-parameters lognormal distribution - a reconsideration. *Austral. J. Statist.*, **15,** 185–190.

[9] Carbonez, A., El-Shaarawi, A. H. and Teugels, J. L. (1999). Maximum microbiological contaminant levels. *Environmetrics*, **10**, 79–86.

[10] Cohen, A. C. (1951). Estimating parameters of logarithmic- normal distributions by maximum likelihood. *J. Amer. Statist. Assoc.*, **46,** 206–212.

[11] Gordon, L. and Pollak, M. (1995). A robust surveillance scheme for stochastically ordered alternatives. *Ann. Statist*, **23**, 1350–1375.

[12] Harter, H. L. and Moore, A. H. (1966). Local maximum likelihood estimation of the parameters of three-parameter lognormal populations from complete and censored samples. *J. Amer. Statist. Assoc.*, **61,** 842–851.

[13] Hill, B. M. (1963). The three parameter lognormal distribution and Bayesian analysis of point source epidemic. *J. Amer. Statist. Assoc.*, **58,** 72–84.

[14] Jandhyala, V.K., Fotopoulos, S.B. and Evaggelopoulos, N. (1999a). Change-point methods for Weibull models with applications to detection of trends in extreme temperatures. *Environmetrics*, **10**, 547–564.

[15] Jandhyala, V.K., Zacks, S. and El-Shaarawi, A. H. (1999b). Change-point methods and their applications: contributions of Ian MacNeill. *Environmetrics*, **10**, 657–676.

[16] Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, **76**, 409–423.

[17] Liu, R. Y. and Tang, J. (1996). Control charts for dependent and independent measurements based on bootstrap methods. *J. Amer. Statist. Assoc.*, **91**, 1694–1700.

[18] MacNeill, I. B., Tang, S. M. and Jandhyala, V. K. (1991). A search for the source of the Nile's change-points. *Environmetrics*, **2**, 341–375.

[19] Malachowski, M. S., Levine, S. P., Herrin, G., Spear R.C., Yost Z. and Yi, Z. (1994). Workplace and environmental air contaminant concentrations measured by open path Fourier transform infrared spectroscopy: a statistical process control technique to detect changes from normal operating conditions. *Air and Waste*, **44**, 673–682.

[20] McGilchrist, C. A. and Fraser, R. (1996). A CUSUM sampling inspection scheme for sewage sludge. *Environmetrics*, **7**, 391–400.

[21] Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist*, **13**, 206–227.

[22] Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist*, **15**, 749–779.

[23] Pollak, M. and Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, **72**, 267–280.

[24] Pollak, M. and Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Ann. Statist*, **19**, 394–416.

[25] Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

[26] Schwartz, S. and Siegel, G. B. (1970). Models for and constraints on decision making. In Atkisson, A. and Gaines, R.S. (eds.): *Air Quality Standards*. Charles E. Merrill, Ohio, 25–51.

[27] Shiryayev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.*, **13**, 22–46.

[28] Srivastava, M. S. and Wu, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts procedures for detecting a shift in the mean. *Ann. Statist*, **21**, 645–670.

[29] Tang, S. M. and MacNeill, I. B. (1993). The effect of serial correlation on tests for parameter change at unknown time. *Annals of Statist.*, **21**, 552–575.

[30] Tartakovsky, A. (1995). Asymptotic properties of CUSUM and Shiryayev's procedures for detecting a change in a nonhomogeneous Gaussian processes. *Math. Meth. Statist*, **4**, 389–404.

[31] Thompson, M. L., Sampson, P. D., Guttorp, P. and Caccia, D. C. (2000). A statistical hypothesis testing formulation of current ozone environmental regulatory standards. Preprint.

[32] Vaughan, W. J. and Russell, C. S. (1983). Monitoring point sources of pollution: answers and more questions from statistical quality control. *Amer. Statist.*, **37**, 476–487.

[33] Wetherill, G. B. (1991). *Statistical Process Control: Theory and Practice*. Chapman and Hall, London.

[34] Yakir, B. (1997). A note on optimal detection of a change in distribution. *Ann. Statist*, **25**, 2117–2126.

[35] Yashchin, E. (1993). Performance of CUSUM control schemes for serially correlated observations. *Technometrics*, **35**, 37–52.