Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Atmospheric Environment 40 (2006) 6528–6540
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case study from the Canadian Arctic

Emma F. Eastoe[a,*], Crispin J. Halsall[b], Janet E. Heffernan[a], Hayley Hung[c]

[a] *Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster LA1 4YF, UK*
[b] *Department of Environmental Science, Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK*
[c] *Meteorological Services of Canada (MSC), Environment Canada, 4905 Dufferin Street, Downsview, Ont., Canada M3H 5T4*

Received 16 March 2006; received in revised form 23 May 2006; accepted 25 May 2006

Abstract

The data sets of four semi-volatile organic compounds (phenanthrene, PCB-28, p,p'-DDE and α-endosulfan) from a multi-year Canadian Arctic air monitoring programme were examined to test the effect of both including and removing censored data (i.e. data that fall below analytical detection limits) on time-trend models. Two approaches were taken with the data, one that included all censored values, known as a survival analysis, and the other with censored values replaced with a fixed constant, referred to here as a replacement analysis. Initially, the results from the time-trend models (depicting seasonality and year-on-year trends) from the two analyses, where replacement involved a small amount of data that fell below instrumental detection limits, showed very few differences. This was effectively due to the small quantity of censoring apparent in each of the data sets (the data sets of 2 compounds had $<10\%$ censoring). However, when the degree of censoring was artificially increased to 50% for two of the compounds (phenanthrene and p,p'-DDE), differences in modelled trend results were evident. By comparing the results of the trend models fitted under both survival and replacement analyses with these highly censored data sets to the actual observed data, it was evident that the survival analysis produced time series models that were far more robust given the quantity of censoring. The application of a Kaplan–Meier (K–M) estimator as a diagnostic tool confirmed the survival analysis approach for producing more robust trend models for both compounds. As a result, we recommend survival analysis and the retention of all censored data within a given data set and this justifies the current approach of retaining all censored data within the Canadian arctic air databases. Blank correction of these types of databases and/or simple exclusion of censored data, could confound time-series analysis.

Keywords: Semi-volatile organic compounds; Pollutants; Censor; Database; Kaplan–Meier plot; Arctic

1. Introduction

An important aspect of air pollution in the arctic atmosphere is the occurrence of semi-volatile organic compounds (SVOCs) that include certain

[*] Corresponding author. Tel.: +44 1524 593327.
E-mail address: e.eastoe@lancaster.ac.uk (E.F. Eastoe).

1352-2310/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2006.05.073

pesticides, chemicals of industrial origin and by-products of combustion. These pollutants are present in remote environments primarily through long-range atmospheric transport (e.g. MacDonald et al., 2000; Hung et al., 2005) and in many cases act as bioaccumulative toxic chemicals that may occur at significant levels in arctic biota (e.g. AMAP, 2004). The Canadian Northern Contaminants Program (NCP) has established a long-term air sampling campaign that has collected atmospheric data on these pollutants since the early 1990s, primarily from Alert, a scientific research, military and meteorological station located on Ellesmere Island, Nunavut, in the Canadian High Arctic.

Data from this site provide one of the longest time-series for atmospheric SVOCs, providing valuable information with which to examine the occurrence and behaviour of these chemicals and assess contamination of the arctic environment. Recently, attempts have been made to model the underlying trend in atmospheric concentrations for SVOCs, whereby the short-term temporal dependence or 'seasonality' in the data is identified and then removed (Hung et al., 2001; Hung et al., 2002, Becker et al., 2006). The presence or lack of a year-on-year trend can then be used to set abatement policies in source regions and/or determine whether international protocols are having an impact on levels in the remote atmosphere.

Difficulties frequently occur when measuring SVOCs in remote environments because the sample chemicals are usually present at very low concentrations (e.g. in the range of $\mathrm{pg\,m^{-3}}$). For air samples taken systematically over the course of a year, the analytical methodology may be unable to measure the true quantity of the chemical that is present in the atmosphere. As a result, there are a large number of samples where concentrations are categorised or banded according to the amount of chemical present and the degree of confidence in the analytical methodology. In this study the treatment of these data is considered by utilising an approach from a branch of statistics called survival analysis (e.g. Lawless, 2003; Helsell, 2005). This allows, with a degree of confidence, utilisation of data that may otherwise be omitted from a valuable database. Such omissions can reduce the effectiveness of time-series analysis and other statistical interpretations.

Currently, within the NCP arctic air programme, samples are not blank corrected, whereby the average mass of an analyte determined from multiple field blanks is subtracted from each weekly air sample. Instead there are two censoring levels—the method detection limit (MDL) and the instrumental detection limit (IDL) expressed as air concentrations. The MDL is defined as the average field blank ($\bar{x}$) plus 3 times the standard deviation around that average ($\bar{x}+3\mathrm{SD}$). A typical observation must therefore lie in one of three bands; $\leqslant$IDL, $>$IDL but at least as low as the MDL or $>$MDL. The IDL is the true censoring level, as this is the level below which the instruments cannot measure at a known confidence level. However, reported weekly air concentrations do occur that fall below the IDL due to the aspirated air volume for that sampling week being greater than the average air volume used to derive the IDL. Data in the middle band ($>$IDL$<$MDL) are treated with some care because, whilst the readings are exact, they are small compared to the noise in the data. It is generally believed that these measurements could just be noise and so they are also treated as being unknown in the same way as those observations that are $<$IDL. Values below the MDL are referred to as *censored* data. Current practice is to 'flag' data that fall below the MDL, and to replace those few values that fall below the IDL by a fixed constant, currently $\frac{2}{3}$ IDL (Hung et al., 2002; Becker et al., 2006). These are then treated as if they were exact observations. Censored data (i.e. $<$MDL) have been included in recent time-series modelling of the Arctic data (e.g. Hung et al., 2001, 2002).

The aim of this study is to use standard methodology from survival analysis to improve the use of censored data and assess the effect of retaining censored data on time series analysis; the latter carried out using a log-normal distribution time-series model. To mimic the effect of blank correction and/or possible deletion of censored data, a replacement analysis was pursued where values that fell below the MDL were simply replaced with a fixed value ($\frac{2}{3}$ MDL). Furthermore, the level of censoring was also artificially increased (by 50%) for selected compounds, and the results compared to the survival methodology where all censored data were retained. Four semi-volatile organic pollutants (phenanthrene, 1,1-dichloro-2,2-*bis*(*p*-chlorophenyl)ethene (p,p'-DDE), α-endosulfan and polychlorinated biphenyl 28 (PCB-28) were considered, since each of these chemicals have air data which fall into the censoring categories outlined earlier and are representative of chemical classes that are targeted within this arctic programme.

2. Survival analysis

Survival analysis provides a statistical method for the treatment of data containing censoring, whereby censoring implies that some (or all) of the data points are not observed exactly. The term 'survival' originates in medical statistics, where data are mainly patient survival times, for which these methods were primarily developed. However the methods may be useful for any data set containing censoring. For the air data considered here, censoring is due to the occurrence of analytes in blank sample media (i.e. MDL) and the inability of instruments to detect extremely low levels of a chemical (i.e. IDL) and is commonly termed *left censoring*.

Suppose that a sequence of weekly observations of any given chemical is denoted by y_i and that it comprises a mixture of exact and censored observations. The subscript i denotes the week of observation. The value, C_i, below which the observation y_i cannot be detected exactly is called the *censoring level*. The censoring level is allowed to change over time, since C_i in this study is the MDL, which varies across years and a function of the occurrence and quantity of analyte on the blank media. The indicator function, δ_i, defined below, is attached to each observation as a mathematical pointer to denote whether or not an observation is censored. It takes value 1 if the associated observation is exact and 0 if it is censored. If there are n observations then for $i = 1, \ldots, n$:

$$\delta_i = \begin{cases} 1 & \text{if } y_i > C_i, \\ 0 & \text{if } y_i \leqslant C_i. \end{cases} \tag{1}$$

For these type of pollutant databases, logged observations are assumed to follow a Normal (μ, σ^2) distribution, with mean (μ) and standard deviation (σ) parameters (e.g. Leister and Baker, 1994). Let $x_i = \ln(y_i)$ be the (natural) log of the observations, then x_i are normally distributed with density function $(f(x_i))$;

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right\}. \tag{2}$$

Explanatory variables are included by modelling the parameters as functions of the desired covariates, so the parameters are indexed by i. Such models require careful choice of the functional form assumed for the parameters.

All models are fitted by classical likelihood inference (e.g. Lindsey, 1995; Knight, 2000), where maximisation of a likelihood function delivers estimates of the parameters and associated standard errors, reflecting uncertainty in the estimation. The two analyses (survival and replacement) using the Normal distribution model of Eq. (2) to the logged data, each used different likelihood functions. For the replacement analysis, which assumes that after replacement of values below the censoring level, all data can be treated as exact, the likelihood (L) was

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} f(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right\}, \end{aligned} \tag{3}$$

where n is the number of observations, $\mu = (\mu_1, \ldots, \mu_n)$ and $\sigma = (\sigma_1, \ldots, \sigma_n)$. For left censored data, the true (but uncensored) value is known only to be less than, or equal to, the censoring level, C_i. The information that censored data contributes to the likelihood is therefore the probability of observing a value *less than or equal to* C_i. The survival analysis likelihood (L_1) is then

$$L_1(\mu, \sigma) = \prod_{\delta_i = 1} f(x_i) \prod_{d_i = 0} F(C_i), \tag{4}$$

where F is the cumulative distribution function (CDF) associated with the density in Eq. (2). The CDF is defined as $F(x) = Pr(X \leqslant x)$, which is the probability of observing a censored value. Thus, the likelihood is the product of the density functions of the exact observations and the CDFs of the censored observations.

The fitted model under the survival analysis was obtained by placing all the air data, including censored values (i.e. <MDL), in L_1 of Eq. (4) and maximising with respect to the parameters μ_i and σ_i to obtain their maximum likelihood estimates (MLEs). In the replacement method, the model was fitted using the database but with values below the MDL simply replaced with $\frac{2}{3}$ MDL and using L from Eq. (3) instead of L_1.

2.1. The Kaplan– Meier estimator

The Kaplan–Meier (K–M) estimator can be used to estimate the survivor function, $S(x) = 1 - F(x)$, of the standardized residuals of each model fitted using either Eqs. (3) or (4) in order to assess the goodness of fit of that model. For the model with parameters μ_i and σ_i fitted to the data, x_i, standardized residuals, z_i, are given by

$$z_i = \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i}, \tag{5}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are parameter estimates. The estimated survivor function for the z_i is compared to the survivor function for the standard Normal distribution, since the estimated distribution should be close to the standard Normal distribution for a well-fitting model, with the standard normal survivor plot within the 95% confidence interval band of the K–M estimate.

The K–M estimator is most easily described in terms of right censored data. The air data in this case is left censored; however left censoring can be turned into right censoring simply by considering the negative residuals. On negation, the censored observations, previously the lowest data points, become the highest. The symmetry of the Normal distribution ensures no mathematical complications arise in doing this. Under the null hypothesis that the fitted model is sufficient, the negative residuals should also follow a standard Normal distribution. A right censored observation is known only to take a value at least as big as the censoring level. If x_i are the logged observations and C_i the censoring levels, then γ_i is the indicator function for the right censored negative observations, where

$$\gamma_i = \begin{cases} 0 & \text{if } x_i > C_i, \\ 1 & \text{if } x_i \leqslant C_i. \end{cases} \tag{6}$$

Let x_j be equal to an exact observation and let $r(x_j)$ be the number of observations which are either exactly observed, or are censored at a value, at least as large as x_j. If d_j denotes the number of observations exactly equal to x_j, the K–M estimate of the survivor function at x_j is

$$\hat{S}(x_j) = \prod_{i:x_i \leqslant x_j} \frac{r(x_i) - d_i}{r(x_i)}. \tag{7}$$

Since the K–M estimates the survivor function it is a decreasing function taking values between 0 and 1. Note also that Eq. (7) is only calculated at actual observations (x_j), so the estimate is a step function. Further details of the K–M estimator are provided by Hougaard (2000).

3. Methods

Table 1 provides information about the four compounds selected for this study which consisted of two pesticides (p,p'-DDE and α-endosulfan), a low molecular weight polycyclic aromatic hydrocarbon (PAH) (phenanthrene) and a polychlorinated biphenyl (PCB-28). For each compound, weekly air concentrations over the period January 1992 to December 1998 were available from Alert, although the phenanthrene data set extended to October 2000. All the data sets have missing observations which can be ignored since they are due to equipment failure/sample loss, and can be assumed to be random. Further details of the high-volume sampling procedures, analytical details and quality controls can be found in Fellin et al. (1996) and subsequent publications (Halsall et al., 1998; Hung et al., 2001, 2002). Field blanks (consisting of blank sample media, subject to the same handling and transport procedures as the samples) were collected each month, resulting in 12 blanks for each year which were used to derive the MDLs for each analyte.

Fig. 1 illustrates the data sets for the four compounds, depicting the censored data as well as the MDL and IDL values; none of the observations

Table 1
Physical–chemical properties and description of contaminants selected for this study and routinely detected in Arctic air

Contaminant (CAS number)	Molecular weight (g mol^{-1})	Vapour pressure* (Pa)	Chemical 'family'
Phenanthrene (85-01-8)	178.24	0.113[a]	Polycyclic aromatic hydrocarbon (PAH)
PCB-28 (7012-37-5)	257.5	0.0269[b]	Polychlorinated biphenyl (2,4,4′-trichlorobiphenyl)
p,p'-DDE (72-55-9)	319.0	0.0034[c]	metabolite of p,p'-DDT (organochlorine pesticide)
α-Endosulfan (959-98-8)	406.9	0.0044[c]	Currently used pesticide (CUP)

*Sub-cooled liquid vapour pressure(P_L) at 298 K.
[a] Mackay et al. (1992).
[b] Li et al. (2003).
[c] Shen and Wania (2005).

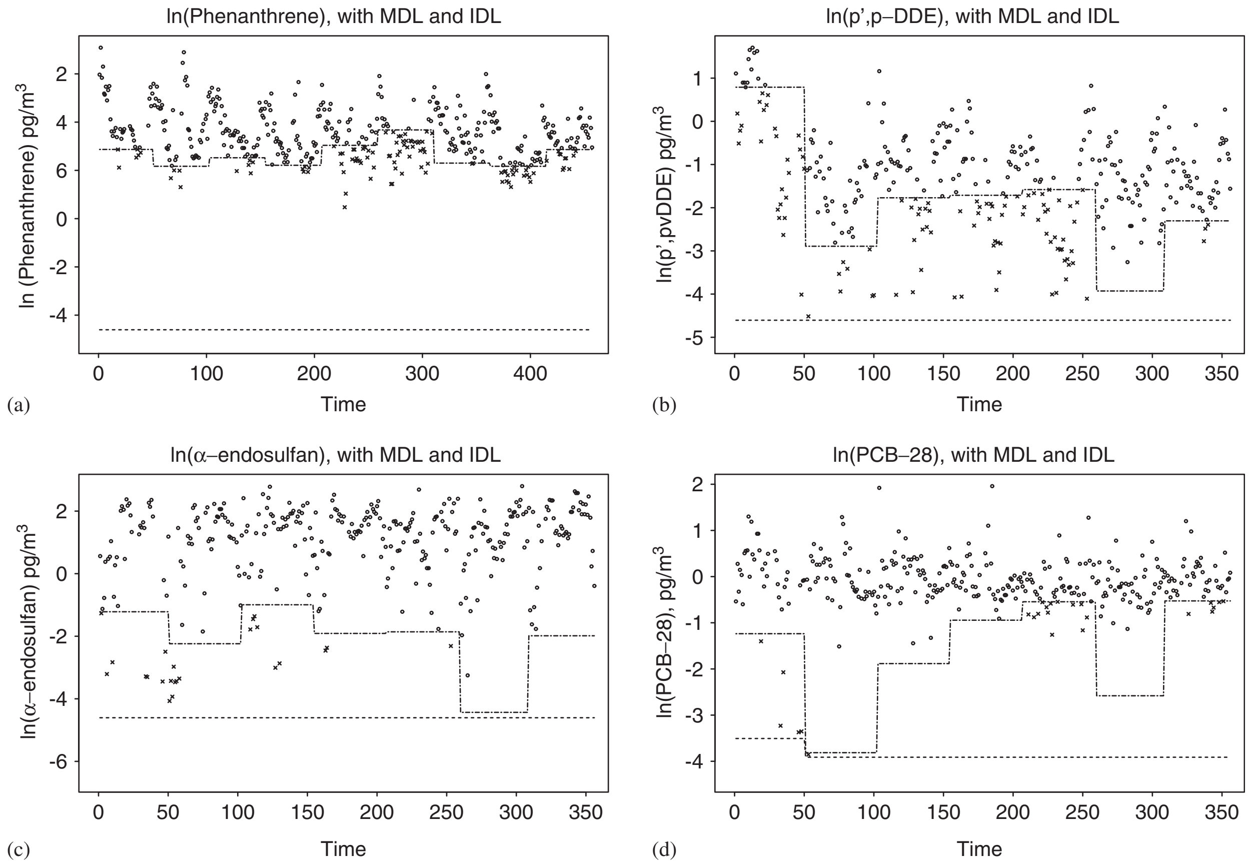


Fig. 1. Time-series of weekly concentrations for (a) phenanthrene, (b) p,p'-DDE, (c) α-endosulfan and (d) PCB-28 measured at Alert, Nunavut, Canada (note the log-scale of the y-axis). The dash-dot line indicates the MDL and the dashed line indicates the IDL. Crosses denote censored data and circles exact data.

were at the IDL and so all censored observations were found to lie in the band ($>$IDL$<$MDL). The phenanthrene data set had 25% censoring, p,p'-DDE had 31%, α-endosulfan 7% and PCB-28, 8%. The censoring often appeared in clusters, due to the lowest values occurring at the same time due to a reduction in concentrations associated with seasonal behaviour in atmospheric concentrations, e.g. Hung et al. (2005).

Two analyses (survival and replacement) were carried out on each of the data sets. In the latter, censored observations that fell below the MDL were simply replaced by $\frac{2}{3}$ MDL and all data were assumed exact. This scenario could arise if a data analyst decided to remove all censored data from the data set. In both analyses, the model selection procedure and diagnostics were the same, where the diagnostics—consisting of residual plots and a K–M estimate for the survival function of residuals—were used to compare the goodness of fit of the trend model (outlined below) to the data using the two methods. A further set of survival and replacement analyses was also carried out on phenanthrene and p,p'-DDE to determine the robustness of the two methods to increased levels of censoring. A hypothetical MDL, selected to censor 50% of the data, was applied to both data sets and survival and replacement analyses were carried out. The goodness of fit of the trend model under this more extreme level of censoring was judged by comparison of the model fit to the data censored at the observed MDL.

3.1. Modelling long-term trends

To model year-on-year trends in these data, any confounding seasonality must first be removed. To do this the parameters of the log-normal distribution are allowed to vary over time. Restricting attention to linear functions of temporal covariates allows exploitation of the widely documented techniques of generalized linear modelling (e.g. Dobson, 1990; McCullagh and Nelder, 1983).

For each compound data set, a naive model with time-constant parameters was first fitted, before attempting more complicated models with parameters which are linear functions of temporal covariates for comparison. Seasonality was modelled by allowing sinusoidal terms in the model parameters, μ and σ; utilised, for example, in the rainfall models of Coe and Stern (1982). Up to three pairs of sinusoids (annual, biannual and quarterly cycles) were allowed. In selecting the best fitting model simplicity was paramount, thus, if there was no evidence to include a sinusoidal pair, then sinusoids with smaller periods were automatically excluded. On removing the seasonal aspect from the data, both parameters were tested for the inclusion of the linear year-on-year time trend. More sophisticated time-trend models have been utilized on the Alert data set for both PCBs and organochlorine (OC) pesticides using Digital Filtration (Hung et al., 2001, 2002) and for PAHs using dynamic harmonic regression (DHR) (Becker et al., 2006), where the latter allowed evolution of multiple sinusoids in the data. In the current study however, the application of a time series model was used to examine the importance of the survival analysis in comparison to the replacement method, and not necessarily to compare to the previous time-series analyses outlined above.

A range of log-normal trend models, each including different covariates in the mean (μ) and standard deviation (σ) parameters, were fitted to the data using both the survival and the replacement analyses, with the results of the most sophisticated model utilised in this study (i.e. the model with most covariates). This model is outlined below in Eqs. (8) and (9), where the logged data (x_i) are normally distributed according to $x_i \sim N(\mu_i, \sigma_i^2)$, where μ_i is given by

$$\begin{aligned}\mu_i = \alpha_1 &+ \alpha_2 \cos\left(\frac{2\pi w_i}{52}\right) + \alpha_3 \sin\left(\frac{2\pi w_i}{52}\right) \\ &+ \alpha_4 \cos\left(\frac{4\pi w_i}{52}\right) + \alpha_5 \sin\left(\frac{4\pi w_i}{52}\right) \\ &+ \alpha_6 \cos\left(\frac{8\pi w_i}{52}\right) + \alpha_7 \sin\left(\frac{8\pi w_i}{52}\right) + \alpha_8 s_i \end{aligned} \qquad (8)$$

and σ_i as

$$\begin{aligned}\ln(\sigma_i) = \beta_1 &+ \beta_2 \cos\left(\frac{2\pi w_i}{52}\right) + \beta_3 \sin\left(\frac{2\pi w_i}{52}\right) \\ &+ \beta_4 \cos\left(\frac{4\pi w_i}{52}\right) + \beta_5 \sin\left(\frac{4\pi w_i}{52}\right) \\ &+ \beta_6 \cos\left(\frac{8\pi w_i}{52}\right) + \beta_7 \sin\left(\frac{8\pi w_i}{52}\right) + \beta_8 s_i, \end{aligned} \qquad (9)$$

where w_i denotes the week and s_i the *normalised* week of observation. The normalisation transforms the time scale to the range (0,1) by dividing the week index (w_i) by the total number of weeks, that is

$$s_i = \frac{w_i}{\max_i(w_i)}. \qquad (10)$$

The vectors $\alpha = (\alpha_1, \ldots, \alpha_8)$ and $\beta = (\beta_1, \ldots, \beta_8)$ are parameters to be estimated and are referred to as the covariate coefficients. Modelling the log of the standard deviation parameter ensures positivity and the best fitting models are chosen by forward selection (McCullagh and Nelder, 1983) using a significance level of 1%. Inclusion of the normalized week covariate (s_i) in the final model implies statistical evidence of a linear time trend. The size and type (positive or negative) of trend is determined by the covariates, α_8 (for the mean) and β_8 (for the standard deviation), which can also be used to determine rate of change.

4. Results and discussion

4.1. Survival analysis

For the four compound data sets all observations below the MDL were treated as being censored at the MDL but all data were retained. Likelihood L_1 from Eq. (4) was used to fit the trend model for each compound described in the previous section. The MLEs of the parameters of the best fitting model for each compound and the associated estimated standard errors are displayed in Table 2. Phenanthrene, p,p'-DDE and α-endosulfan displayed evidence of annual and biannual seasonal cycles, whereas PCB-28 displayed only an annual cycle. Phenanthrene was the only compound to show evidence of a declining year-on-year mean trend, although its standard deviation remained constant over the period studied. The standard deviation also remained constant for PCB-28, whereas the standard deviation of α-endosulfan displayed up to

Table 2
Maximum likelihood estimates (MLEs) for trend models of best fit for each analysis of phenanthrene, p,p'-DDE, α-endosulfan and PCB-28

Chemical	Model	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
Phenanthrene	Survival	3.87 (0.0824)	0.623 (0.0572)	0.195 (0.0598)	0.736 (0.0585)	0.135 (0.0581)			−0.963 (0.142)
	Replacement	3.91 (0.0757)	0.576 (0.0519)	0.0552 (0.0519)	0.695 (0.0517)	0.110 (0.519)			−0.931 (0.129)
p,p'-DDE	Survival	−1.39 (0.0480)	0.578 (0.694)	−0.117 (0.0661)	−0.318 (0.0658)	−0.133 (0.0645)			
	Replacement	−0.852 (0.130)							0.508 (0.0650)
α-endosulfan	Survival	0.975	−0.702 (0.0966)	−0.301 (0.0641)	−0.638 (0.0749)	−0.310 (0.0800)			
	Replacement	1.02 (0.0551)	−0.631 (0.0870)	−0.228 (0.0634)	−0.591 (0.0715)	−0.293 (0.0783)			
PCB-28	Survival	−0.134 (0.0285)	0.00487 (0.0411)	0.114 (0.0399)					
	Replacement	−0.361 (0.0310)	−0.0132 (0.0434)	0.146 (0.0434)					

Chemical	Model	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Phenanthrene	Survival	−0.213 (0.0400)							
	Replacement	−0.273 (0.0339)							
p,p'-DDE	Survival	0.301 (0.108)	0.218 (0.0727)	0.0915 (0.0745)	−0.00164 (0.0713)	−0.271 (0.0670)			−0.964 (0.171)
	Replacement	−0.0498 (0.0641)	−0.664 (0.190)	0.287 (0.0795)					−0.835 (0.135)
α-endosulfan	Survival	−0.00563 (0.179)	0.0352 (0.0413)	0.129 (0.0621)	0.266 (0.0586)	0.173 (0.0603)	−0.130 (0.0629)	−0.261 (0.0603)	
	Replacement	−0.0399 (0.0389)	0.271 (0.0569)	0.109 (0.0589)	0.206 (0.0563)	0.140 (0.0556)	−0.137 (0.0591)	−0.272 (0.0570)	
PCB-28	Survival	−0.131 (0.0818)							−0.918 (0.139)
	Replacement	−0.886 (0.0388)	−0.0731 (0.0538)	0.185 (0.0566)					

Figures in brackets are the estimated standard errors of the parameter estimates. Blank cells correspond to terms not included in the model. α and β parameters are defined in Eqs. (8) and (9).

quarterly cycles and that of p,p'-DDE had up to biannual cycles. These findings compare favourably with the more sophisticated time-series models of Digital Filtration and DHR applied to PCBs/OC pesticides and PAHs, respectively, although for phenanthrene the declining trend determined by DHR was found not to be statistically significant. It should be noted that the normalised time covariate given in Eq. (10) was simply used to determine the year-on-year trend, whereby normalisation was required to ease the computational aspect of model fitting. This produced the same results as using a non-normalised time covariate, except that some care was required with the interpretation of the results. It is worth noting that time-series plots of the standardised residuals from all the best fitting models (Section 3) suggested that the serial correlation had been largely removed by the modeling procedure.

The four sets of plots in Fig. 2 illustrate the fitted mean trend, superimposed on the raw data. The K–M estimate of the survivor curve of the residuals,

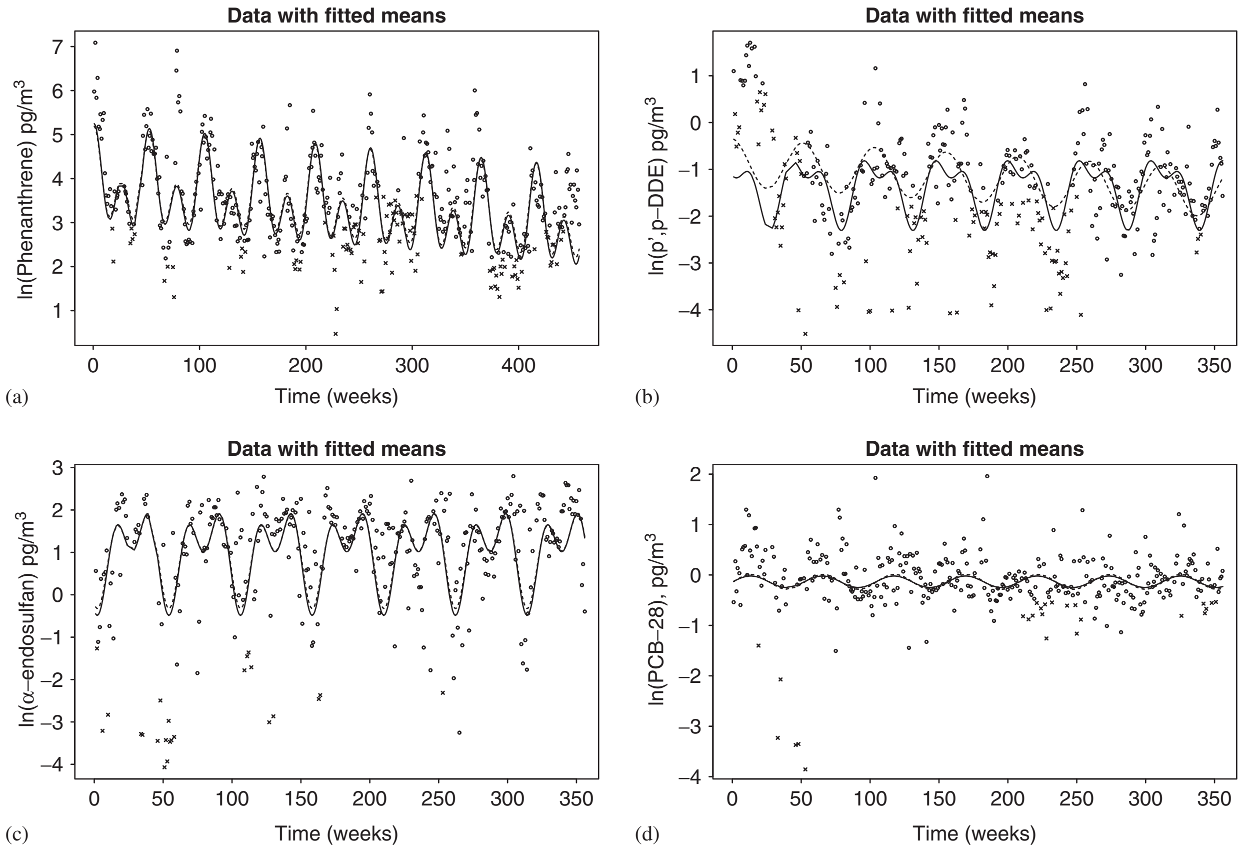


Fig. 2. The fitted means from the best fitting models under both the survival and the replacement analysis for (a) phenanthrene, (b) p,p'-DDE, (c) α-endosulfan and (d) PCB-28 data sets. The full (dashed) line indicates the mean of the survival (replacement) analysis. Circles indicate exact data and crosses represent data below the MDL.

with the survivor function of the standard normal distribution, is shown in Fig. 3. The K–M plot includes the 95% confidence intervals of the estimate. The survivor curve for the standard normal distribution almost always falls inside the confidence intervals for the survivor curves fitted to the model residuals. This suggests that the residuals are approximately standard normal, once censoring is considered, implying that the models fit the data well, supporting the inclusion of censored data for time-series analysis.

4.2. Replacement analysis

In this approach, the data sets were analysed with the censored data removed. This meant all values below the MDL were replaced with $\frac{2}{3}$ MDL and subsequently treated as exact. The same model selection procedure and diagnostics were applied as before. As the data were all treated as exact, likelihood L from Eq. (3) was used to fit the models and the results are shown in Table 2. The covariates found in the best-fitting models under this analysis were compared with their counterparts in the survival analysis. Under the replacement analysis the best-fitting trend models for phenanthrene and α-endosulfan contained the same covariates as their counterparts under the survival analysis. In contrast, under the replacement analysis, the best-fitting model for p,p'-DDE no longer had annual or biannual cycles in the mean parameter, as was found under the survival method, nor did it have a biannual cycle in the standard deviation parameter. The best-fitting model for PCB-28 under the replacement analysis also differed, since the standard deviation parameter now contained an annual cycle, but no year-on-year trend. Under this analysis both phenanthrene and p,p'-DDE had overall mean year-on-year time trends, decreasing for phenanthrene and actually increasing for p,p'-DDE, which differs from Hung et al. (2002), where all the censored data for p,p'-DDE were retained in the trend analysis, and no trend was evident. The standard deviations for p,p'-DDE also displayed different seasonal and year-on-year trends under this approach. The best-fitting trends for the replacement analysis are illustrated in Fig. 2 and show that the trend models actually fit the data in the replacement analysis reasonably well. This is not surprising as the degree of censoring for these compounds is not that high (i.e. it does not exceed 50% for any of the compounds). It is interesting to note that the most obvious difference in the trend models under the two analyses occurs for the data set with the greatest quantity of censoring (i.e. p,p'-DDE). Furthermore, in this case, the diagnostic plots (Figs. 2 and 3) show that a better fit to the data is given by the survival analysis models. This suggests that a survival analysis has its greatest advantage for data sets which are heavily censored, which is particularly relevant for some of the high molecular weight compounds monitored in Arctic air, where censored data may exceed 50% of the total data set.

4.3. Increasing the level of censoring

To further examine the effect that increasing the degree of censoring has on data interpretation and the robustness of the survival analysis, the data sets for phenanthrene and p,p'-DDE were subject to 50% censoring. This was achieved by deriving a hypothetical or artificial MDL, set as the median-MDL (i.e. the median of the whole time-series) for each compound. This produced 50% censoring, but with the benefit that the true values of some of the censored data points were known (i.e. those observations between the true MDL and the median MDL). These 'median-censored' data sets were analysed using both the survival and replacement

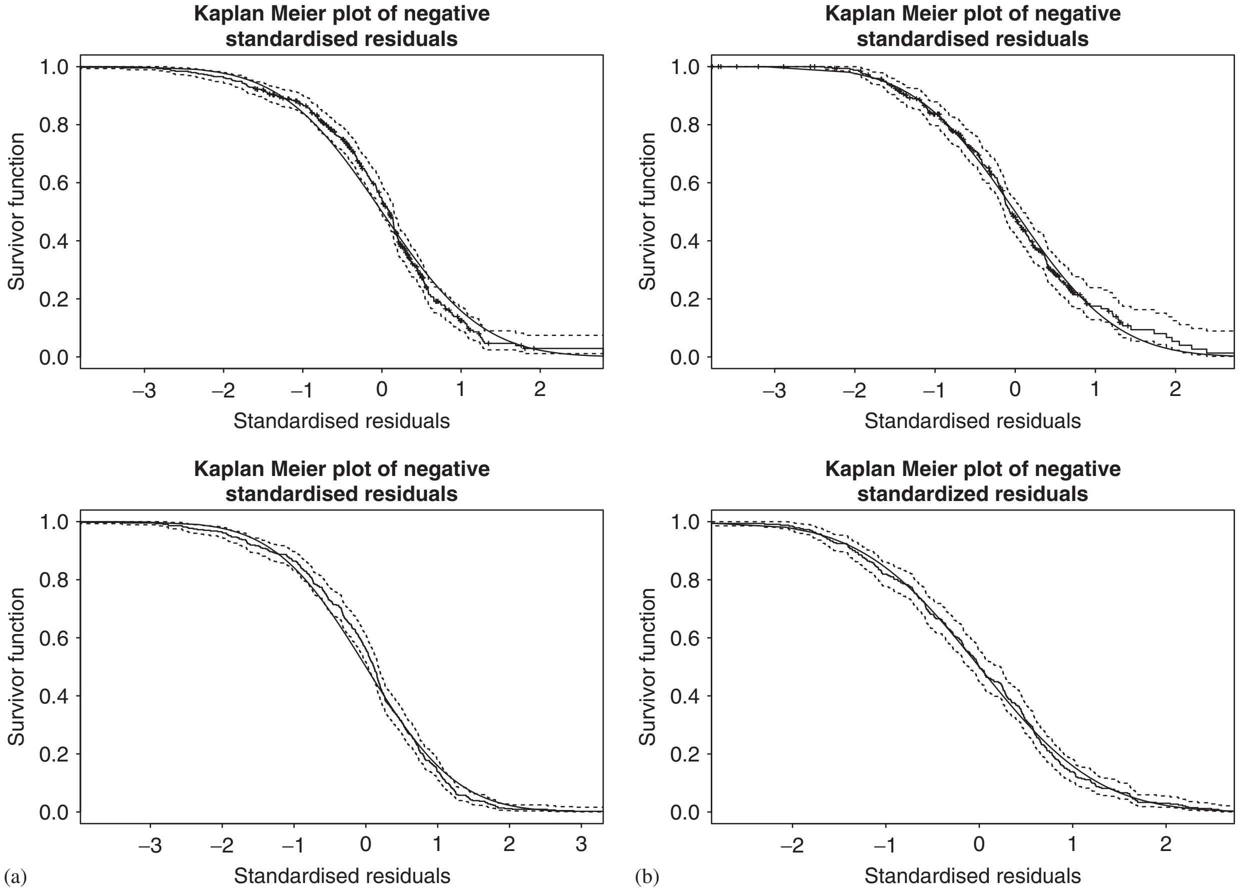


Fig. 3. Kaplan–Meier estimates for the survivor curve of the negative, standardised residuals from the best fitting models under both the survival and the replacement analysis for (a) phenanthrene, (b) p,p'-DDE, (c) α-endosulfan and (d) PCB-28 data sets. Vertical lines represent the residuals from censored observations and the dashed lines show the 95% confidence interval. The bold line indicates the standard normal survivor function.

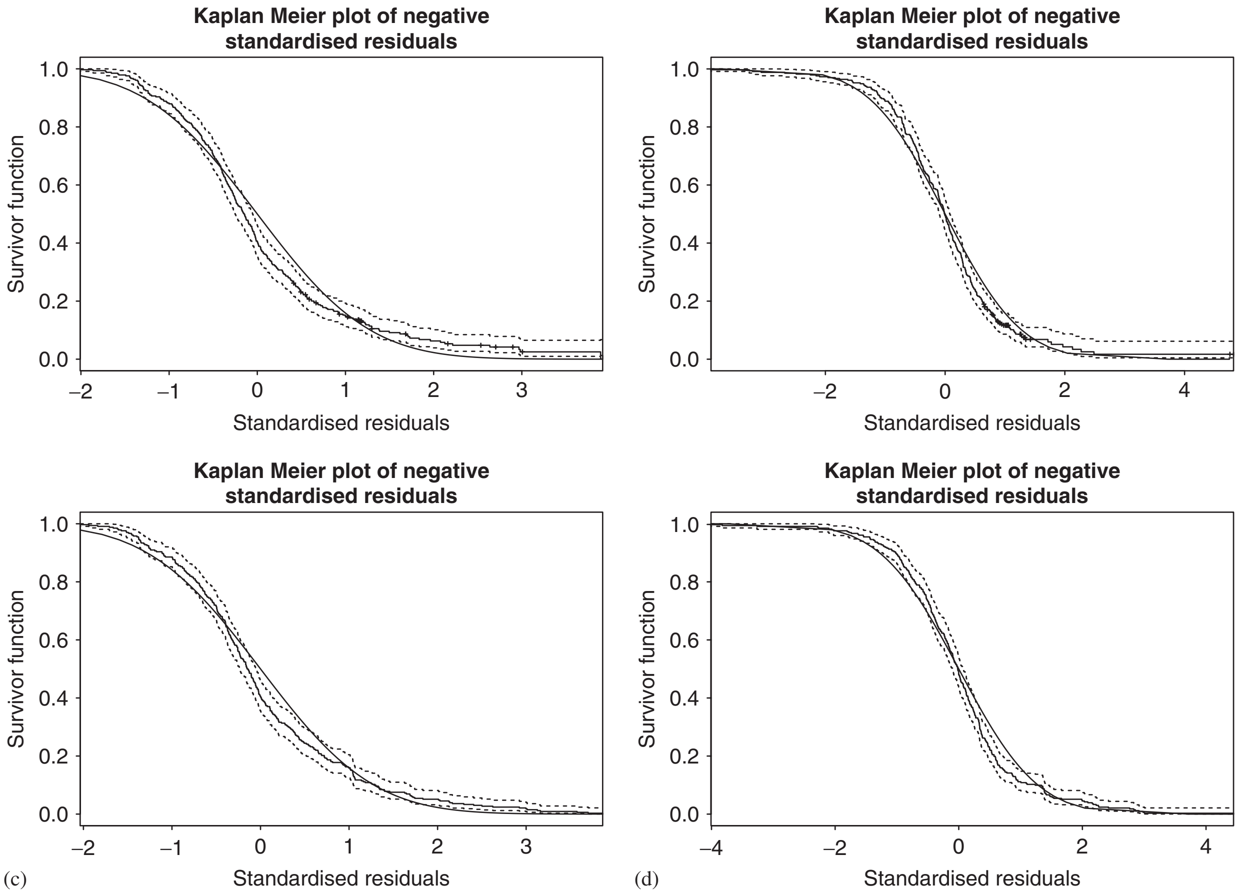


Fig. 3. (*Continued*)

analyses, in which the replacement value for the new censored data sets was $\frac{2}{3}$ median-MDL. The results for these fitted models can be seen in Table 3. Fig. 4 shows the fitted mean and Fig. 5 the KM plots for the best-fitting trend models under the two analysis methods (survival and replacement) for the p,p'-DDE data set and a clear difference can be observed in the best-fitting trend model for the two analyses. A similar situation was also observed for phenanthrene (figure not shown), where a decreasing year-on-year mean trend was found for phenanthrene under both analyses, but the survival analysis resulted in a yearly decrease of 0.158 in logged concentrations, compared to a yearly decrease of 0.076 under the replacement analysis.

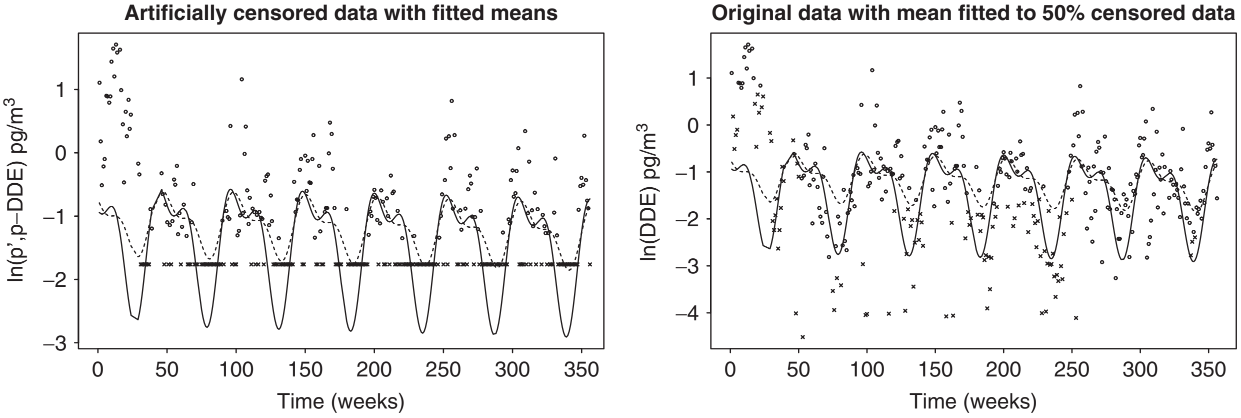
For these data sets, it is possible to judge how well the models obtained under 50% censoring describe the true observed (and less censored) data set. Plots of the p,p'-DDE data set censored at the true MDL, and superimposed with the 50% censored model mean are shown in Fig. 4, for both the survival and replacement methods. If the fitting method is robust to censoring then the model obtained by fitting to a data set with a high proportion of censored (replaced) values would still fit well to the complete data set when the values that were censored or replaced for the fitting are reintroduced. The model trend for the replacement analysis does not fit closely to any of the low values of the complete data set which are actually uncensored by the true MDL, whereas the survival analysis model shows a good fit to the whole range of the complete data set. Fig. 5 also shows KM plots of the negative standardised residuals for the complete data sets of the models under the two methods, fitted with the 50% censored data. It is clear from these plots that the trend model for the survival analysis, even with 50% of the data censored, fits the complete data set very well, whereas the replacement model has a much poorer fit, with the estimated KM plot lying almost always outside of the 95% confidence intervals.

Table 3
Maximum likelihood estimates (MLEs) for the trend model of best fit for each analysis of phenanthrene and p,p'-DDE, using the artificial censoring level (median MDL) to censor 50% of the data

Chemical	Model	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
Phenanthrene	Survival	3.91 (0.0977)	0.745 (0.0877)	0.343 (0.106)	0.741 (0.0753)	0.0417 (0.0737)			−1.32 (0.165)
	Replacement	3.91 (0.0685)	0.489 (0.0429)	0.0521 (0.0440)	0.522 (0.0439)	0.0867 (0.0428)			−0.636 (0.109)
p,p'-DDE	Survival	−1.36 (0.201)	0.896 (0.163)	0.0310 (0.0862)	−0.444 (0.115)	−0.157 (0.0932)			−0.210 (0.289)
	Replacement	−1.05 (0.0740)	−0.116 (0.121)	0.163 (0.107)	0.0798 (0.0968)	−0.249 (0.0908)			−0.860 (0.231)

Chemical	Model	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Phenanthrene	Survival	−0.115 (0.0564)	−0.281 (0.0696)	−0.338 (0.0864)					
	Replacement	−0.297 (0.0720)							−0.335 (0.124)
p,p'-DDE	Survival	0.295 (0.121)	−0.116 (0.121)	0.163 (0.107)	0.0798 (0.0968)	−0.249 (0.0908)			−0.860 (0.231)
	Replacement	0.0973 (0.0845)	0.489 (0.0613)	0.113 (0.0626)	−0.273 (0.0595)	−0.380 (0.0597)			−1.24 (0.144)

Figures in brackets are the estimated standard errors of the parameter estimates. Blank cells correspond to terms not included in the model. α and β parameters are defined in Eqs. (8) and (9).



(a) Artifcially censored data (b) Complete data

Fig. 4. The fitted means from the best fitting models under both the survival (full line) and the replacement analysis (dashed line) for the artificially censored p,p'-DDE data set, using the median-MDL to achieve 50% censoring. The first plot shows the fitted mean parameter superimposed on the artificially censored data. In this plot the crosses are the censored data replaced by $\frac{2}{3}$ median. The second plot shows the fitted mean parameter superimposed on the complete data set. Circles indicate exact data and crosses represent data below the MDL.

5. Conclusions

The evidence provided by these examples supports the use of survival analysis methods, whereby all the censored data are utilised within a given data set. Such techniques provide a highly reasonable alternative method for analysing air pollution data which contain censored values. In other words, procedures like blank correction or censored-data removal may ultimately harm trend analysis and

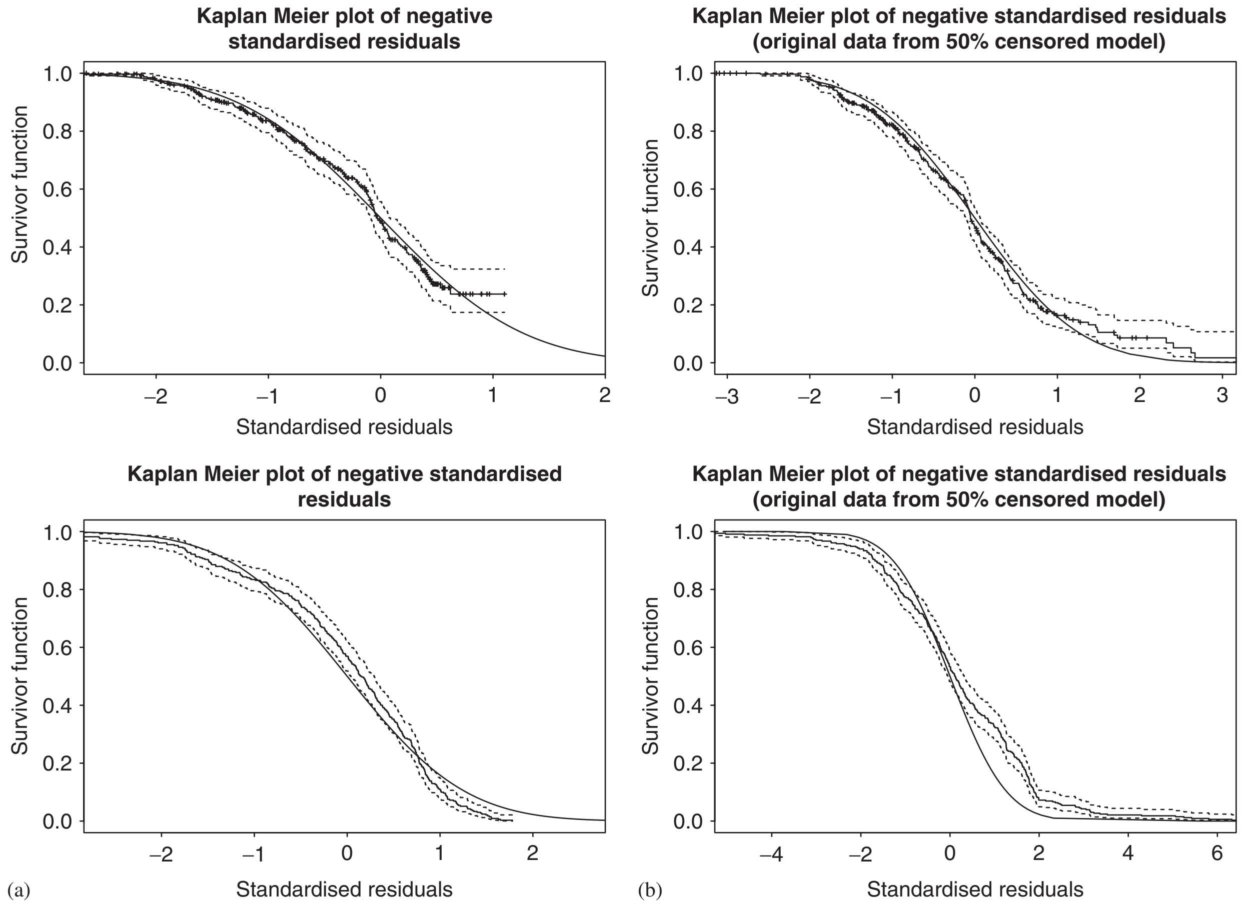


Fig. 5. Kaplan–Meier estimates for the survivor curve of the negative, standardised residuals from the best-fitting models under both the survival and the replacement analyses for the artificially censored p,p'-DDE data set, using the median MDL for 50% censoring. Estimates are for residuals calculated from (a) the artificially censored data, using survival analysis (top) and replacement analysis (bottom) and (b) the complete data, using survival analysis (top) and replacement analysis (bottom). Vertical lines represent the residuals from censored observations and the dashed lines show the 95% confidence interval. The bold line indicates the standard normal survivor function.

data interpretation. With low levels of censoring, evident in the data sets for the compounds chosen here, the survival analysis produced results that were comparable to the replacement method, where the censored values were simply replaced with a fixed value. However, when the level of censoring was increased by imposing an artificial MDL (median MDL), the survival analysis models proved to be far more robust. We recommend that an analysis of any data set of this nature should retain concentrations that may otherwise fall below some arbitrary analytical reference point to allow the fullest exploitation of the data for trend analysis. Furthermore, this will allow better understanding of chemical behaviour, particularly where there is evidence of substantial seasonal or intra-seasonal behaviour and vindicates the approach taken by the Canadian NCP in retaining censored data within their Arctic data sets. Survival analysis methods allow trend models to be built which include the full amount of information present in the data, including those observations censored by the MDL.

Acknowledgements

Emma Eastoe was funded on a studentship by the UK EPSRC. The authors would like to express their gratitude to the Canadian Arctic Environmental Strategy Northern Contaminants Program, Department of Indian Affairs and Northern Development (DIAND) for continued funding of the air monitoring stations. We would also like to thank all site operators, as well as the teams at AirZone One Inc., Canada, and the Freshwater Institute, Canada, for preparing, collecting and analyzing the arctic air samples.

References

AMAP, 2004. AMAP Assessment 2002: Persistent Organic Pollutants in the Arctic. Arctic Monitoring and Assessment Programme (AMAP), Oslo, Norway, xvi + 310pp.

Becker, S., Halsall, C.J., Tych, W., Hung, H.H., Attwell, S., Blanchard, P., LI, H., Fellin, P., Stern, G., Billeck, B., 2006. Resolving the long-term trend of PAHs in the Canadian Arctic atmosphere. Environmental Science and Technology 40, 3217–3222.

Coe, R., Stern, R.D., 1982. Fitting models to daily rainfall data. Journal of Applied Meteorology 21, 1024–1031.

Dobson, A.J., 1990. An Introduction to Generalized Linear Models. Chapman & Hall/CRC Press, New York.

Fellin, P., Barrie, L.A., Dougherty, D., Toom, D., Muir, D., Grift, N., Lockhart, L., Billeck, B., 1996. Air monitoring in the arctic: results for selected persistent organic pollutants for 1992. Environmental Toxicology and Chemistry 15, 253–261.

Halsall, C.J., Bailey, R., Stern, G.A., Barrie, L.A., Muir, D.C.G., Fellin, P., Rosenberg, B., Rovinski, F.Ya., Kononov, E.Ya., Pastukhov, B.V., 1998. Multi-year observations of organohalogen pesticides in the Arctic atmosphere. Environmental Pollution 102, 51–62.

Helsell, D.R., 2005. More than obvious: better methods for interpreting nondetect data. Environmental Science and Technology 39, 419A–423A.

Hougaard, P., 2000. Analysis of Multivariate Survival Data. Springer, New York.

Hung, H.H., Blanchard, P., Halsall, C.J., Bidleman, T.F., Stern, G.A., Fellin, P., Muir, D.C.G., Barrie, L.A., Jantunen, L.M., Helm, P.A., Ma, J., Konoplev, A., 2005. Temporal and spatial variability of atmospheric POPs in the Candian Arctic: results from a decade of monitoring. The Science of the Total Environment 342, 119–144.

Hung, H.H., Halsall, C.J., Blanchard, P., 2001. Are PCBs in the Canadian Arctic atmosphere declining? Evidence from 5 years of monitoring. Environmental Science and Technology 35, 1303–1311.

Hung, H.H., Halsall, C.J., Blanchard, P., Li, H.H., Fellin, P., Stern, G., Rosenberg, B., 2002. Temporal trends of organochlorine pesticides in the Canadian Arctic atmosphere. Environmental Science and Technology 36, 862–868.

Knight, K., 2000. Mathematical Statistics. Chapman & Hall/CRC Press, New York.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, second ed. Wiley, New York.

Leister, D.L., Baker, J.E., 1994. Atmospheric deposition of organic contaminants in the Chesapeake Bay. Atmospheric Environment 28, 1499–1520.

Li, N.Q., Wania, F., Lei, Y.D., 2003. A comprehensive and critical compilation, evaluation, and selection of physical–-chemical property data for selected polychlorinated biphenyls. Journal of Physical and Chemical Reference Data 4, 1545–1590.

Lindsey, J., 1995. Introductory Statistics: A Modelling Approach. Chapman & Hall/CRC Press, New York.

Macdonald, R.W., Barrie, L.A., Bidleman, T.F., Diamond, M.L., Gregor, D.J., Semkin, R.G., Strachan, W.M.J., Li, Y.F., Wania, F., Alaee, M., Alexeeva, L.B., Backus, S.M., Bailey, R., Bewers, J.M., Gobeil, C., Halsall, C.J., Harner, T., Hoff, J.T., Jantunen, L.M.M., Lockhart, W.L., Mackay, D., Muir, D.C.G., Pudykiewicz, J., Reimer, K.J., Smith, J.N., Stern, G.A., Schroeder, W.H., Wagemann, R., Yunker, M.B., 2000. Contaminants in the Canadian Arctic: 5 years of progress in understanding sources, occurrence and pathways. The Science of the Total Environment 254, 93–234.

Mackay, D., Shui, W.-Y., Ma, K., 1992. Illustrated Handbook of Physical–Chemical Properties and Environmental Fate of Organic Chemicals. Vol II: Polynuclear Aromatic Hydrocarbons, Polychlorianted Dioxin, and Dibenzofurnas. Lewis Publishers, Chelsea, MI, USA.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman & Hall/CRC Press, New York.

Shen, L., Wania, F., 2005. Compilation, evaluation, and selection of physical–chemical property data for organochlorine pesticides. Journal of Chemical Engineering and Data 3, 742–768.