THE UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS
TECHNICAL REPORT # 217

SEASONAL CONFOUNDING AND RESIDUAL CORRELATION IN ANALYSES OF HEALTH EFFECTS OF AIR POLLUTION

BY

ISABELLA R. GHEMENT
NANCY E. HECKMAN
A. JOHN PETKAU

MARCH 2006

Seasonal Confounding and Residual Correlation in Analyses of Health Effects of Air Pollution

Isabella R. Ghement, Nancy E. Heckman and
A. John Petkau

Department of Statistics
University of British Columbia
Vancouver, British Columbia, V6T 1Z2, CANADA

March, 2006

Grant sponsor: Natural Sciences and Engineering Research Council of Canada (NSERC); grant numbers: NSERC RGPIN 4586-00 and 7969-99.

ABSTRACT

To investigate the health effects of air pollution via a partially linear model, one must choose the correct amount of smoothing for accurate estimation of the linear pollution effects. This choice is complicated by the dependencies between the various covariates and by the potential residual correlation. Most existing approaches to making this choice are inadequate, as they neither target accurate estimation of the linear pollutant effects nor handle residual correlation. In this paper, we illustrate two new adaptive and objective methods for determining the appropriate amount of smoothing. We construct valid confidence intervals for the linear pollutant effects; these intervals account for residual correlation with minimum input from the user. We use our inferential methods to investigate the same-day effects of PM10 on daily mortality in two data sets for the period 1994 to 1996: one collected in Mexico City, an urban area with high levels of air pollution, and the other collected in Vancouver, British Columbia, an urban area with low levels of air pollution. For Mexico City, our methodology does not detect a PM10 effect. In contrast, for Vancouver, a PM10 effect corresponding to an expected 2.4% increase (95% confidence interval ranging from 0.0% to 4.7%) in daily mortality for every 10 $\mu g/m^3$ increase in PM10 is identified.

KEYWORDS AND PHRASES: air pollution; bandwidth selection; generalized additive model; mortality; partially linear model; particulate matter; residual correlation; seasonal confounding; smoothing

1 Introduction

Many community-level time series studies have provided evidence that air pollution has adverse health effects on humans (see, for instance, Pope et al., 1995, Goldberg et al., 2003, and Bell et al., 2004). Disentangling the association between air pollution and the health outcome of interest is a delicate problem. This association may be confounded by highly correlated factors such as season, temperature and relative humidity, and may also be masked by the correlated nature of the time series response.

Various statistical models can be used to describe the relationship between air pollution and the health outcome of interest (e.g. daily non-accidental mortality rate), adjusted for relevant seasonal and weather confounders. The most widely used in recent years have been generalized additive models (GAM); see, for example, Schwartz (1994), Dominici et al. (2002b), Ramsay et al. (2003a, 2003b) and Touloumi et al. (2004). In a GAM model, the effects of the pollutants of interest on a suitable transformation of the mean response are typically presumed to be linear for the sake of interpretation, whereas those of the remaining covariates (e.g. day of study, temperature, relative humidity) are often presumed simply to be of unspecified smooth form. The degree of smoothing for estimating the effects of these covariates must be specified when fitting a GAM model.

Inference about the linear air pollution effects in such a GAM model is the primary objective. Therefore, the choice of the degree of smoothness of the estimated smooth effects should be informed by this objective. In addition, this choice should account for the dependencies between the various covariates in the model and for potential residual correlation. Most smoothing choice approaches in the air pollution literature focus on different objectives, such as doing well at estimating the smooth effects or minimizing the amount of residual correlation. Recently, Dominici et al. (2004) and Peng et al. (2006) developed methods targeted at choosing the right degree of smoothness for accurate estimation of the linear air pollution effects. However, to our knowledge, no attempts have been made to adapt these methods to situations where residual correlation is present.

In the context of investigating the health impacts of PM10, Peng et al. (2006) study some of the most common methods for choosing the amount of smoothing, including a simplified version of the mean squared error minimization procedure of Dominici et al. (2004), which they call the GCV-PM10 method. This method, designed to eliminate seasonal confounding, chooses the amount

of smoothing by minimizing a generalized cross-validation criterion intended to assess how well the pollutant series of interest can be predicted by a smooth function of day of study. Using simulated data, Peng et al. (2006) show that the GCV-PM10 method outperforms other methods they consider, in the sense of producing estimated pollutant effects with lower mean squared errors.

As we shall see, the GCV-PM10 method fails to provide a suitable choice of amount of smoothing for the two data sets considered in this paper. We will describe and illustrate new methodology for choosing the appropriate amount of smoothing to achieve accurate estimation of the linear effects in a GAM model containing a single nonparametric smooth term when residual correlation is present. Ghement and Heckman (2006) investigated some of the asymptotic properties of this methodology and conducted simulation studies of the finite sample behavior. The need to account for residual correlation is emphasized by Schwartz (2006).

To illustrate the importance of the choice of degree of smoothness, we refer to a data set collected in Mexico City over a period of three years, from January 1, 1994 to December 31, 1996, that is discussed in more detail in Section 5. The data consist of daily counts of non-accidental deaths and daily levels of ambient concentration of PM10 ($\mu g/m^3$), temperature (^oC) and relative humidity (%). As seen in Figure 1, the daily counts of non-accidental deaths and the daily levels of PM10 show pronounced annual patterns, which peak at roughly the same time. The daily levels of temperature and relative humidity also exhibit strong annual patterns. Note that the concentrations of PM10 in Mexico City in this study were quite high: the 50th and 90th percentiles of daily PM10 concentrations were 63 and 96 $\mu g/m^3$, respectively. Also, the number of daily non-accidental deaths ranged largely from 90 to 130 per day.

Preliminary model selection, detailed in Appendix B, suggests that an adequate GAM model for the Mexico City data is:

$$log(D_t) = \beta_0 + \beta_1 PM_t + m(t) + \beta_2 T_t + \beta_3 H_t + \epsilon_t,\ t = 1, \ldots, 1096, \tag{1}$$

where D_t is the observed number of non-accidental deaths in Mexico City on day t, and PM_t, T_t and H_t denote the daily measures of PM10, temperature and relative humidity, respectively. Model (1) is referred to as a partially linear model as it contains only one smooth nonparametric effect, m. A detailed exposition of partially linear models is provided by Härdle et al. (2000).

Figure 2 illustrates the impact of the amount of smoothing used for estimating the seasonal effect m in model (1) on 95% confidence intervals for β_1. These intervals are similar to those obtained by software typically used for analyzing this type of data. In particular, the standard errors used to evaluate these 95% confidence intervals do not account for possible error correlation. Each of these confidence intervals is centered about a local linear backfitting estimate for β_1, computed as described in Section 2.

Figure 2 shows, in a dramatic fashion, that changing the amount of smoothing for estimating m greatly affects the inferences on β_1, the short-term PM10 effect on log mortality. The amount of smoothing is determined by a bandwidth, but can be translated into an equivalent number of degrees of freedom. Large bandwidths, which correspond to few degrees of freedom, lead to the conclusion that the data provide strong evidence in favour of a PM10 effect on log mortality in Mexico City, after adjusting for seasonal and weather confounding. In contrast, with smaller bandwidths, the data do not provide evidence in support of a PM10 effect. The sensitivity of conclusions to the amount of smoothing – even when ignoring potential residual correlation – reinforces the need to make this choice in an objective fashion to ensure valid inferences on the PM10 effect.

Our implementation of the GCV-PM10 method of Peng et al. (2006) yields too small a choice of bandwidth for the Mexico City data. Indeed, Figure 3 shows that the GCV-PM10 method for a local linear kernel regression smoother specifies the smallest amount of smoothing considered, resulting in a predicted PM10 series that tends to interpolate the PM10 measurements. Using the GCV-PM10 method with other smoothing methods instead of local linear kernel regression yields similar overfitting.

The remainder of the paper is organized as follows. Section 2 describes our proposed model, a partially linear model with correlated errors, together with a method for estimating the linear and smooth effects in this model. Section 3 presents adaptive methods for choosing the degree of smoothness of the estimated smooth effect when accurate estimation of the linear effects is the objective. Section 4 provides a method for constructing approximate confidence intervals for the linear effects of interest. Sections 5 and 6 apply this methodology to the analysis of two air pollution/mortality data sets, including the data set collected in Mexico City. The paper concludes with a brief discussion.

2 Partially Linear Models with Correlated Errors

Given the data $(Y_t, X_{tj}, t),\ t = 1, \ldots, n, j = 1, \ldots, p$, where the Y_t are measurements on a continuous response (e.g. log mortality) and the $X_{t1}, \ldots, X_{tp}$ are measurements on p covariates (e.g. PM10, temperature, relative humidity), a partially linear model with correlated errors is given by:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_p X_{tp} + m(t) + \epsilon_t,\ t = 1, \ldots, n. \tag{2}$$

Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is an unknown parameter vector, m is an unknown, smooth, real-valued function and the error terms satisfy $E(\epsilon_t) = 0$, $Var(\epsilon_t) = \sigma_\epsilon^2 > 0$ and $Corr(\epsilon_t, \epsilon_{t'}) = \boldsymbol{\Psi}_{t,t'}$. To ensure identifiability in model (2), we assume that m satisfies the restriction $\sum_{t=1}^n m(t) = 0$. The methodology to be described applies to estimating linear combinations of the linear effects, but our focus in this paper is on estimating β_1. Thus, we treat the smooth effect m and the potential error correlation as nuisances.

To estimate β_1, we employ the backfitting algorithm of Hastie and Tibshirani (1990). The algorithm uses an iterative estimation scheme to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{m} = (m(1), \ldots, m(n))^\top$. The choice of smoothing method used for estimating $\boldsymbol{m}$ determines the smoother matrix $\boldsymbol{S}_h^c$ appearing in the algorithm; this matrix is indexed by the smoothing parameter h which controls the degree of smoothness of the estimated $\boldsymbol{m}$. We use local linear kernel regression; see Appendix A for the definition of $\boldsymbol{S}_h^c$ in this case. In the remainder of this paper, h will be referred to as a bandwidth.

We refer to the resulting estimators of $\boldsymbol{\beta}$ and $\boldsymbol{m}$, denoted by $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{\boldsymbol{m}}_h$, as local linear backfitting estimators. These estimators are the fixed points of the backfitting equations:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_h &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \widehat{\boldsymbol{m}}_h), \\
\widehat{\boldsymbol{m}}_h &= \boldsymbol{S}_h^c (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_h),
\end{aligned}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix corresponding to the parametric part of model (2).

Usually, $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{\boldsymbol{m}}_h$ are found by iterating the updating steps $\widehat{\boldsymbol{\beta}}_h^{(k)} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{Y} - \widehat{\boldsymbol{m}}_h^{(k-1)})$ and $\widehat{\boldsymbol{m}}_h^{(k)} = \boldsymbol{S}_h^c (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_h^{(k)})$ over k. However, iteration is not required to compute $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{\boldsymbol{m}}_h$.

Straightforward algebraic manipulation of the backfitting equations yields the following explicit expressions for these estimators:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_h &= (\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S}_h^c)\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S}_h^c)\boldsymbol{Y} \equiv \boldsymbol{M}_h\boldsymbol{Y}, \\ \widehat{\boldsymbol{m}}_h &= \boldsymbol{S}_h^c(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{M}_h)\boldsymbol{Y}, \end{aligned} \tag{3}$$

provided the $(p+1)\times(p+1)$ matrix $\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{S}_h^c)\boldsymbol{X}$ is invertible. If this condition holds, the solution is unique.

3 Choosing the Amount of Smoothing

Selecting the appropriate bandwidth h for computing the estimator of β_1, denoted $\widehat{\beta}_{1,h}$, is crucial. Ghement and Heckman (2006) show that, in the presence of error correlation and dependencies between the linear and smooth components of the model, the 'optimal' h for estimating β_1 is smaller than that for estimating m. Opsomer and Ruppert (1999) obtained a similar result for the case when the errors are uncorrelated. Peng et al. (2006) also illustrate the need for under-smoothing via simulations and point out that this should not be interpreted as a problem, but as a "fact of semiparametric life of which the practitioner should be keenly aware". Adaptive methods devised for accurate estimation of m, such as cross-validation, are unlikely to yield a satisfactory choice of amount of smoothing for accurate estimation of β_1.

To choose h values appropriate for the estimation of β_1, we will use two of the adaptive bandwidth selection methods developed and recommended by Ghement and Heckman (2006) and discussed in Section 3.2 below. Both methods select h to minimize an estimator of the conditional mean squared error of $\widehat{\beta}_{1,h}$, given $\boldsymbol{X}$. Both methods require preliminary estimators of $\boldsymbol{m}$, σ_ϵ^2 and $\boldsymbol{\Psi}$, described in Section 3.1.

3.1 Preliminary Estimation of $\boldsymbol{m}$, σ_ϵ^2 and $\boldsymbol{\Psi}$

Our preliminary estimators of $\boldsymbol{m}$, σ_ϵ^2 and $\boldsymbol{\Psi}$ depend on a bandwidth b chosen to yield accurate estimation of $\boldsymbol{m}$. This bandwidth b is not to be confused with the bandwidth h of Section 3.2,

which is chosen to achieve accurate estimation of β_1.

We estimate $\boldsymbol{m}$ via $\widehat{\boldsymbol{m}}_b$, the local linear backfitting estimator with bandwidth b chosen via cross-validation, modified to account for possible error correlation. Specifically, if l is a positive integer quantifying our belief in the extent of the serial correlation, we take the bandwidth b to be the minimizer of the leave-$(2l+1)$-out cross-validation criterion:

$$MCV_l(b) = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \boldsymbol{X}_t^\top \widehat{\boldsymbol{\beta}}_b - \widehat{m}_b^{(-t;\,l)}(t)\right)^2.$$

Here, $\widehat{\boldsymbol{\beta}}_b$ is the local linear backfitting estimator of $\boldsymbol{\beta}$ obtained from all the data and $\widehat{m}_b^{(-t;\,l)}(t)$ is the local linear estimator of $m(t)$ computed from the 'data' $(t', Y_{t'} - \boldsymbol{X}_{t'}^\top \widehat{\boldsymbol{\beta}}_b)$ for which $|t - t'| > l$. An explicit expression for $\widehat{m}_b^{(-t;\,l)}(t)$ can be found in Ghement and Heckman (2006).

We use $\widehat{\boldsymbol{m}}_b$ and $\widehat{\boldsymbol{\beta}}_b \equiv (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{Y} - \widehat{\boldsymbol{m}}_b)$ to estimate σ_ϵ^2 and $\boldsymbol{\Psi}$, as follows. We assume that the errors in model (2) are realizations from a covariance-stationary autoregressive process of order R, having mean 0 and variance σ_ϵ^2. We first evaluate the residuals $\widehat{\boldsymbol{\epsilon}}_b = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{m}}_b$. Next, we use $\widehat{\boldsymbol{\epsilon}}_b$ to estimate R via the finite sample criterion developed by Broersen (2000). We then estimate σ_ϵ^2 and the autoregressive parameters in $\boldsymbol{\Psi}$ via Burg's algorithm, described, for instance, in Brockwell and Davis (1991).

3.2 The Plug-In and Global Empirical Methods

The two methods for choosing h discussed here involve estimating $MSE(\widehat{\beta}_{1,h}|\boldsymbol{X}) = Bias^2(\widehat{\beta}_{1,h}|\boldsymbol{X}) + Var(\widehat{\beta}_{1,h}|\boldsymbol{X})$ and then minimizing this as a function of h via a search over a grid of h values, $\mathcal{H}$. Note that, by (3), $\widehat{\beta}_{1,h} = \boldsymbol{c}^\top\widehat{\boldsymbol{\beta}}_h = \boldsymbol{c}^\top\boldsymbol{M}_h\boldsymbol{Y}$, with $\boldsymbol{c} = (0, 1, 0, \ldots, 0)^\top$, so

$$Bias(\widehat{\beta}_{1,h}|\boldsymbol{X}) = \boldsymbol{c}^\top\boldsymbol{M}_h\boldsymbol{m}, \tag{4}$$

and

$$Var(\widehat{\beta}_{1,h}|\boldsymbol{X}) = \sigma_\epsilon^2\boldsymbol{c}^\top\boldsymbol{M}_h\boldsymbol{\Psi}\boldsymbol{M}_h^\top\boldsymbol{c}. \tag{5}$$

Both methods estimate the conditional variance of $\widehat{\beta}_{1,h}$ by replacing σ_ϵ^2 and $\boldsymbol{\Psi}$ in (5) with the estimates from Section 3.1 but they differ in how they estimate the conditional bias of $\widehat{\beta}_{1,h}$.

The *plug-in method* estimates the bias by replacing $\boldsymbol{m}$ in (4) with $\widehat{\boldsymbol{m}}_b$ from Section 3.1.

The *global empirical method* is a modification of the empirical bias bandwidth selection method, developed by Opsomer and Ruppert (1999) in the case of uncorrelated errors. It estimates the bias using an empirical modelling approach, inspired by the fact that, as $h \to 0$ and $n \to \infty$,

$$E(\widehat{\beta}_{1,h}|\boldsymbol{X}) \approx a_0 + \sum_{t=2}^{T} a_t h^t, \tag{6}$$

where $a_0 = \beta_1$ and $a_t, t = 2, \ldots, T$, are unknown constants. In this approach, the bias is estimated by fitting the model (6) to the 'data' $\left\{\left(h, \widehat{\beta}_{1,h}\right) : h \in \mathcal{H}'\right\}$ using ordinary least squares. This results in estimated coefficients $\widehat{a}_0, \widehat{a}_2, \ldots, \widehat{a}_T$ and an estimator for the bias as a function of h:

$$\widehat{Bias}(\widehat{\beta}_{1,h}|\boldsymbol{X}) = \widehat{E(\widehat{\beta}_{1,h}|\boldsymbol{X})} - \widehat{a}_0 = \sum_{t=2}^{T} \widehat{a}_t h^t.$$

In practice, we take $\mathcal{H}'$ to be the same as the $\mathcal{H}$ used in the grid search.

The global empirical method may do a better job of estimating the conditional bias of $\widehat{\beta}_{1,h}$, as it does not rely on a preliminary estimator for $\boldsymbol{m}$. However, this method requires the specification of T in the expansion (6) and a range of h values, $\mathcal{H}'$, to be used in the least squares fitting step. In practice, the choice of $\mathcal{H}'$ has not been a problem. Both methods depend on l, the tuning parameter in the modified cross-validation criterion used for computing the preliminary estimators $\widehat{\boldsymbol{m}}_b$, $\widehat{\sigma}_\epsilon^2$ and $\widehat{\boldsymbol{\Psi}}$. In their simulations, Ghement and Heckman (2006) found both methods to perform well for values of l that are not too close to 0.

4 Inference for β_1

We propose to conduct inference on β_1 based on the approximate 95% confidence interval:

$$\widehat{\beta}_{1,h} \pm z_{\alpha/2}\widehat{SE}(\widehat{\beta}_{1,h}|\boldsymbol{X}), \tag{7}$$

where $\widehat{SE}(\widehat{\beta}_{1,h}|\boldsymbol{X})$ is the square root of the plug-in estimator of the conditional variance in (5) and h is chosen via one of the bandwidth selection methods of Section 3. Here $\widehat{SE}(\widehat{\beta}_{1,h}|\boldsymbol{X})$ is computed with the preliminary estimators $\widehat{\boldsymbol{m}}_b$, $\widehat{\sigma}^2_\epsilon$ and $\widehat{\boldsymbol{\Psi}}$ of Section 3.1 because these estimators are obtained using the bandwidth b optimized for accurate estimation of $\boldsymbol{m}$.

This confidence interval does not account for the uncertainty associated with the data-dependent choice of h and the estimation of the error correlation structure. Nonetheless, Ghement and Heckman (2006) show that this method has good coverage properties in their simulations.

In the next two sections, we apply this inferential methodology to the analysis of two data sets: the Mexico City data already discussed and a similar set of data from Vancouver to be introduced in Section 6. Our goal is to determine if the pollutant PM10 has a significant same-day effect on mortality in these cities after controlling for seasonal and weather confounding.

5 Analysis of Mexico City Data

The Mexico City mortality data were obtained from the Instituto Nacional de Estadística Geografía e Informática. The air quality data were obtained from the Centro de Cómputo de la Dirección de Prevención y Control Contra la Contaminación, México D.F. During the study period, the monitoring network included 8 stations for PM10 and 10 stations for temperature and relative humidity (5 of these were PM10 stations) operating within Mexico City. PM10 ($\mu g/m^3$), temperature (^oC) and relative humidity (%) were recorded on an hourly basis. For each of these variables, missing hourly data were imputed as described in Vedal et al. (2003). These hourly data were converted to daily averages at each station and then averaged across stations to produce the covariates used in our analysis.

The focus of our analysis is to conduct inference about β_1, the linear PM10 effect on log mortality in Mexico City in model (1). We estimate β_1 via $\widehat{\beta}_{1,h}$, the local linear backfitting estimate of β_1 defined in Section 2, with h values chosen via the plug-in and global empirical methods of Section 3. Both methods require preliminary estimation of $\boldsymbol{m}$, σ^2_ϵ and $\boldsymbol{\Psi}$.

We obtain the preliminary estimates for $\boldsymbol{m}$, σ^2_ϵ and $\boldsymbol{\Psi}$ as described in Section 3.1, with the tuning parameter l of the leave-$(2l+1)$-out cross-validation taking on the values $0, 1, \ldots, 10$ to study sensitivity to this choice. For each l, the bandwidth b minimized the leave-$(2l+1)$-out

cross-validation criterion over the range 7 to 365 days. Figure 4 suggests that $l = 0$ and 1 yield under-smoothed estimates of $\boldsymbol{m}$ (the selected bandwidths were 16 days), so we focus on values of l from 2 to 10 (for which the selected bandwidths varied between 34 and 59 days). The order of the autoregressive process fitted to the corresponding model residuals was estimated as $R = 0$ for $l = 2, \ldots, 7$, $R = 2$ for $l = 8$ and 9, and $R = 3$ for $l = 10$.

The preliminary estimates for $\boldsymbol{m}$, σ_ϵ^2 and $\boldsymbol{\Psi}$, depending on $l = 2, \ldots, 10$, are used to estimate $MSE(\widehat{\beta}_{1,h}|\boldsymbol{X})$ by the global empirical and plug-in methods. For the global empirical method, we take $T = 4$ in the expansion (6). To specify the grid $\mathcal{H}' = \mathcal{H}$ of h values for the global empirical method, we assessed the quality of the global polynomial fits produced by this method for $\mathcal{H}$'s of the form $\{7, 8, \ldots, N\}$, where $N = 30(30)210, 250$ or 548. With $T = 4$, all values of N between 90 and 250 provided different choices of bandwidth but essentially the same conclusions regarding the linear PM10 effect β_1; the results presented below correspond to $N = 90$. We then used this same grid for our grid search minimization of the estimate of $MSE(\widehat{\beta}_{1,h}|\boldsymbol{X})$. For these data, the plug-in method is also not very sensitive to the range of the grid for the grid search, as long as this range is reasonably wide. We therefore used the grid $\mathcal{H} = \{7, 8, \ldots, 90\}$ for both methods.

Figure 5 shows the estimated squared bias, variance and mean squared error curves calculated by the plug-in and the global empirical methods for different values of l. The plug-in method yields $h = 24$ days for $l = 2, \ldots, 9$ and $h = 22$ days for $l = 10$. The global empirical method yields $h = 38$ days for all l. Thus, for both methods, the resulting choices of h are remarkably stable as l varies from 2 to 10, although the plug-in choices are smaller than the global empirical choices.

Figure 6 shows the corresponding 95% confidence intervals for β_1 obtained from formula (7). Despite their differing choices of bandwidth, the two methods lead to very similar point estimates and confidence intervals. The standard errors here are comparable with those in Figure 2, reflecting only a modest degree of correlation in the residuals once the systematic components in the data have been taken into account. For this data set, the choice of l has little influence on the results, as the preliminary estimates for σ_ϵ^2 and $\boldsymbol{\Psi}$, and both types of bandwidth choices, are affected very little by the specification of l.

Figure 6 provides no indication of a same-day PM10 effect on log mortality. Further, the magnitude of the PM10 effect is reasonably precisely determined; the range of plausible values indicated by these confidence intervals corresponds to changes in daily mortality of roughly -0.4%

to +0.4% for an increase of 10 $\mu g/m^3$ in the daily concentration of $PM10$.

6 Vancouver Data

This section presents the analysis of an air pollution/mortality data set collected in Vancouver, British Columbia, from January 1, 1994 to December 31, 1996. The data consist of daily counts of non-accidental deaths and daily levels of ambient concentration of PM10 ($\mu g/m^3$), temperature (^oC), relative humidity (%), barometric pressure ($kilopascals$) and rain (% hr/day). The mortality data were obtained from the BC Vital Statistics Agency through the Centre for Health Services and Policy Research at the University of British Columbia and the air quality data were obtained from the Greater Vancouver Regional District. PM10 is a daily average across 10 monitoring sites, temperature is a daily average across 16 sites, relative humidity and barometric pressure are daily averages across 7 sites, and rainfall is a daily average across 8 sites. Additional details about the methods used to collect and pre-process these data can be found in Vedal et al. (2003).

The scatterplots of all variables versus day of study display annual cyclical patterns (Figure 7). The concentrations of PM10 in Vancouver during the study were much lower than in the Mexico City data: the 50th and 90th percentiles of daily PM10 concentrations were 13 and 23 $\mu g/m^3$, respectively. The daily total numbers of deaths were also lower than in Mexico City, ranging largely from 30 to 40 per day.

Preliminary model selection, outlined in Appendix B, identifies the following partially linear model as appropriate for the Vancouver data:

$$log(D_t) = \beta_0 + \beta_1 \cdot PM_t + m(t) + \beta_2 \cdot T_t + \beta_3 \cdot H_t + \beta_4 \cdot P_t + \beta_5 \cdot R_t + \epsilon_t,\ t = 1, \ldots, 1096, \quad (8)$$

where D_t is the number of non-accidental deaths in Vancouver on day t, and PM_t, T_t, H_t, P_t and R_t denote daily measures of PM10, temperature, relative humidity, barometric pressure and rain, respectively.

Figure 8, constructed by the same methods as Figure 2, illustrates the impact of choice of bandwidth for estimating the seasonal effect m in model (8) on estimation of β_1, the linear PM10 effect in this model. The fluctuation of the confidence intervals here is not nearly as dramatic as

in Figure 2. For all values of the bandwidth considered, these naive confidence intervals indicate reasonable evidence of a PM10 effect in Vancouver.

Figure 9 shows that the GCV-PM10 method of Peng et al. (2006) once again results in too small a value of h, which will be inappropriate for estimating β_1 accurately. To choose the correct amount of smoothing for accurate estimation of β_1, we apply the same methodology utilized for the Mexico City data. We implement this methodology with $l = 0, 1, \ldots, 10$, $\mathcal{H} = \{7, 8, \ldots, 90\}$ and $T = 4$ – just as for the Mexico City data.

Examining plots (Figure 10) of the preliminary estimates of $\boldsymbol{m}$, obtained with leave-$(2l+1)$-out cross-validation choices of smoothing for $l = 0, 1, \ldots, 10$, suggests that $l = 10$ (for which the selected bandwidth was roughly 121 days) yields an over-smoothed estimate of $\boldsymbol{m}$, so we concentrate on values of l from 0 to 9 (for which the selected bandwidths varied between 36 and 52 days). For all values of l in this range, the order R of the autoregressive process fitted to the corresponding model residuals is estimated to be zero.

Figure 11 displays the estimated squared bias, variance and mean squared error curves used for determining the plug-in and global empirical choices of bandwidth for $\widehat{\beta}_{1,h}$, the local linear backfitting estimate of β_1. Comparing Figure 11 to Figure 5 indicates that the choice of bandwidth for the Vancouver data is not as straightforward as for the Mexico City data. For the Vancouver data, the global minimizer of the mean squared error curves produced by each method is the largest value of h considered, 90 days. Visual inspection of the corresponding $\widehat{\boldsymbol{m}}_h$ suggests $h = 90$ is larger than desired. Therefore, for the global empirical method, we choose the first local minimizer of the mean squared error curves, as suggested by Ruppert (1997); the resulting choices of bandwidth are 39 days for all l from 0 to 9. No such adjustment is available for the plug-in method, which fails to select an appropriate choice of bandwidth for these data. We take our plug-in choices of bandwidth to be 90 days for all l from 0 to 9.

Figure 12 shows the corresponding 95% confidence intervals for β_1 obtained from formula (7). Despite the considerably different choices of bandwidth for this data set, the two methods lead to essentially identical point estimates and confidence intervals. Although the PM10 effect is much less precisely determined here than was the case for the Mexico City data set (the estimated standard errors here are almost 6 times larger), Figure 12 indicates that, for all values of l considered, there is reasonable evidence of a same-day PM10 effect in Vancouver. We conclude that increases in

the daily concentrations of PM10 are associated with increases in daily mortality in Vancouver. Specifically, for a 10 $\mu g/m^3$ increase in the daily concentration of PM10, we estimate an expected 2.4% increase in the daily mortality rate with 95% confidence interval ranging, roughly, from 0.0% to 4.7%.

7 Discussion

To investigate the health effects of air pollution via a partially linear model, one must choose the correct amount of smoothing for accurate estimation of the linear pollution effects. This choice is complicated by the dependencies between the various covariates and by the potential residual correlation.

Most methods available in the literature fail to choose the appropriate amount of smoothing because they focus on inadequate objectives such as accurate estimation of the smooth effect or minimization of the amount of residual correlation. The GCV-PM10 method of Peng et al. (2006), intended to overcome this deficiency, was identified as 'best' amongst a number of existing methods on the basis of a simulation study. Unfortunately, in the two data sets investigated here – one collected in Mexico City, an urban area with high levels of air pollution, the other collected in Vancouver, an urban area with low levels of air pollution – this method failed to choose an appropriate amount of smoothing.

We used two adaptive and objective methods of Ghement and Heckman (2006) for determining appropriate choices of amount of smoothing. These methods produced essentially identical results within each data set. Unlike other methods available in the literature, these methods specifically target accurate estimation of the linear pollutant effects and appropriately take into account residual correlation.

We also constructed valid confidence intervals for the linear pollutant effects. These intervals have the feature that they account for residual correlation with minimum input from the user. For Mexico City, our results suggested the absence of a same-day PM10 effect. In contrast, for Vancouver, we found evidence of a same-day PM10 effect. The disparity of the results for the two cities is not overly surprising in view of the considerable variation in such effects across geographical locations, possibly due to the heterogeneous geographical nature of PM10, previously reported in

the literature; see, for example, Dominici et al. (2002a, Figure 2), Touloumi et al. (2004, Figure 3) and Welty and Zeger (2005, Figure 3). It is interesting to speculate why the estimated same-day PM10 effect on mortality in Vancouver, an urban area with low levels of air pollution, is rather larger than typical PM10 effects that have been reported previously.

Acknowledgements

The authors thank Professor Sverre Vedal, Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, and Dr. Eduardo Hernández-Garduño, Division of Tuberculosis Control, British Columbia Centre for Disease Control, Vancouver, for kindly providing the Mexico City and Vancouver data.

References

Bell ML, Samet JM, Dominici F. 2004. Time-series studies of particulate matter. *Annual Review of Public Health* **25**: 247–280.

Brockwell PJ, Davis RA. 1991. *Time Series: Theory and Methods, Second Edition.* Springer–Verlag: New York.

Broersen PMT. 2000. Finite sample criteria for autoregressive order selection. *IEEE Transactions on Signal Processing* **48**: 3550–3558.

Dominici F, Daniels M, Zeger SL, Samet JM. 2002a. Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of American Statistical Association* **97**: 100–111.

Dominici F, McDermott A, Hastie T. 2004. Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association* **99**: 938–948.

Dominici F, McDermott A, Zeger SL, Samet, JM. 2002b. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology* **156**: 193–203.

Fan J. 1993. Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics* **21**: 196–216.

Fan J, Gijbels I. 1992. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics* **20**: 2008–2036.

Ghement IR, Heckman NE. 2006. Inference in partially linear models with correlated errors. *Manuscript in preparation.*

Goldberg, MS, Burnett, RT, Stieb, D. 2003. A review of time-series studies used to evaluate the short-term effects of air pollution on human health. *Review of Environmental Health* **18**: 269–303.

Härdle W, Liang H, Gao J. 2000. *Partially Linear Models.* Physica–Verlag: Heidelberg.

Hastie TJ, Tibshirani RJ. 1990. *Generalized Additive Models.* Chapman & Hall: New York.

Opsomer JD, Ruppert D. 1999. A root-n consistent estimator for semi-parametric additive modelling. *Journal of Computational and Graphical Statistics* **8**: 715–732.

Peng RD, Dominici F, Louis TA. 2006. Model choice in multi-site time series studies of air pollution and mortality (with discussion). *Journal of the Royal Statistical Society Series A* **169**: Part 2, 179–203.

Pope CA, Dockery DW, Schwartz J. 1995. Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation Toxicology* **7**: 1–18.

Ramsay T, Burnett R, Krewski D. 2003a. The effect of concurvity in generalized additive models linking mortality and ambient air pollution. *Epidemiology* **14**: 18–23.

Ramsay T, Burnett R, Krewski D. 2003b. Exploring bias in a generalized additive model for spatial air pollution data. *Environmental Health Perspectives* **111**: 1283–1288.

Ruppert D. 1997. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association* **92**: 1049–1062.

Schwartz J. 1994. Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics* **22**: 471–488.

Schwartz J. 2006. Discussion of "Model choice in multi-site time series studies of air pollution and mortality". *Journal of the Royal Statistical Society Series A* **169**: Part 2, 198–200.

Touloumi G, Atkinson R, Le Tertre A, Samoli E, Schwartz J, Schindler C, Vonk J, Rossi G, Saez M, Rabszenko D, Katsouyanni K. 2004. Analysis of health outcome time series data in epidemiological studies. *Environmetrics* **15**: 101-117.

Vedal S, Brauer M, White R, Petkau J. 2003. Air pollution and daily mortality in a city with low levels of pollution. *Environmental Health Perspectives* **111**: 45–51.

Welty L, Zeger S. 2005. Are the acute effects of PM10 on mortality in NMMAPS the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology* **162**: 80–88.

Appendix A: Definition of Local Linear Smoother Matrix $\boldsymbol{S}_h^c$

Local linear kernel regression is an effective method for estimating a smooth regression function $\mu(t)$ of unspecified form from data $(t, R_t), t = 1, \ldots, n$, where $E(R_t) = \mu(t)$. See, for instance, Fan and Gijbels (1992) and Fan (1993). Local linear kernel regression estimates $\mu(t_0)$ by a weighted average of the observations within the smoothing window $(t_0 - h, t_0 + h)$, where h is a user-specified parameter known as the bandwidth. The weights depend on a kernel function K chosen so that most weight is given to those observations with t closest to t_0.

If $\widehat{\mu}(\cdot)$ is the local linear kernel regression estimator of $\mu(\cdot)$, then we can write $(\widehat{\mu}(1), \ldots, \widehat{\mu}(n))^\top = \boldsymbol{S}_h(R_1, \ldots, R_n)^\top$, where $\boldsymbol{S}_h$ is an $n \times n$ hat matrix whose (t, t')-th element is given by:

$$[\boldsymbol{S}_h]_{t,t'} = \frac{K\left((t-t')/h\right)\left[\mathcal{S}_{n,2}(t) - (t-t')\mathcal{S}_{n,1}(t)\right]}{\mathcal{S}_{n,2}(t)\mathcal{S}_{n,0}(t) - \mathcal{S}_{n,1}(t)^2}.$$

Here,

$$\mathcal{S}_{n,l}(t) = \sum_{t''=1}^{n} K\left(\frac{t-t''}{h}\right)(t-t'')^l,\ l = 0, 1, 2.$$

The smoother matrix $\boldsymbol{S}_h^c$ appearing in the backfitting algorithm described in Section 2 is the centered version of the hat matrix $\boldsymbol{S}_h$; that is, $\boldsymbol{S}_h^c = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^\top/n)\boldsymbol{S}_h$, with $\boldsymbol{I}$ being the $n \times n$ identity matrix and $\boldsymbol{1}$ being an $n \times 1$ vector of ones.

In all analyses reported in this paper, we used the Epanechnikov kernel, given by

$$K(u) = \begin{cases} (3/4)(1-u^2), & if\ |u| < 1; \\ 0, & if\ |u| \geq 1. \end{cases}$$

Appendix B: Preliminary Model Selection

In what follows, we provide an overview of our preliminary model selection for the two data sets considered.

Our models treat log mortality as a continuous response and assume the relationship between PM10 and log mortality is linear for ease of interpretation. In all models, the seasonal effect is assumed to be smooth, of unspecified form. The models differ in their specification of the weather effects confounding this relationship.

All models were fitted using the S-Plus function *gam* with locally weighted linear regression (*loess*) as the smoothing method and the more stringent convergence parameters recommended by Dominici et al. (2002b). Loess differs from local linear kernel regression in that it uses a span instead of a bandwidth to control the size of the smoothing window. The span, expressed as a fixed proportion of nearest neighbors, determines the proportion of data points to be included in the smoothing window when performing local linear regression.

Competing models were compared via approximate F-tests, with the residual degrees of freedom obtained from the trace of the hat matrices associated with the model fits (Hastie and Tibshirani, 1990). Strictly speaking, the F-tests are valid only if the error terms in these models can be regarded as independent and identically distributed, having a normal distribution with mean 0 and constant variance σ^2; hence these tests provide only rough guidance in our context.

B.1 Mexico City Data

The first model we considered for the Mexico City data was a partially linear model, incorporating a linear PM10 effect β_1 and a smooth seasonal effect m of unspecified form:

$$log(D_t) = \beta_0 + \beta_1 PM_t + m(t) + \epsilon_t,\ t = 1, \ldots, 1096. \tag{9}$$

To identify a reasonable range of spans for estimating m by a univariate loess smoother, we fitted model (9) with spans ranging from 0.05 to 0.50 in increments of 0.01 and examined plots of the fitted m and corresponding model residuals. This suggested spans in the range 0.09 to 0.12, as spans smaller than 0.09 led to visually noisy fits, while spans larger than 0.12 led to over-smoothed fits that failed to reflect important seasonal features of the data (Figure 13). The reduction in residual variability achieved by including m in the model decreased from 47.9% to 46.4% as the span increased from 0.09 to 0.12. The degrees of freedom expended to achieve this reduction decreased from 19.9 to 14.7 as the span increased from 0.09 to 0.12.

The second model we considered for the Mexico City data was:

$$log(D_t) = \beta_0 + \beta_1 PM_t + m(t) + m_1(T_t, H_t) + \epsilon_t,\ t = 1, \ldots, 1096, \tag{10}$$

with m_1 a smooth bivariate weather surface. We compared model (10) to model (9) using approximate F-tests, allowing the span s for estimating m to range between 0.09 and 0.12, and the span s_1 for estimating m_1 to range between 0.05 and 0.50, in increments of 0.01. As seen in Figure 14, for $s = 0.09$ the resulting p-values were consistently greater than 0.40 (the reduction in residual variability associated with the addition of m_1 decreased from 7.6% to 0.8% and the degrees of freedom consumed to achieve this reduction decreased from 79.6 to 9.4 as s_1 increased from 0.05 to 0.50). Other values of s yielded similar results.

As m_1 contributed very little to model (10), we found no apparent need to include either temperature or relative humidity in the model. However, as these weather variables are typically included in models for air pollution/mortality data, we preferred model (1) to model (9); the cost of this expansion of the model is modest – only two degrees of freedom. Residual plots for fits of model (1) with spans s for estimating m of 0.09 (Figure 15), 0.10, 0.11 or 0.12 revealed no systematic

structure, suggesting that the PM10, seasonal and weather components of the model account for the structure in the data reasonably well. The extent of the serial correlation in these sets of residuals was small (the lag 1-4 serial correlations varied between 0.02 and 0.09 for the different values of s; see Figure 16 for $s = 0.09$), likely due to the fact that most of the short-term temporal variation in the log mortality counts is accounted for by m, the seasonal component. Comparing to the extent of serial correlation in the Mexico City raw log mortality counts supported this belief (the lag 1 serial correlation was about 0.54 and the serial correlations diminished quite slowly with increasing lag, as seen in Figure 17).

B.2 Vancouver Data

Our preliminary model selection for the Vancouver data considered the models:

$$log(D_t) = \beta_0 + \beta_1 \cdot PM_t + m(t) + \epsilon_t, \tag{11}$$

$$log(D_t) = \beta_0 + \beta_1 \cdot PM_t + m(t) + m_1(T_t, H_t) + \epsilon_t, \tag{12}$$

$$log(D_t) = \beta_0 + \beta_1 \cdot PM_t + m(t) + \beta_2 \cdot T_t + \beta_3 \cdot H_t + \epsilon_t, \tag{13}$$

$$log(D_t) = \beta_0 + \beta_1 \cdot PM_t + m(t) + \beta_2 \cdot T_t + \beta_3 \cdot H_t + m_2(P_t, R_t) + \epsilon_t, \tag{14}$$

as well as model (8). Here, m is a smooth seasonal effect of unspecified form and m_1 and m_2 are smooth bivariate surfaces of unspecified form. As for Mexico City, $t = 1, \ldots, 1096$.

We first fitted model (11) to the Vancouver data. Using the same approach as for the Mexico City data, we determined that the seasonal effect m in this model could be reasonably estimated with spans s in the range 0.10 to 0.20 (Figure 18). The resulting estimates of m consumed degrees of freedom decreasing from 17.9 for $s = 0.10$ to 8.7 for $s = 0.20$. The reduction in residual variability achieved by including m in the model decreased from 17.4% to 15.3% as s increased from 0.10 to 0.20. Note that the seasonal component is much stronger in Mexico City, as is also evident from Figures 1 and 7.

Next, we compared model (12) to (11) via approximate F-test conducted with spans s for estimating m ranging between 0.10 and 0.20 and spans s_1 for estimating m_1 ranging between 0.32 and 0.50, in increments of 0.01 (spans s_1 smaller than 0.32 were deemed inappropriate for estimating the surface m_1, as they yielded visually rough surfaces that consumed high numbers of degrees of freedom: $df \geq 14.9$; Figure 19 displays the estimated weather surface m_1 for $s = 0.15$ and $s1 = 0.05, 0.15, 0.30$ and 0.50). For all values of s and s_1 considered, the resulting p-values were larger than 0.30. For $s = 0.15$, the reduction in residual variability associated with the addition of m_1 decreased from about 1.2% to about 0.8% as s_1 increased from 0.32 to 0.50; the degrees of freedom expended to achieve this reduction decreased from 13.6 to 9.2 as s_1 increased from 0.32 to 0.50 (Figure 20). Other values of s yielded similar results. As the smooth surface m_1 contributed little to model (12), we concluded that model (12) did not offer a clear improvement over model (11). Since including effects for temperature and relative humidity is common and linear effects cost only 2 degrees of freedom, we prefer model (13) to model (11).

Even though we do not need to include either barometric pressure or rain in our model (the reduction in residual variability achieved by including m_2 in the model decreased from about 3.1% to about 0.9% and the degrees of freedom expended to achieve this reduction decreased from about 28.6 to 9.3 as s_2 increased from 0.15 to 0.50), we chose to proceed with model (8), which includes linear effects for barometric pressure and rain. Residual plots for fits of model (8) with spans s for estimating m in the range 0.10 to 0.20 showed that this model fits the Vancouver data reasonably well (the residual plots corresponding to $s = 0.15$ are shown in Figure 21). We found very little indication of serial correlation in these sets of residuals (the lag 1-4 serial correlations varied between -0.03 and 0.03; see Figure 22 for $s = 0.15$), reflecting the relatively weak serial correlation in the Vancouver raw log mortality counts (the lag 1 serial correlation was only about 0.18, as seen in Figure 23).

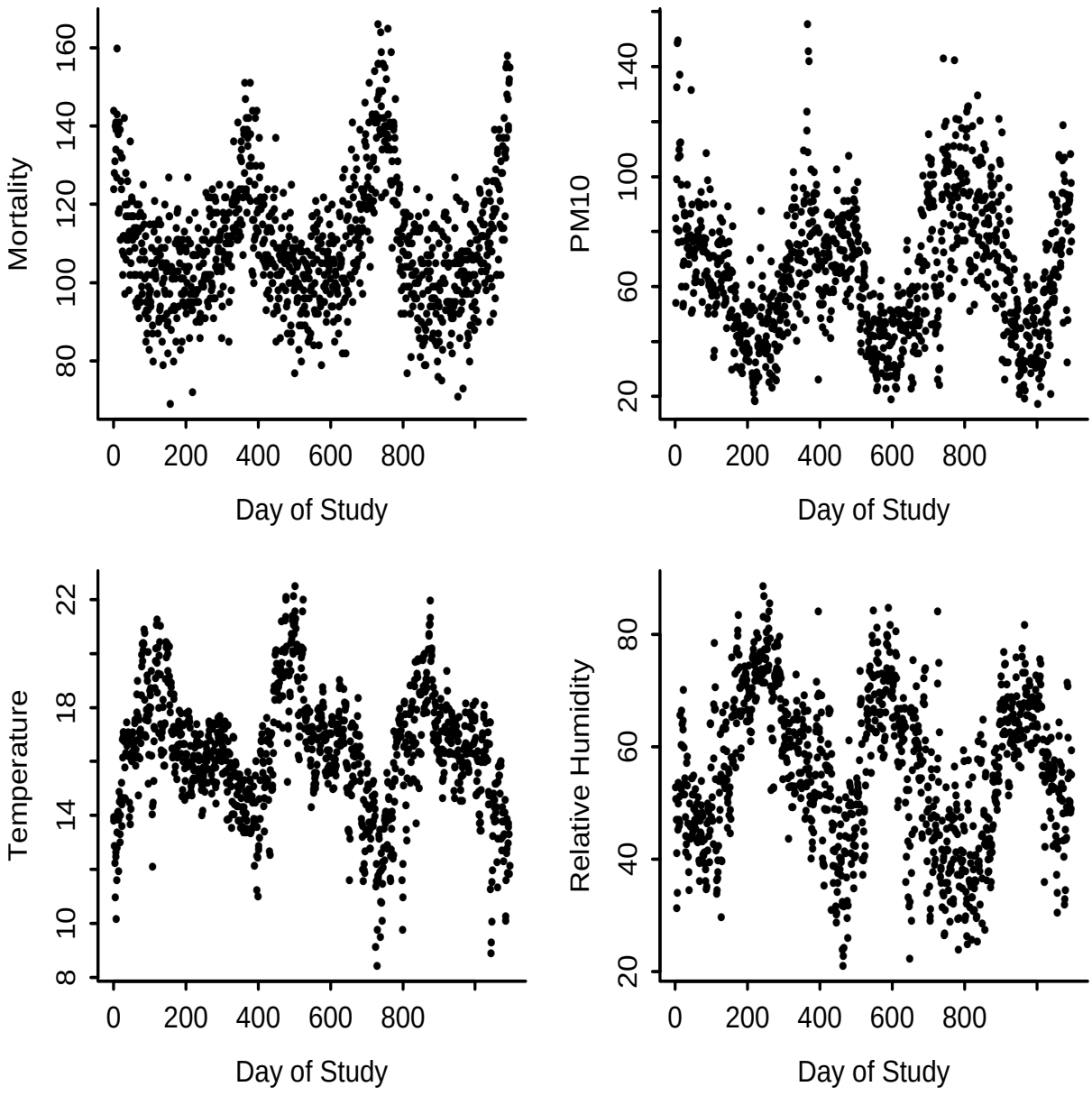

Figure 1: Scatterplots of mortality, PM10, temperature and relative humidity versus day of study for the Mexico City data.

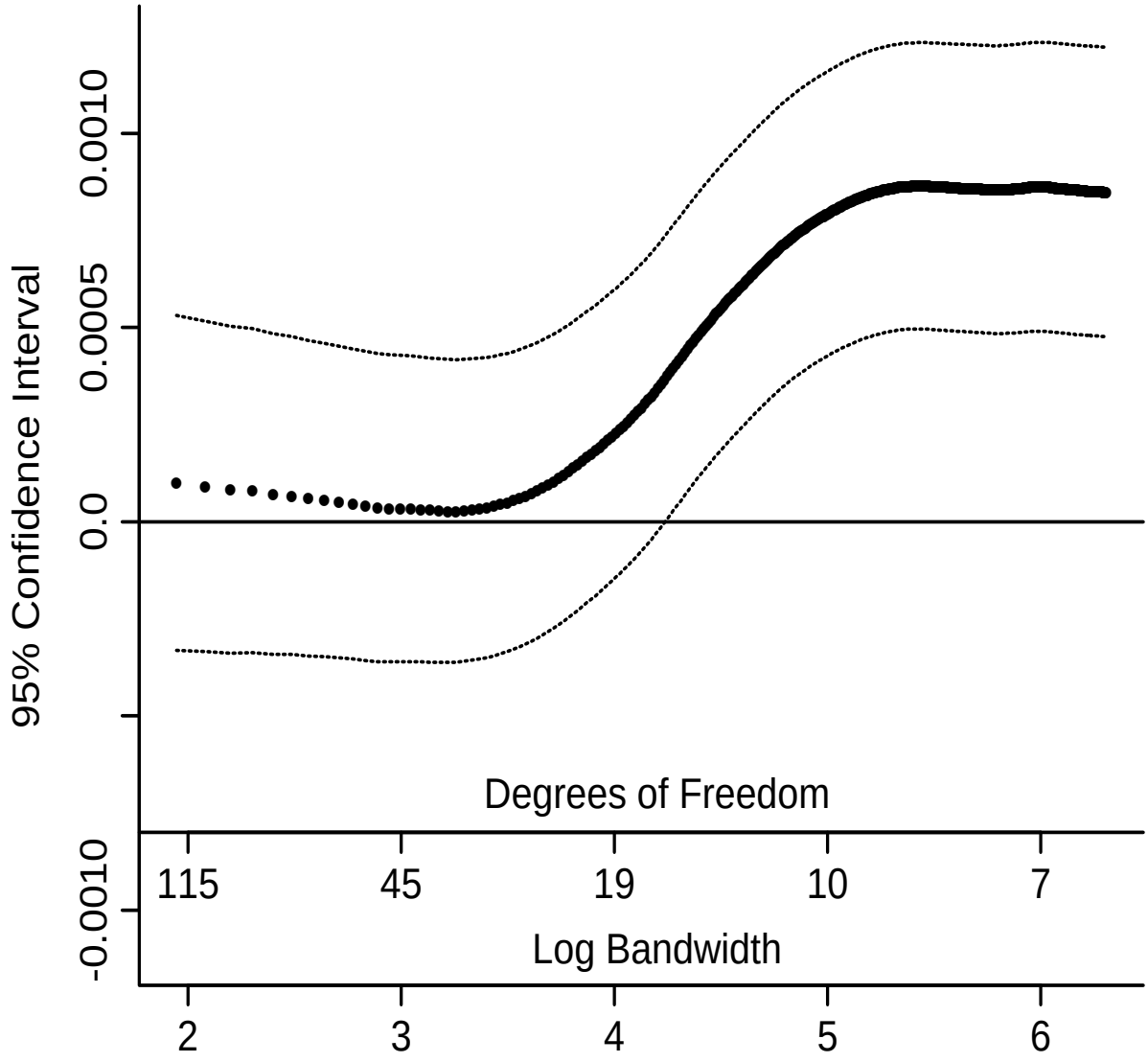

Figure 2: Local linear backfitting estimates and naive 95% confidence intervals for the linear PM10 effect β_1 in model (1), describing the Mexico City data, as a function of either the log bandwidth used for estimating the seasonal effect m or the degrees of freedom consumed for estimating the model. The naive 95% confidence intervals were calculated ignoring residual correlation. The bandwidth was allowed to range from 7 days to 365 days.

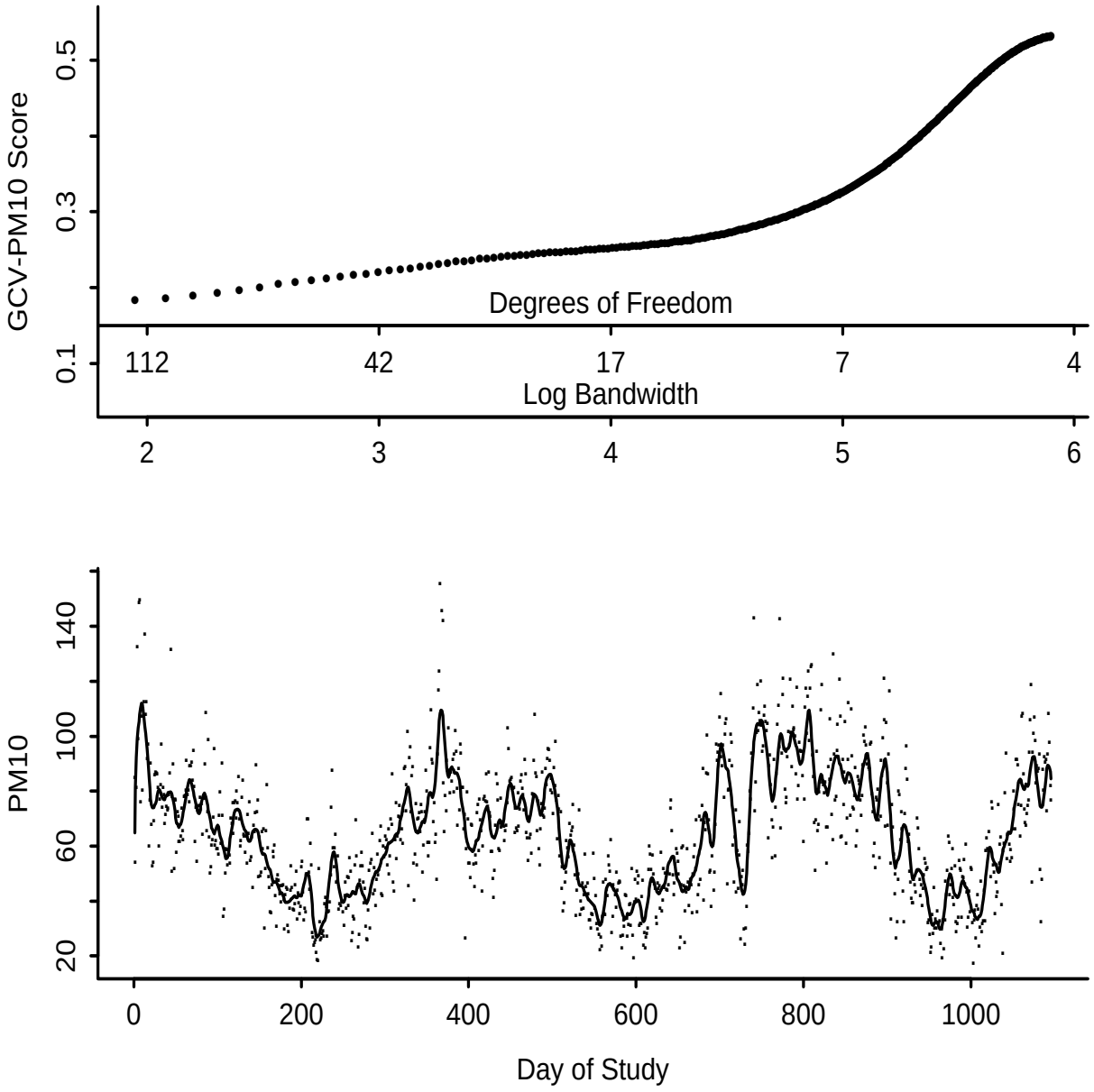

Figure 3: Mexico City data. *Top:* Generalized cross-validation score as a function of either the log bandwidth used for smoothing PM10 on day of study or the degrees of freedom consumed by the smooth of PM10. The smoothing was performed via local linear kernel regression. The bandwidth was allowed to range from 7 days to 365 days. *Bottom:* Scatterplot of PM10 versus day of study. The smooth of PM10 versus day of study for the GCV-PM10 choice of bandwidth (7 days) is superimposed.

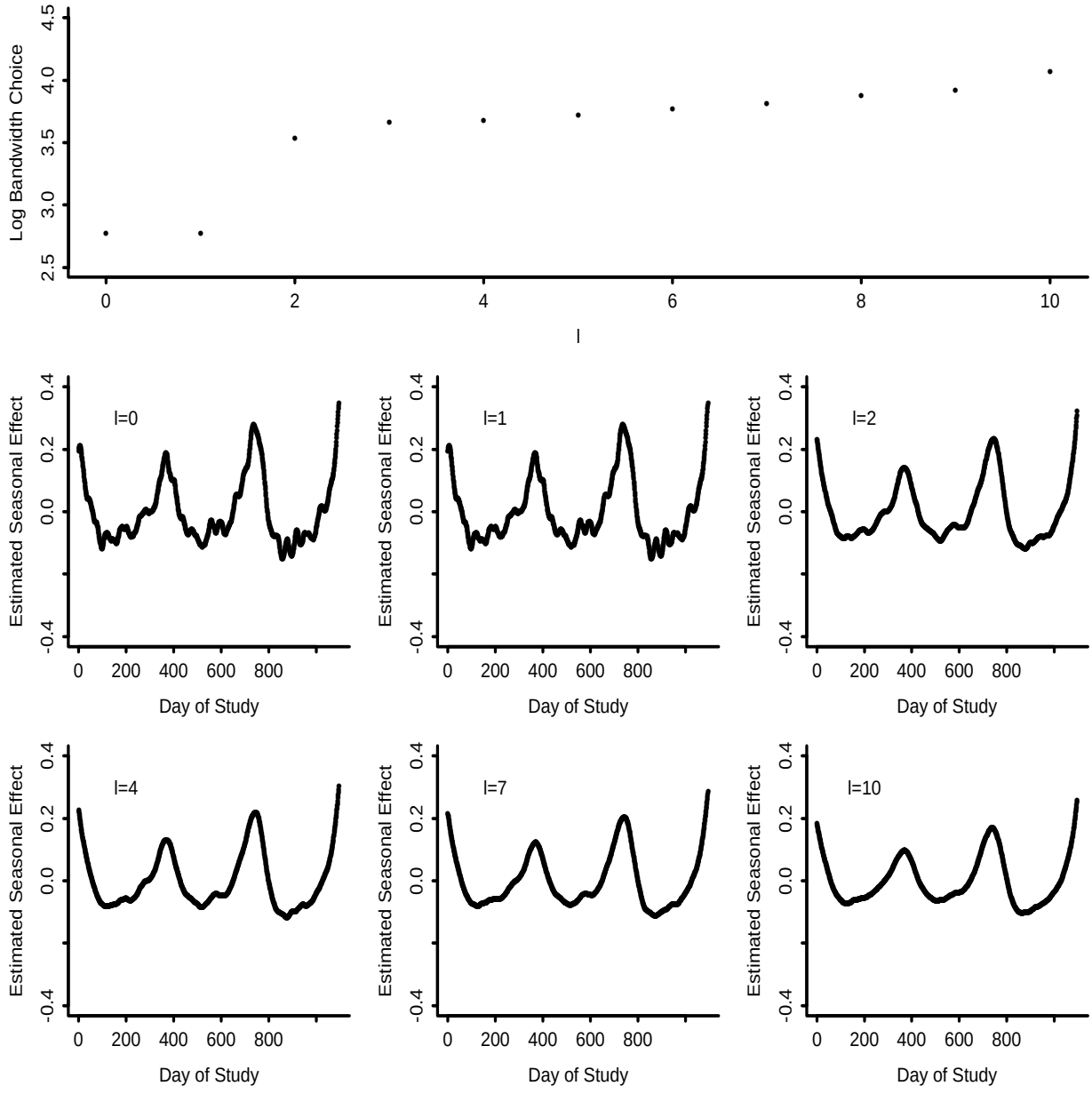

Figure 4: *Top panel:* Log bandwidth for estimating the seasonal effect m in model (1), describing the Mexico City data, as chosen by leave-$(2l+1)$-out crossvalidation, where $l = 0, \ldots, 10$. *Bottom panels:* Preliminary estimates of the seasonal effect m, obtained with selected leave-$(2l+1)$-out cross-validation choices of bandwidth.

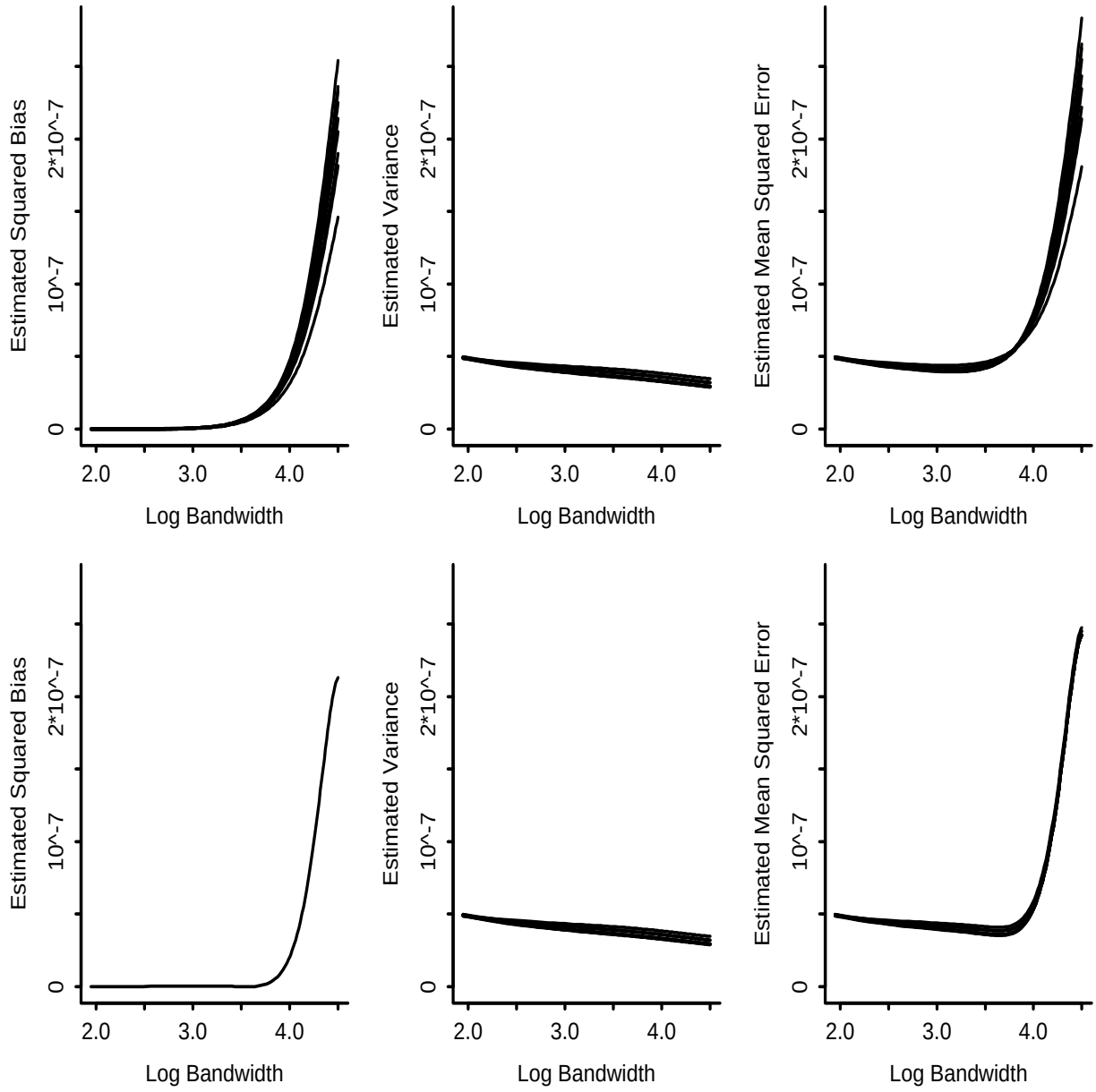

Figure 5: Mexico City data. The estimated squared bias, variance and mean squared error of $\widehat{\beta}_{1,h}$, the local linear backfitting estimate of the linear PM10 effect β_1 in model (1), versus *log* h, for different values of l ($l = 2, \ldots, 10$) in the leave-$(2l+1)$-out cross-validation. *Top:* Plug-in method. *Bottom:* Global empirical method.

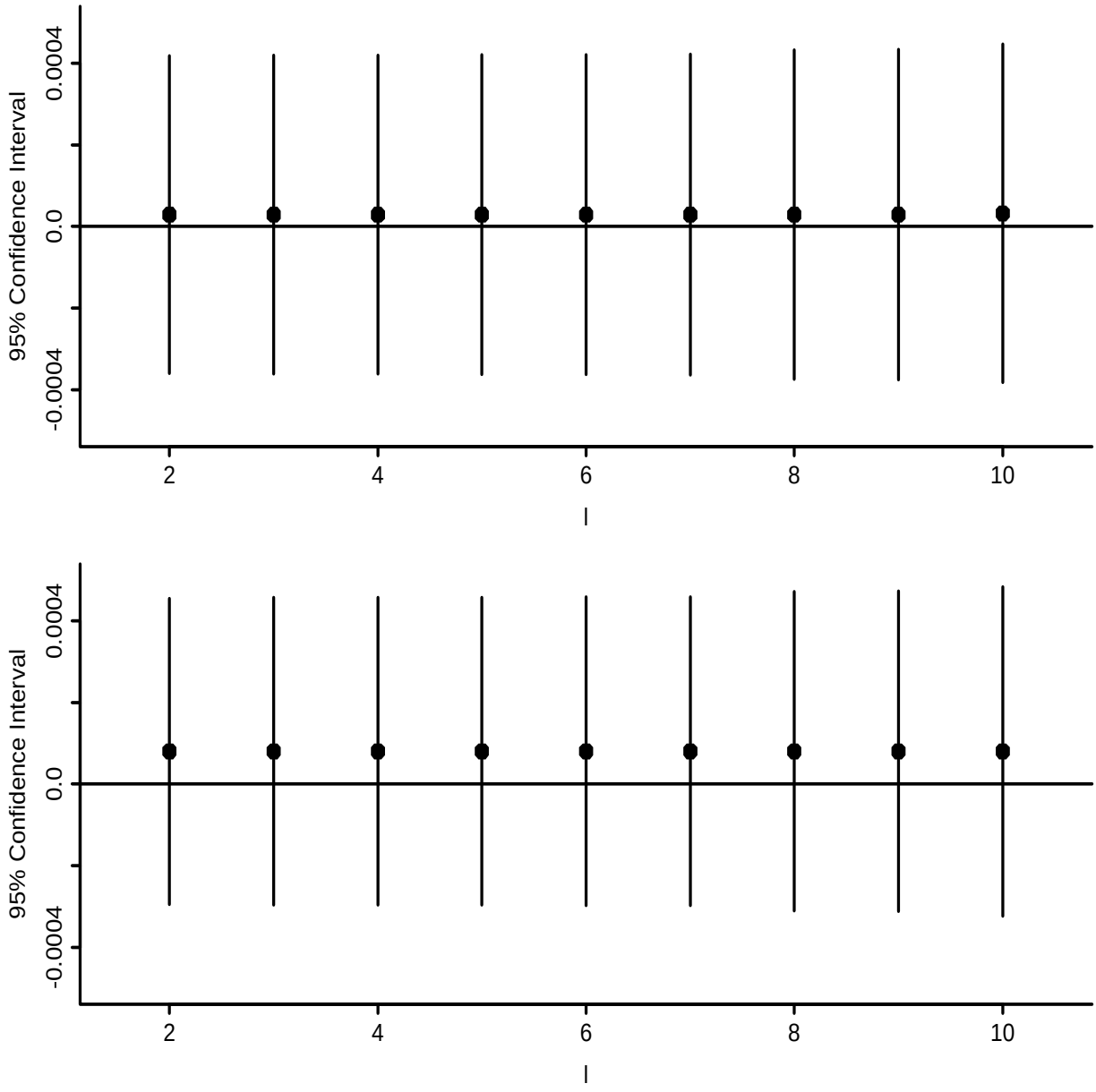

Figure 6: Mexico City data. 95% confidence intervals for β_1, the linear PM10 effect in model (1), as a function of l, the tuning parameter in leave-$(2l+1)$-our cross-validation. The values of l range from 2 to 10. *Top:* Plug-in method. *Bottom:* Global empirical method.

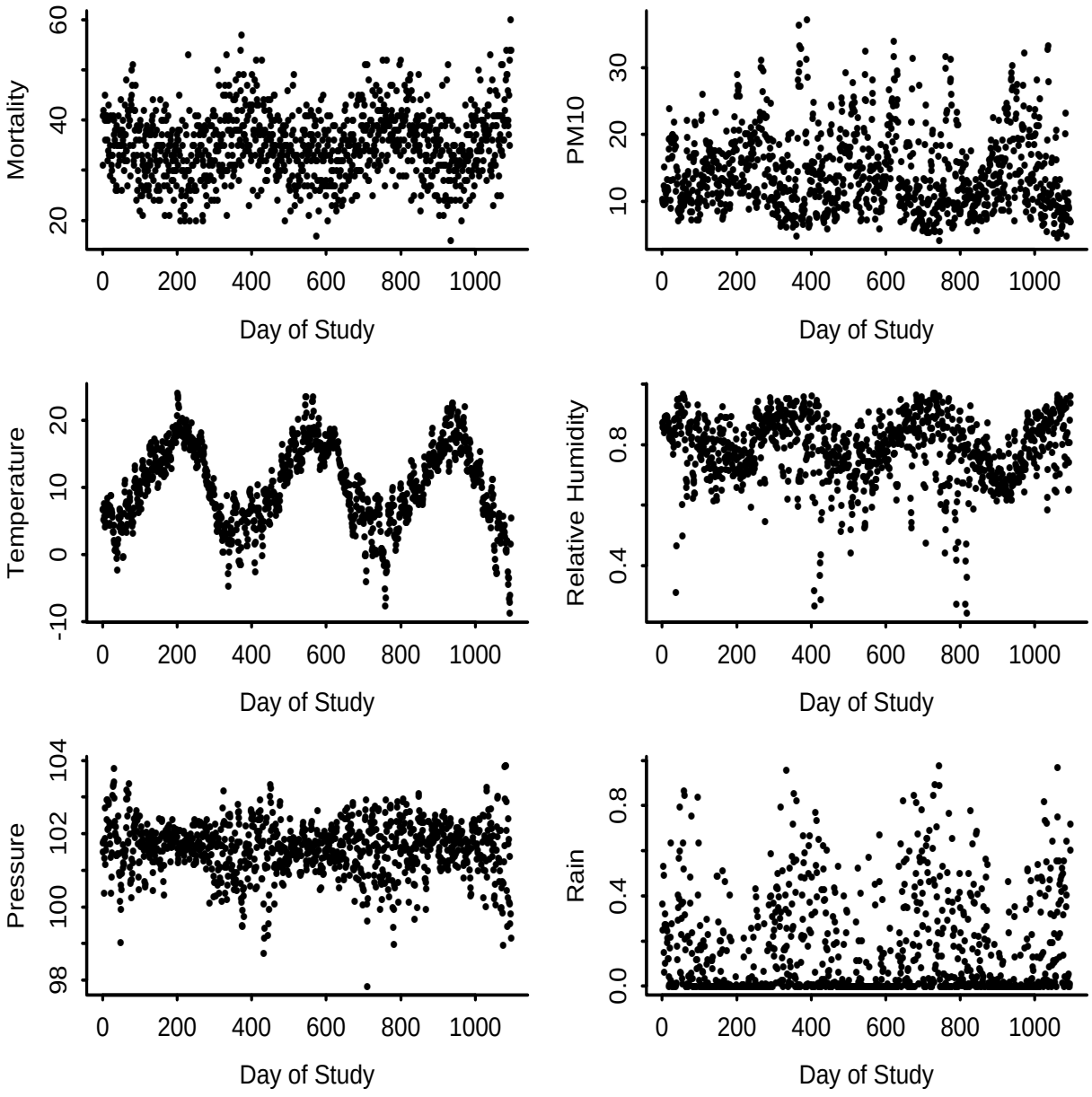

Figure 7: Scatterplots of mortality, PM10, temperature, relative humidity, barometric pressure and rain versus day of study for the Vancouver data.

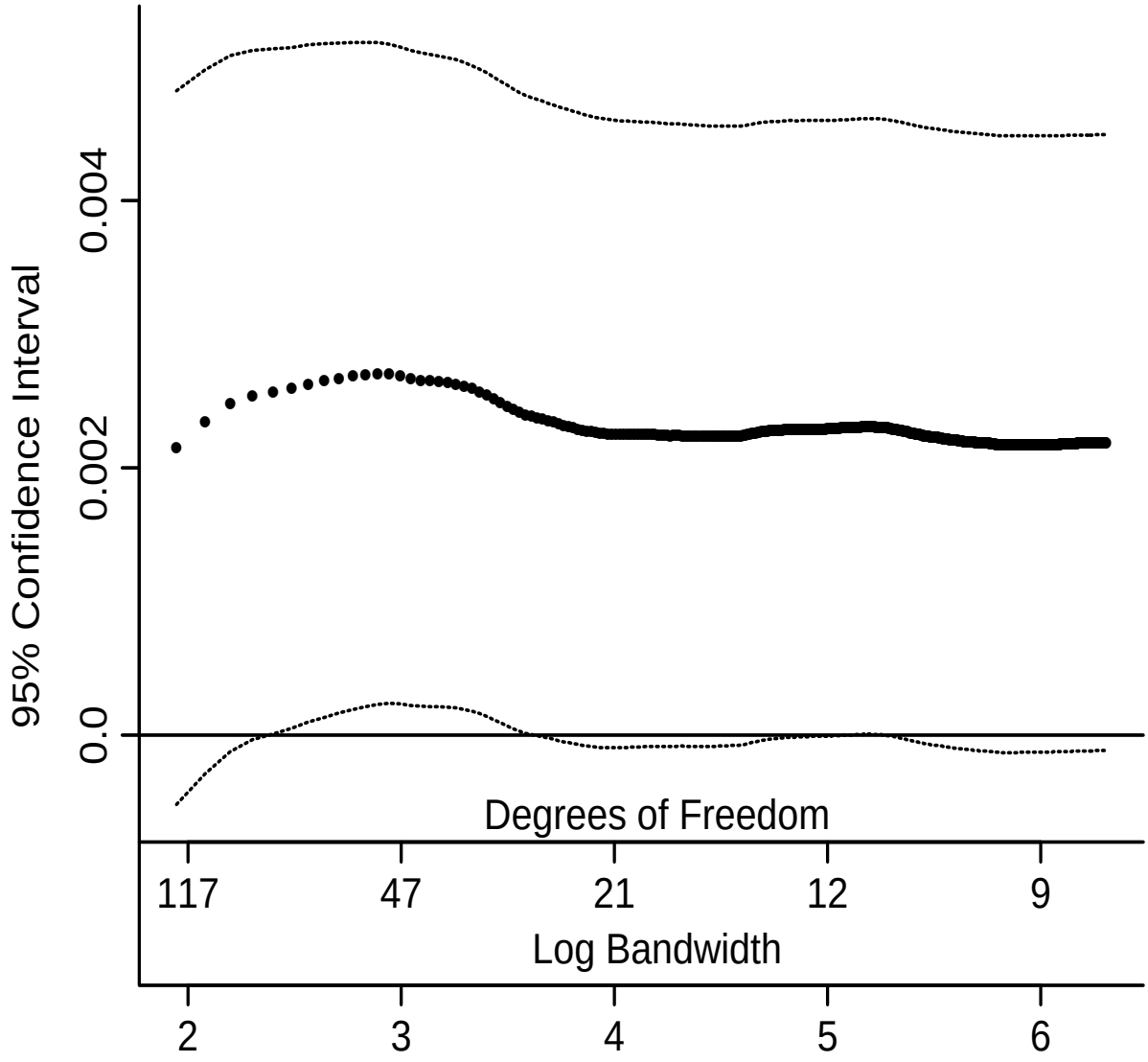

Figure 8: Local linear backfitting estimates and naive 95% confidence intervals for the linear PM10 effect β_1 in model (8), describing the Vancouver data, as a function of either the log bandwidth used for estimating the seasonal effect m or the degrees of freedom consumed for estimating the model. The naive 95% confidence intervals were calculated ignoring residual correlation. The bandwidth was allowed to range from 7 days to 365 days.

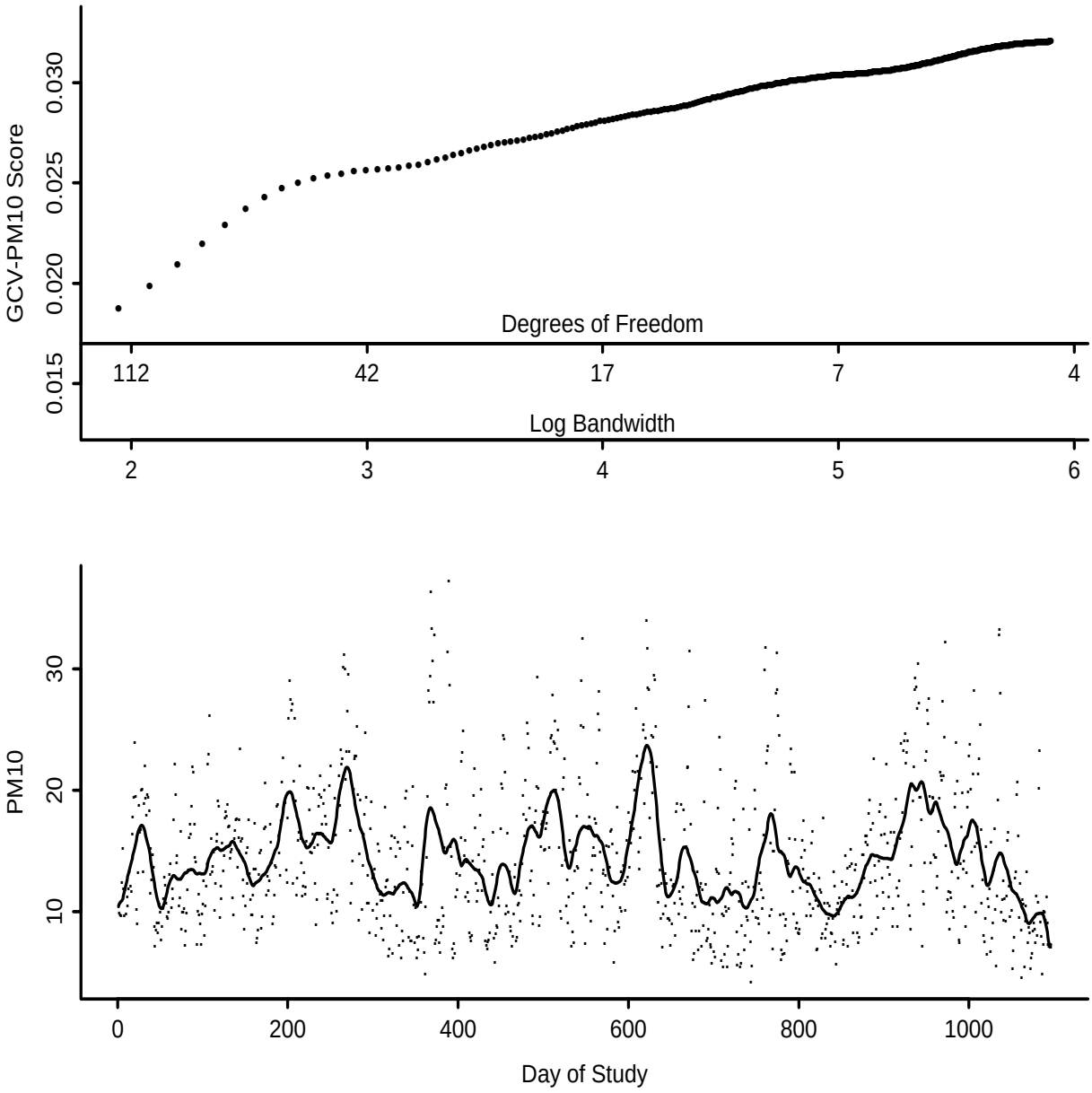

Figure 9: Vancouver data. *Top :* Generalized cross-validation score as a function of either the log bandwidth used for smoothing PM10 on day of study or the degrees of freedom consumed by the smooth of PM10. The smoothing was performed via local linear kernel regression. The bandwidth was allowed to range from 7 days to 365 days. *Bottom:* Scatterplot of PM10 versus day of study. The smooth of PM10 versus day of study for the GCV-PM10 choice of bandwidth (7 days) is superimposed.

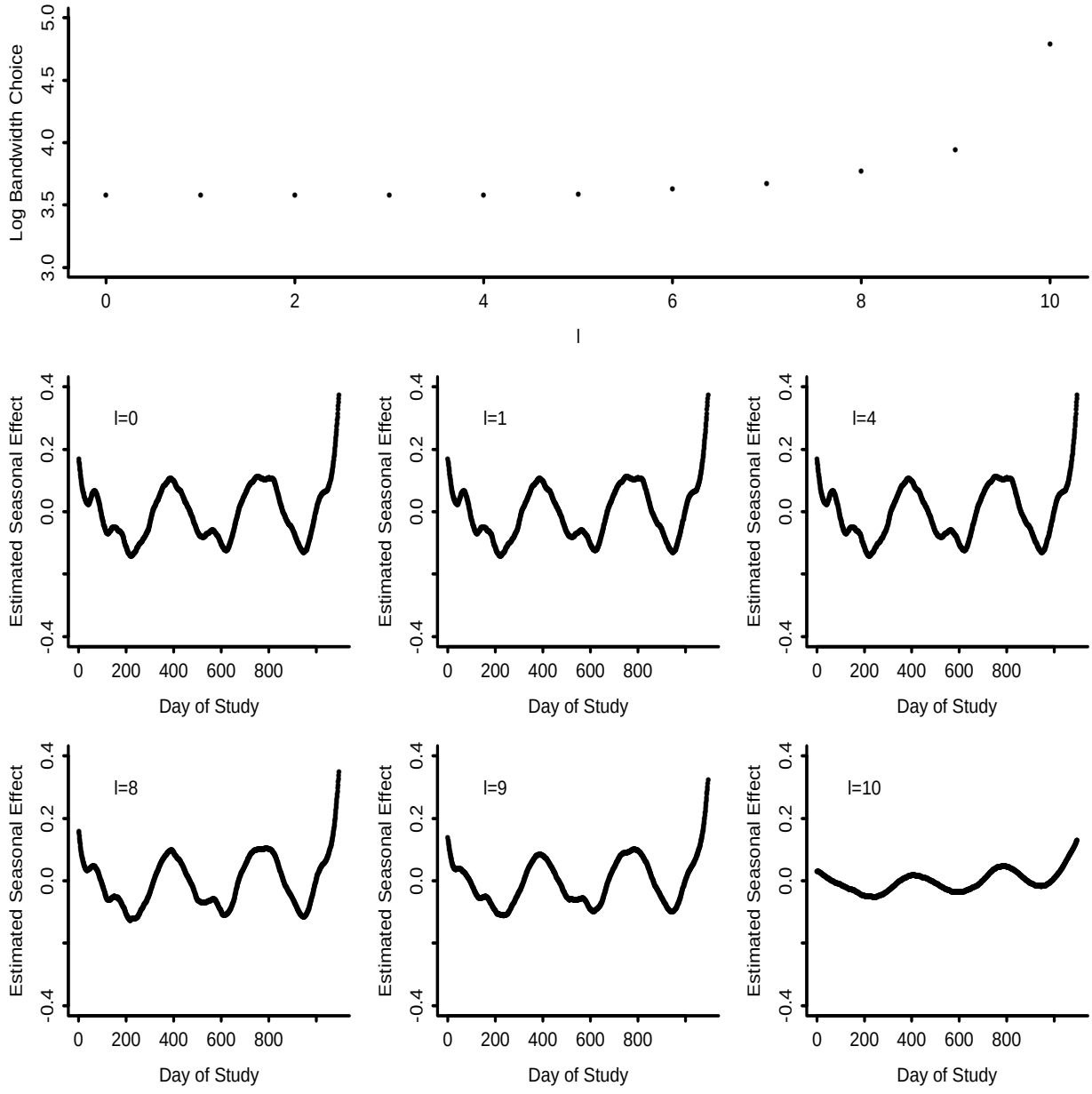

Figure 10: *Top panel:* Log bandwidth for estimating the seasonal effect m in model (8), describing the Vancouver data, as chosen by leave-$(2l+1)$-out crossvalidation. *Bottom panels:* Preliminary estimates of the seasonal effect m, obtained with selected leave-$(2l+1)$-out cross-validation choices of bandwidth.

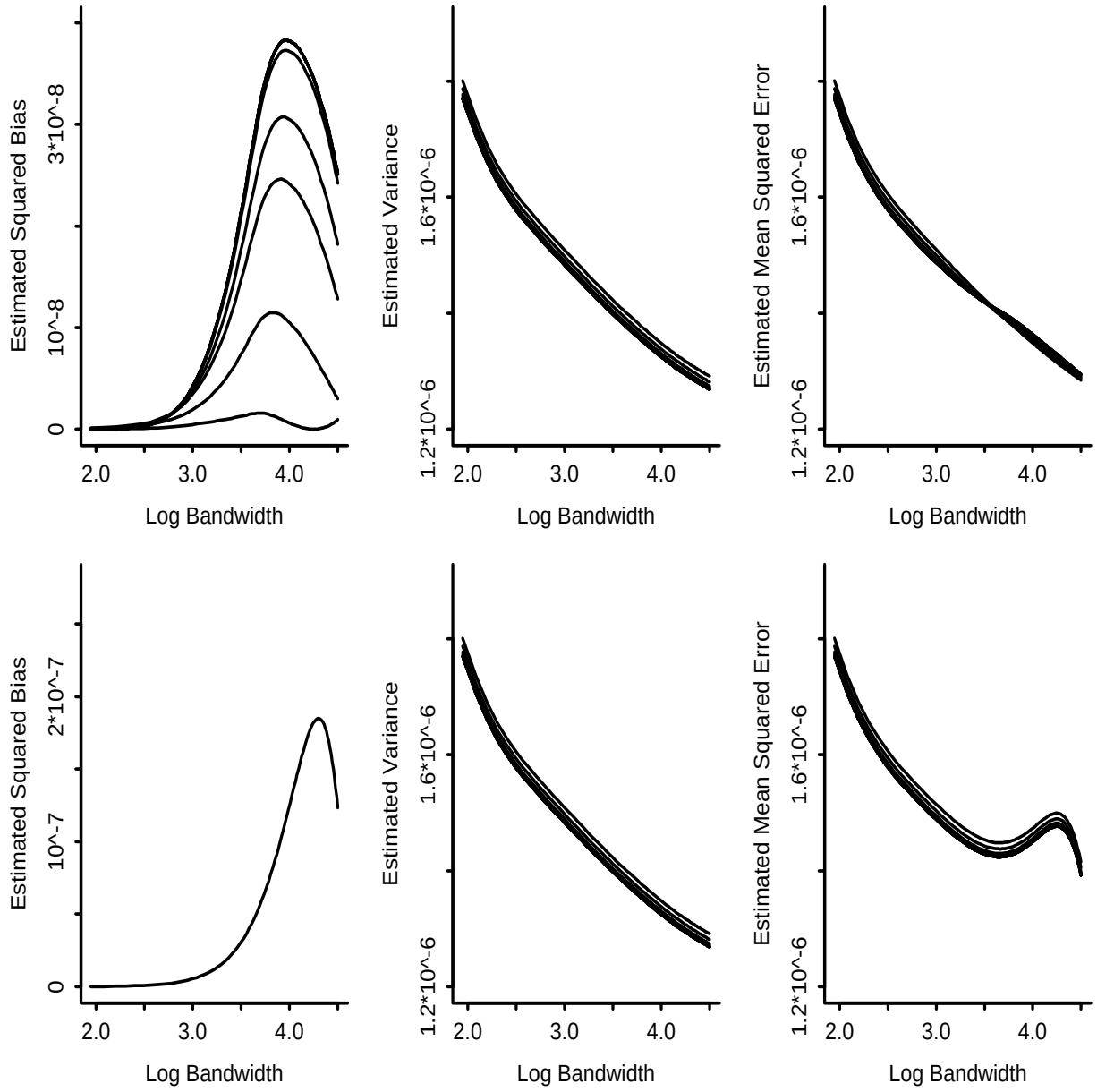

Figure 11: Vancouver data. The estimated squared bias, variance and mean squared error of $\widehat{\beta}_{1,h}$, the local linear backfitting estimate of the linear PM10 effect β_1 in model (8), versus *log* h, for different values of l ($l = 0, 1, \ldots, 9$) in the leave-$(2l+1)$-out cross-validation. *Top:* Plug-in method. *Bottom:* Global empirical method.

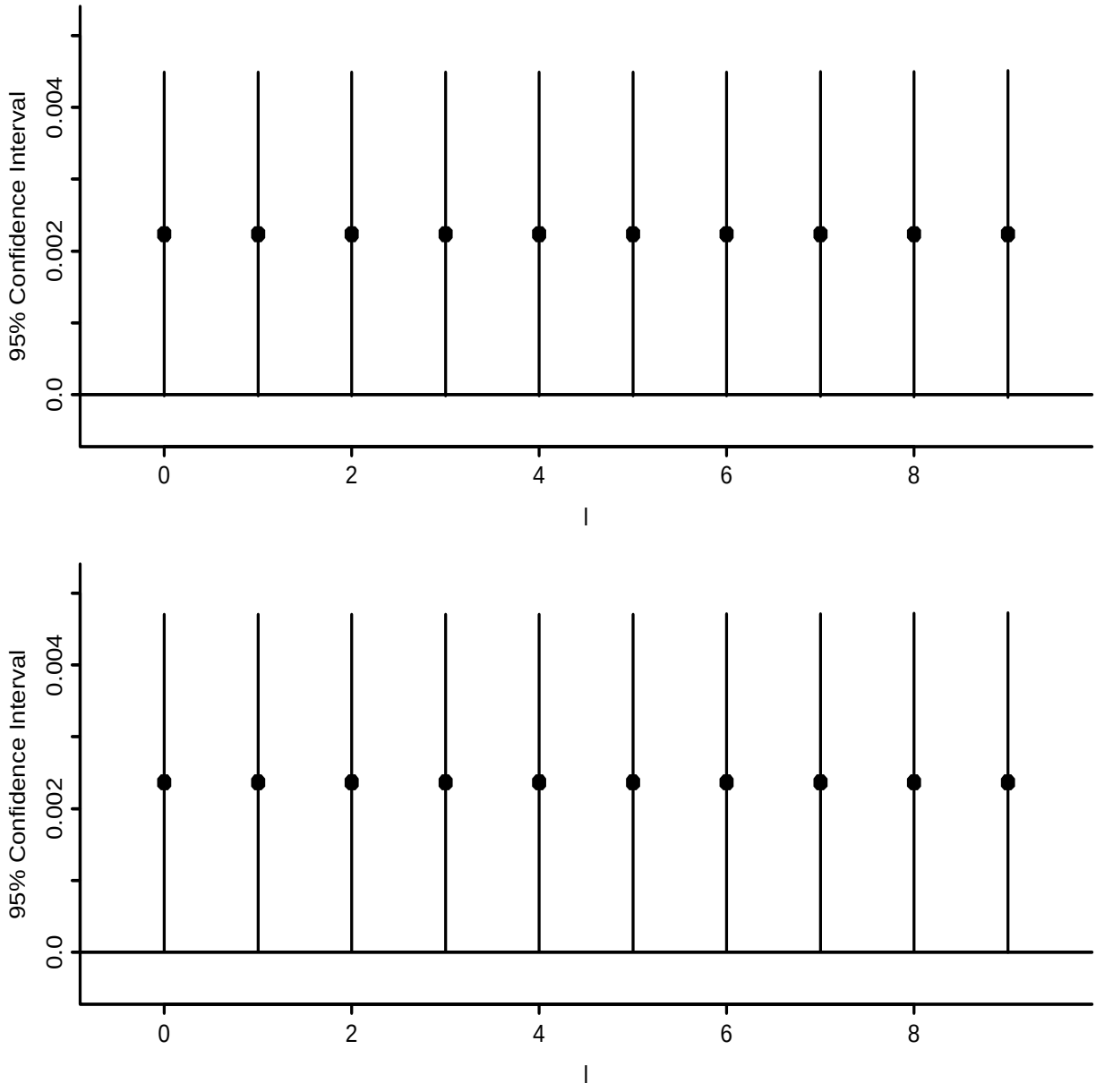

Figure 12: Vancouver data. 95% confidence intervals for β_1, the linear PM10 effect in model (8), as a function of l, the tuning parameter in leave-$(2l+1)$-out cross-validation. The values of l range from 0 to 9. *Top:* Plug-in method. *Bottom:* Global empirical method.

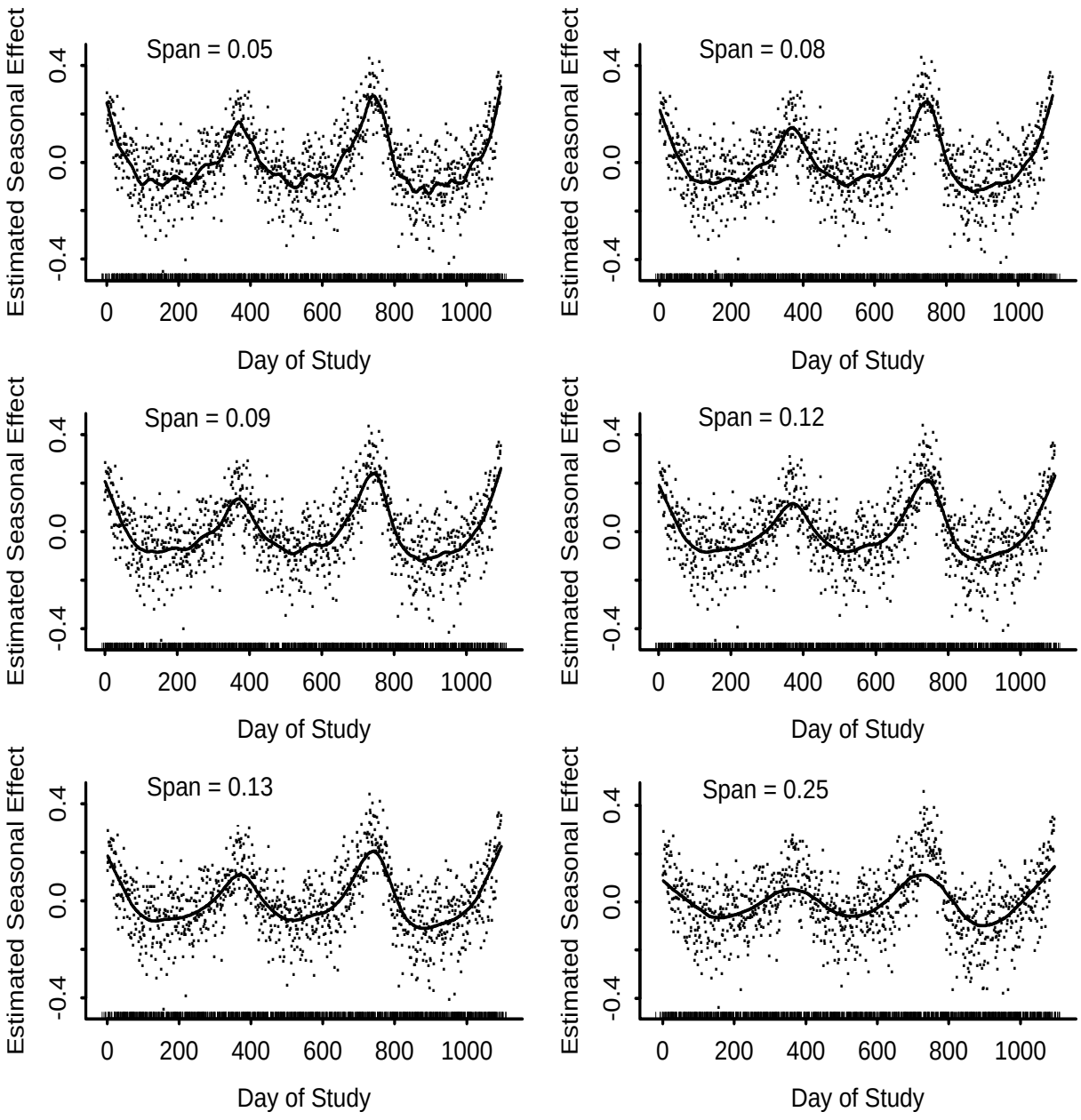

Figure 13: Plots of the fitted seasonal effect m in model (9), describing the Mexico City data, for various spans. Partial residuals, obtained by subtracting the fitted parametric part of the model from the responses, are superimposed as dots.

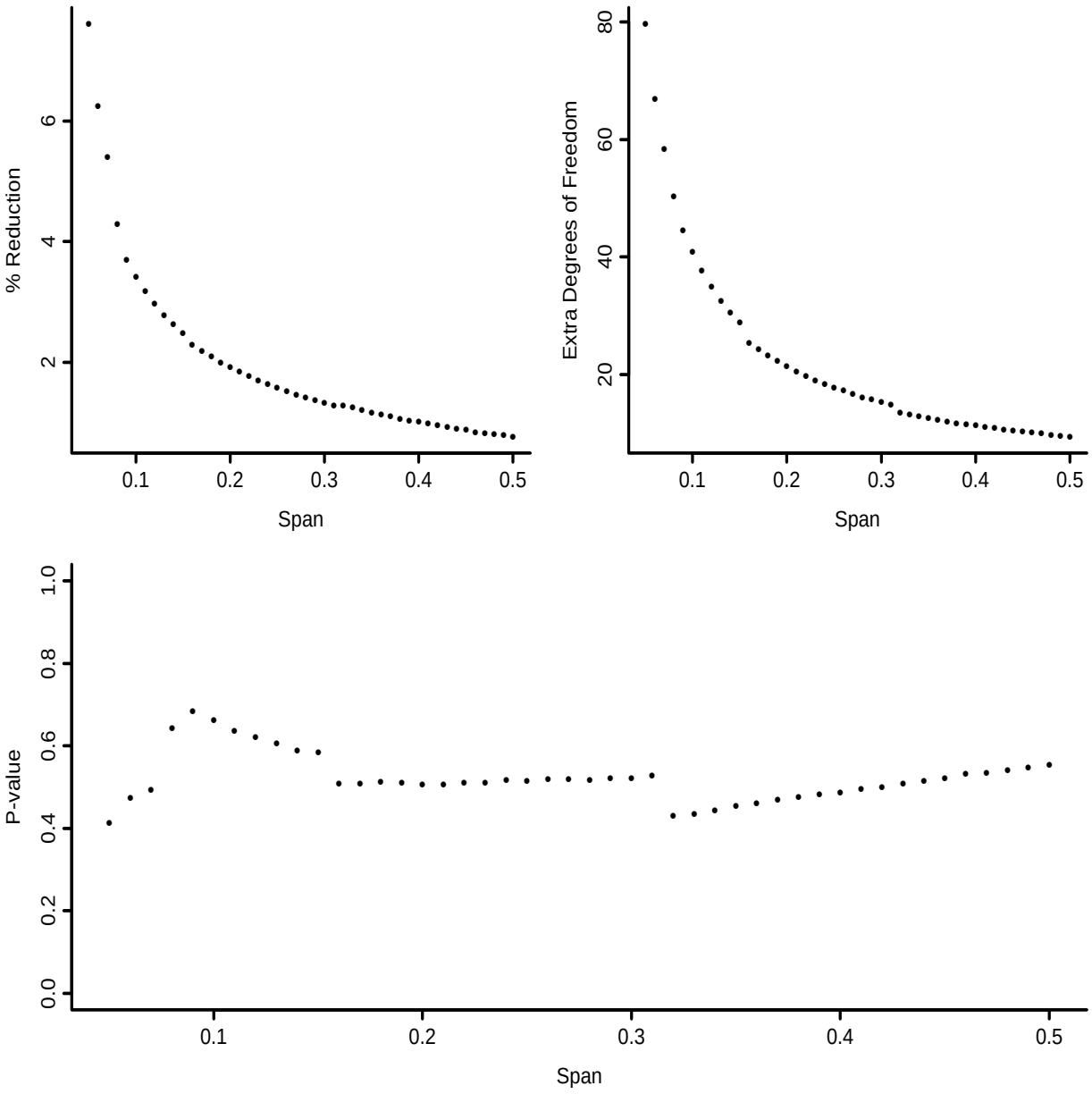

Figure 14: *Top:* Percent reduction in residual variability achieved by switching from model (9) to (10), together with the degrees of freedom expanded to achieve this reduction. Both of these models describe the Mexico City data. *Bottom:* P-values associated with a series of crude F-tests for testing model (10) against model (9). The seasonal effect m was estimated with a span of 0.09.

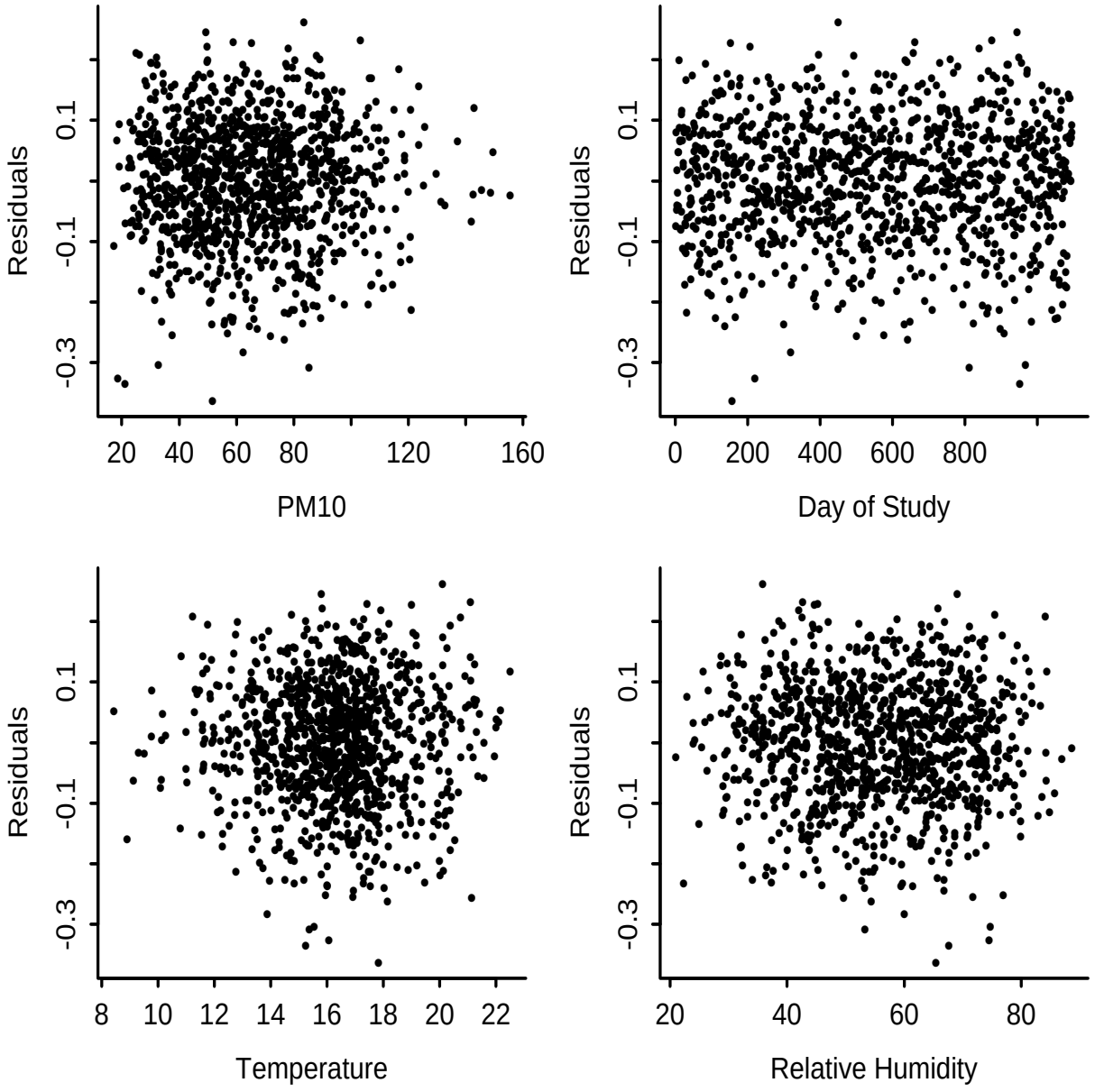

Figure 15: Plot of residuals associated with model (1), describing the Mexico City data, versus PM10 (top row) and day of study (bottom row). The span used for estimating the seasonal effect m in model (1) is 0.09.

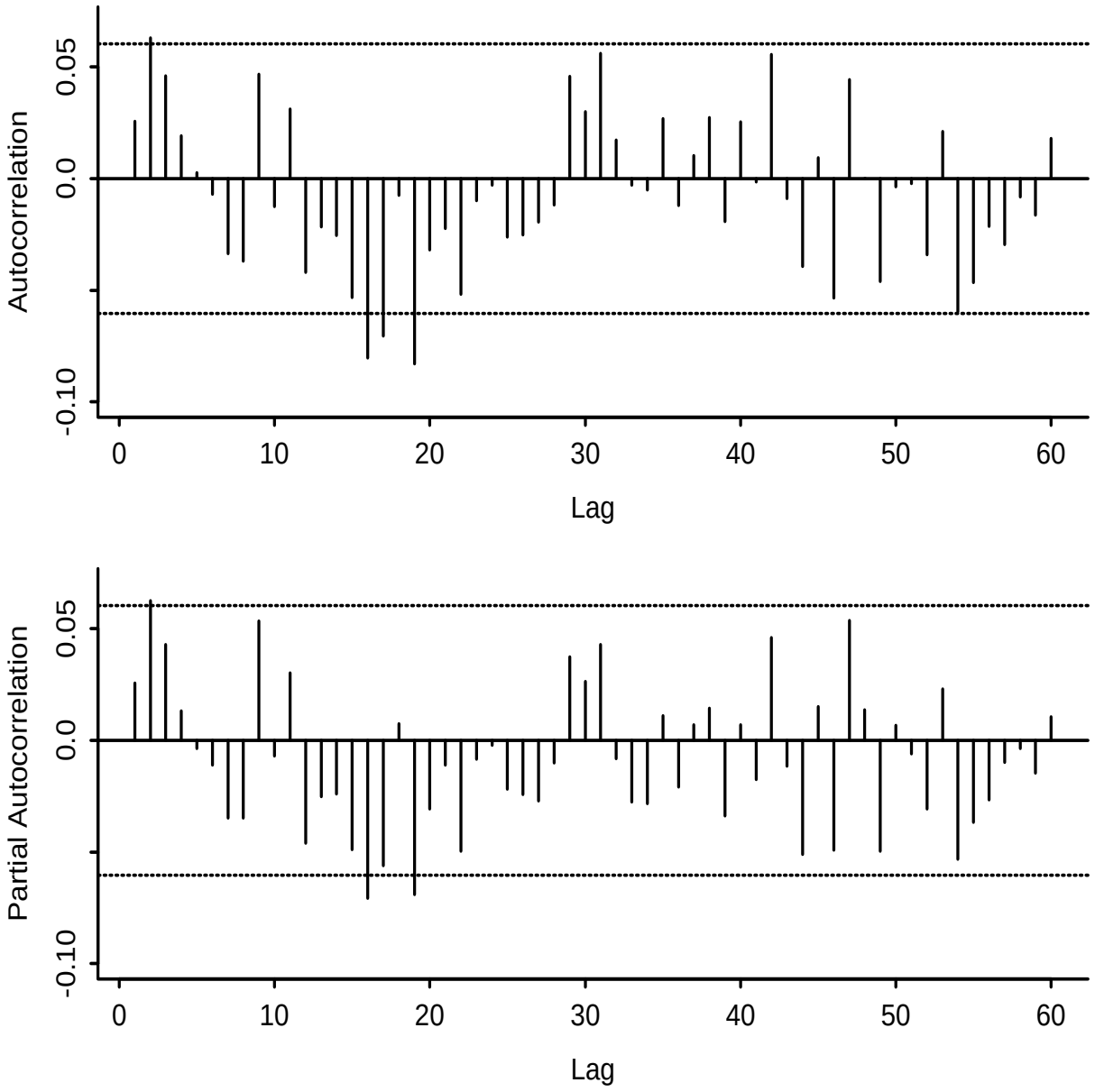

Figure 16: *Top:* Autocorrelation and partial autocorrelation plots for the residuals associated with model (1), describing the Mexico City data. These residuals were obtained by estimating the seasonal effect m with a span of 0.09.

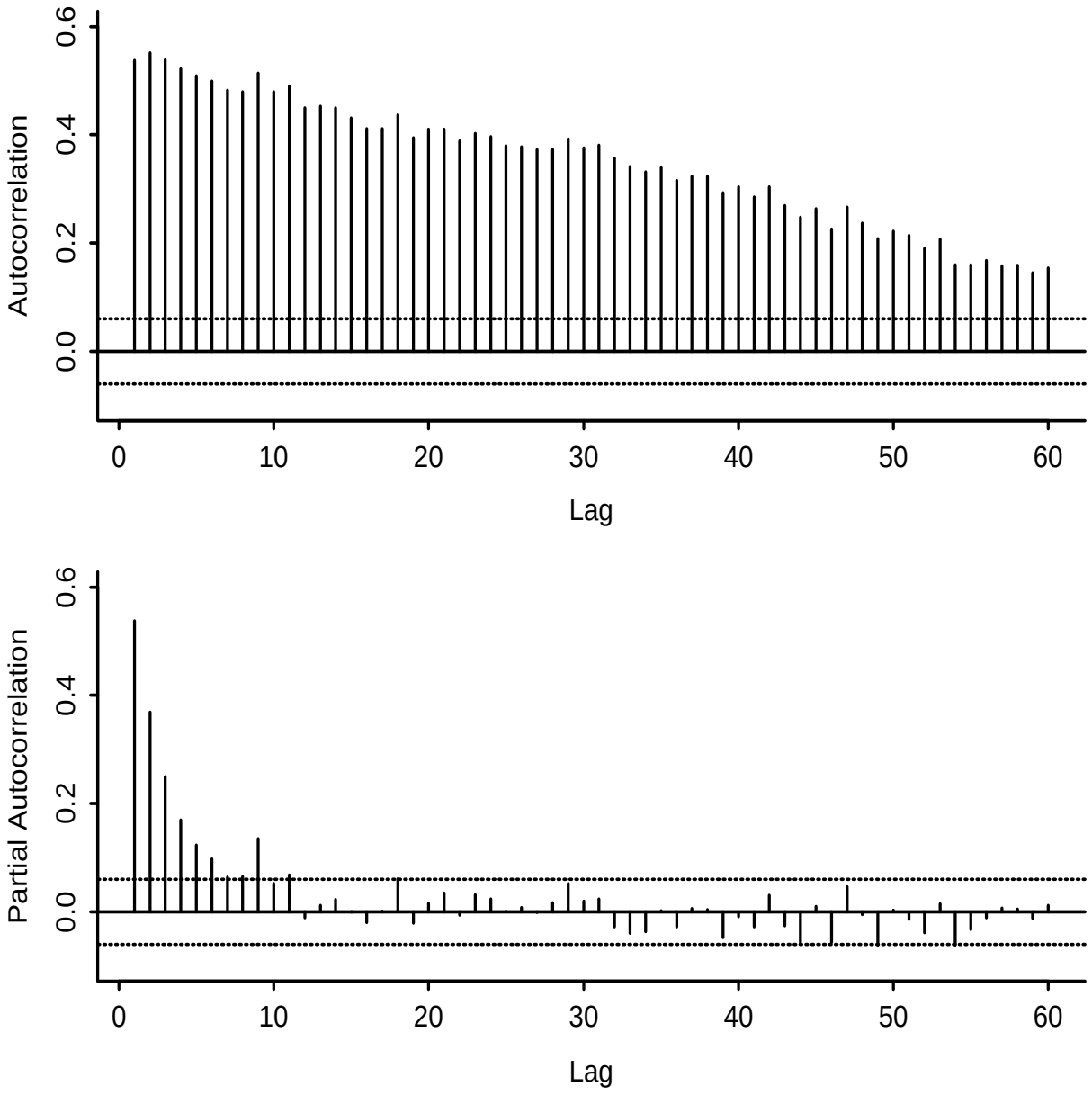

Figure 17: Autocorrelation and partial autocorrelation plots for the Mexico City log mortality counts.

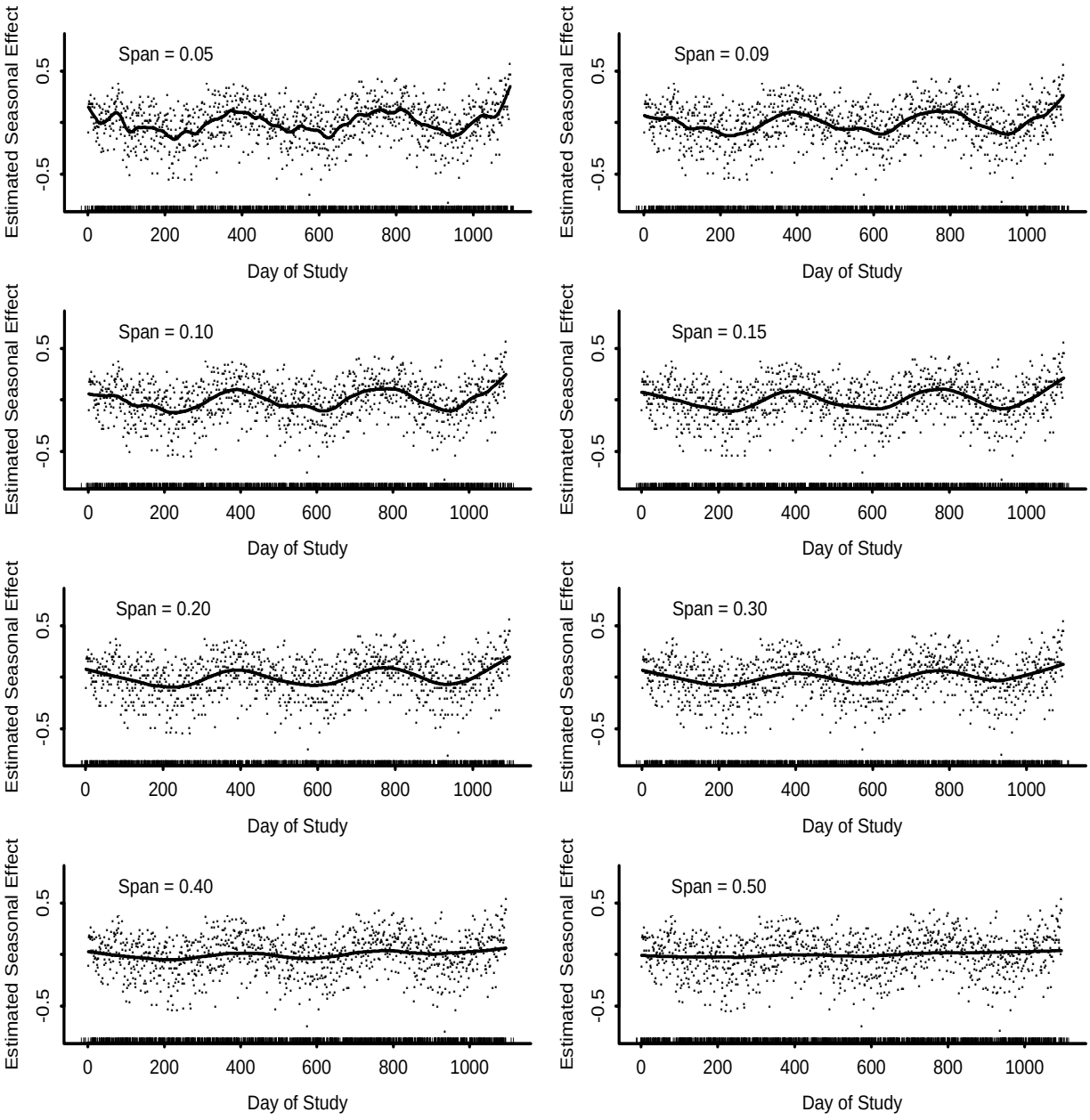

Figure 18: Plots of the fitted seasonal effect m in model (11), describing the Vancouver data, for various spans. Partial residuals, obtained by subtracting the fitted parametric part of the model from the responses, are superimposed as dots.

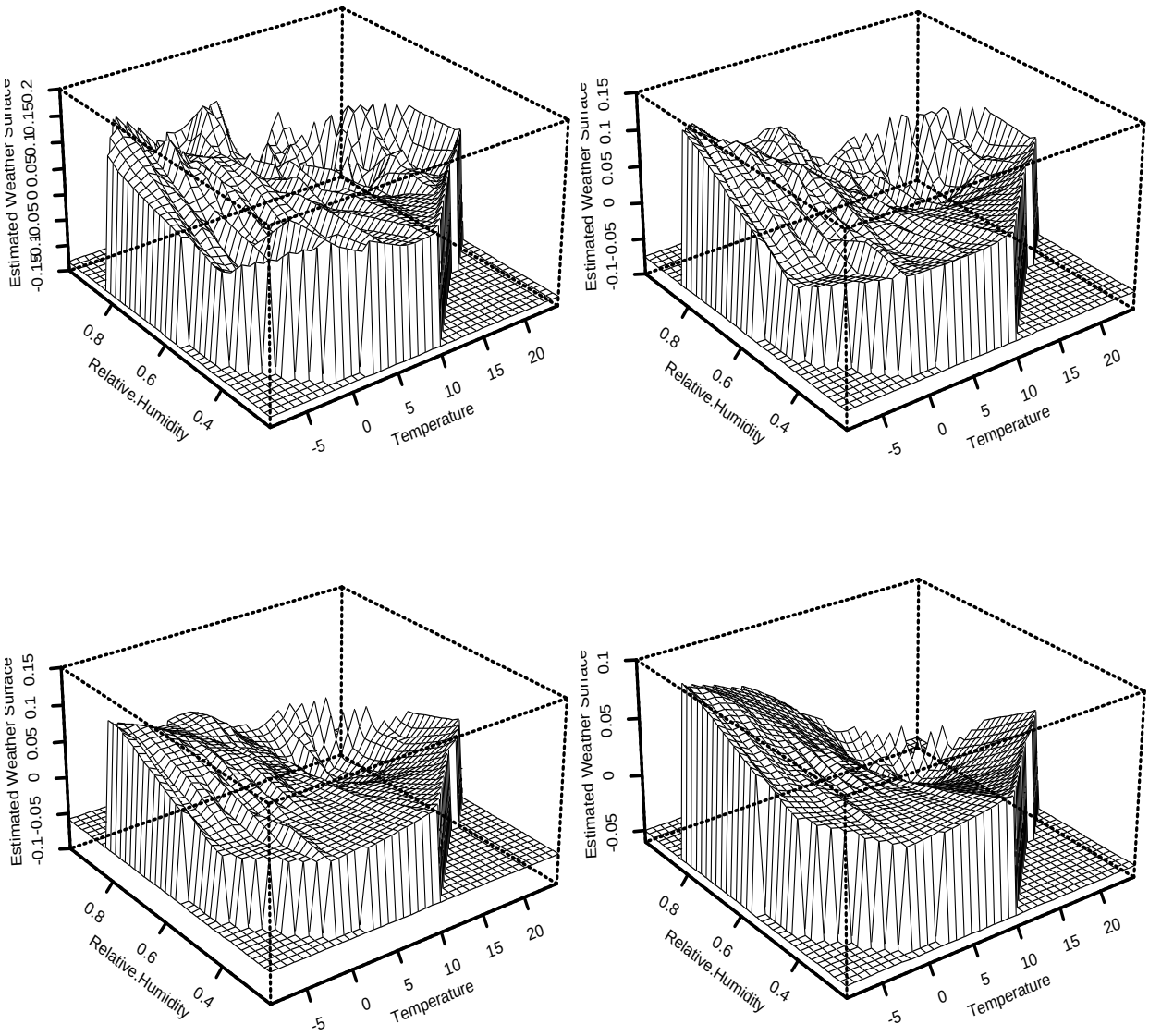

Figure 19: Plots of the estimated weather surface m_1 in model (12), describing the Vancouver data, for spans of 0.05, 0.15, 0.30 and 0.50. The span used for estimating the seasonal effect m is 0.15.

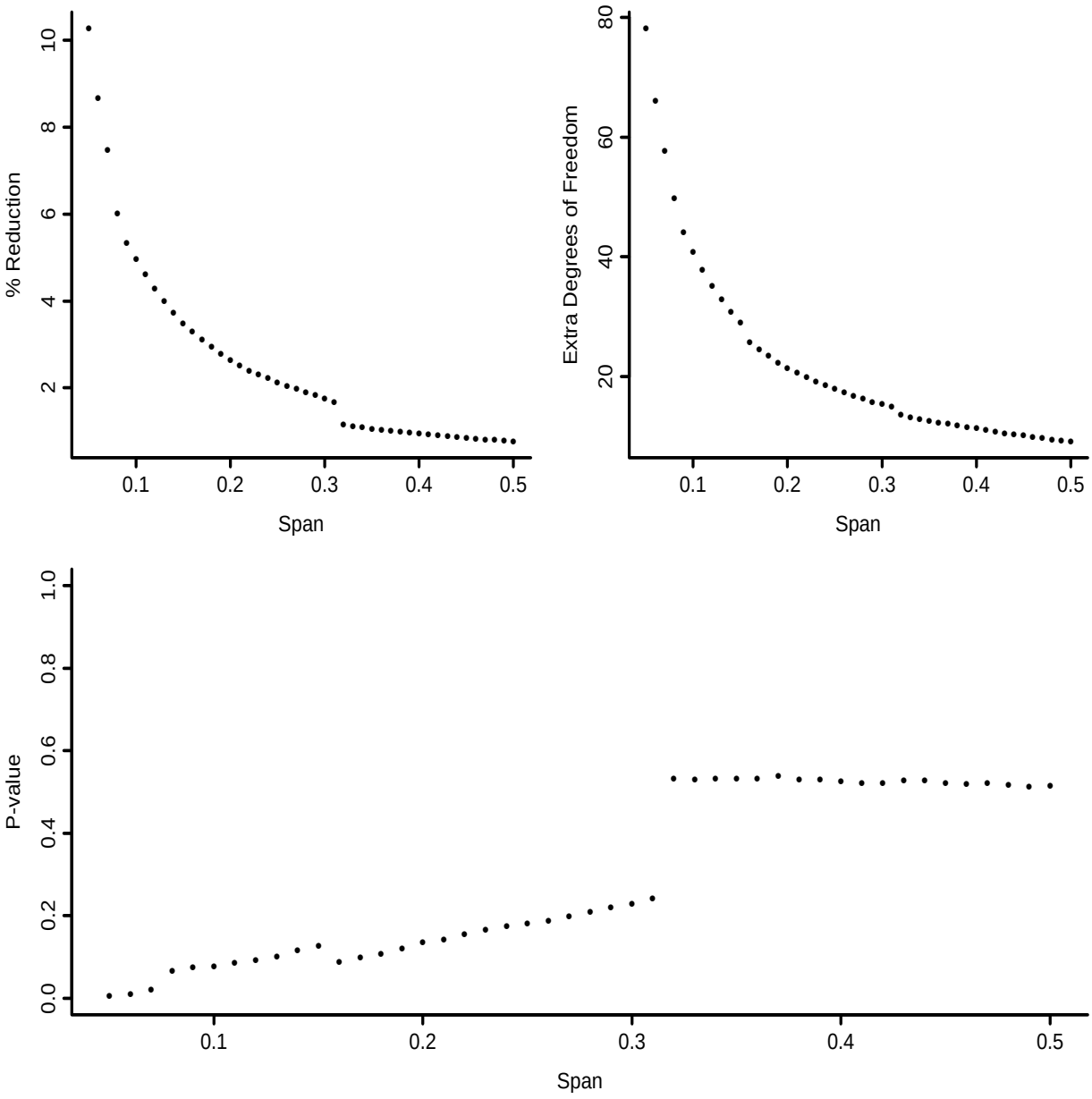

Figure 20: *Top:* Percent reduction in residual variability achieved by switching from model (11) to (12), together with the degrees of freedom expanded to achieve this reduction. Both of these models describe the Vancouver data. *Bottom:* P-values associated with a series of crude F-tests for testing model (12) against model (11). The seasonal effect m was estimated with a span of 0.15.

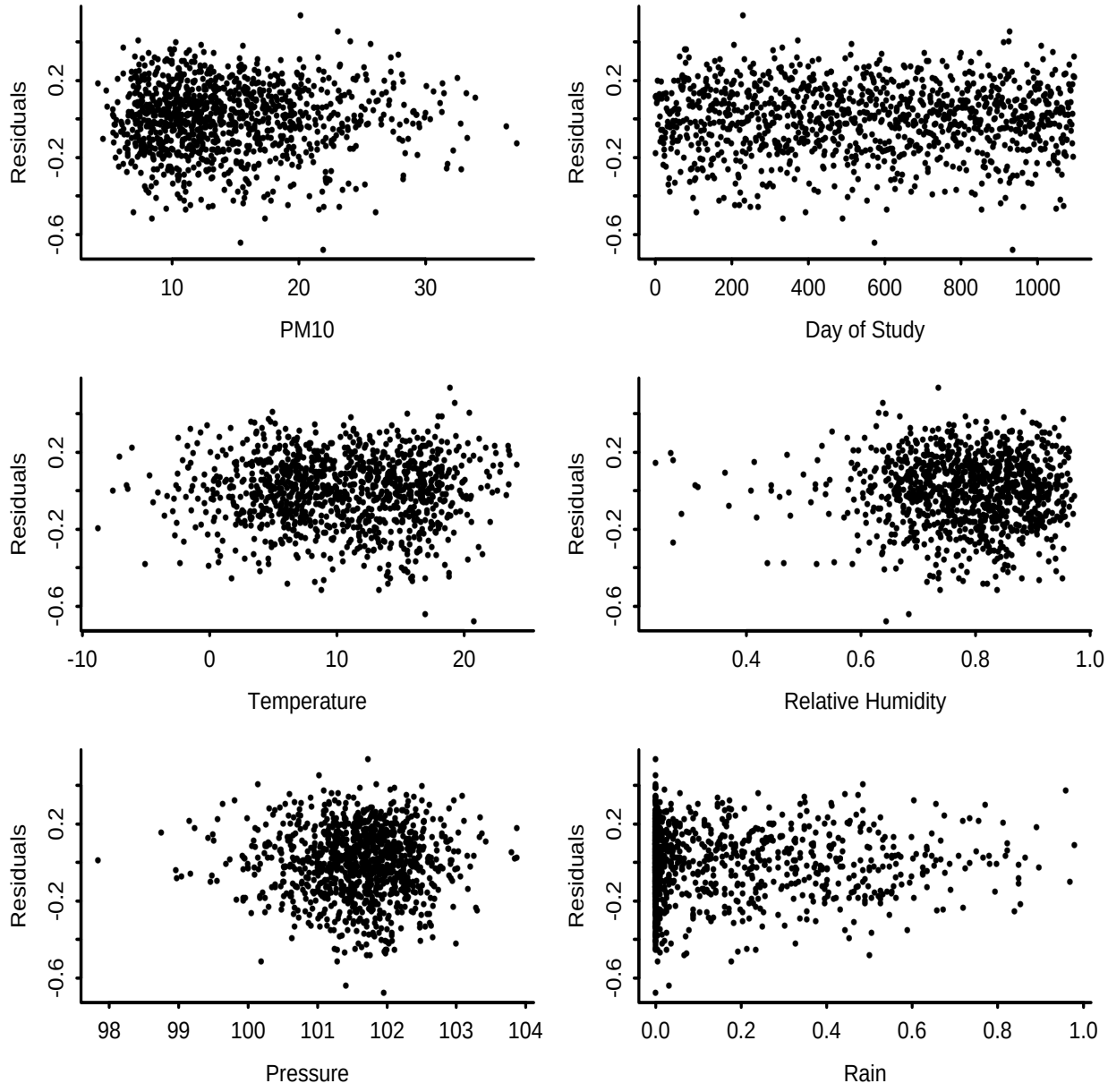

Figure 21: Plot of residuals associated with model (8), describing the Vancouver data, versus all predictors. The span used for estimating the seasonal effect m in model (8) is 0.15.

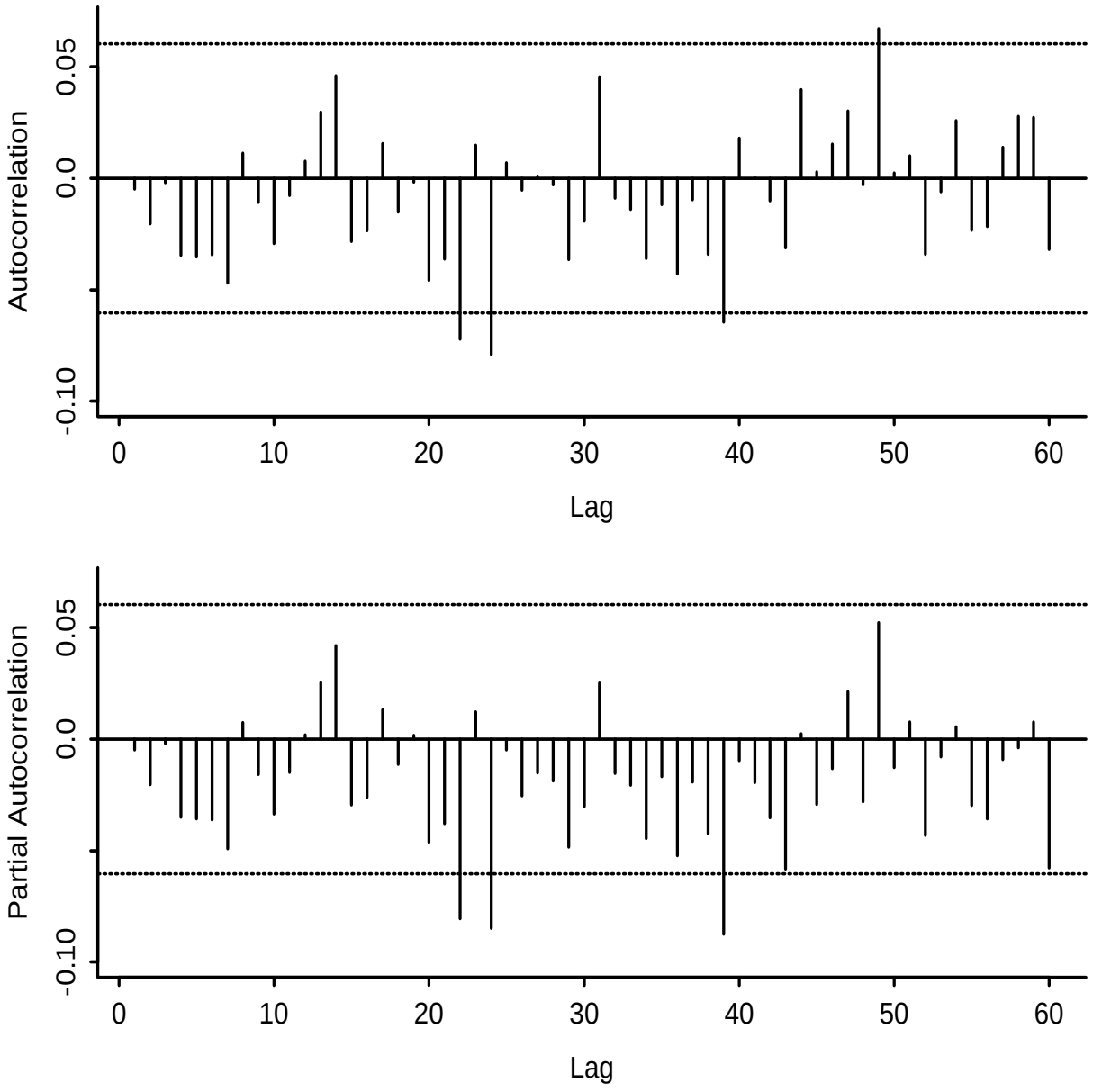

Figure 22: Autocorrelation and partial autocorrelation plots for the residuals associated with model (8), describing the Vancouver data. These residuals were computed by estimating the seasonal effect m_1 with a span of 0.15.

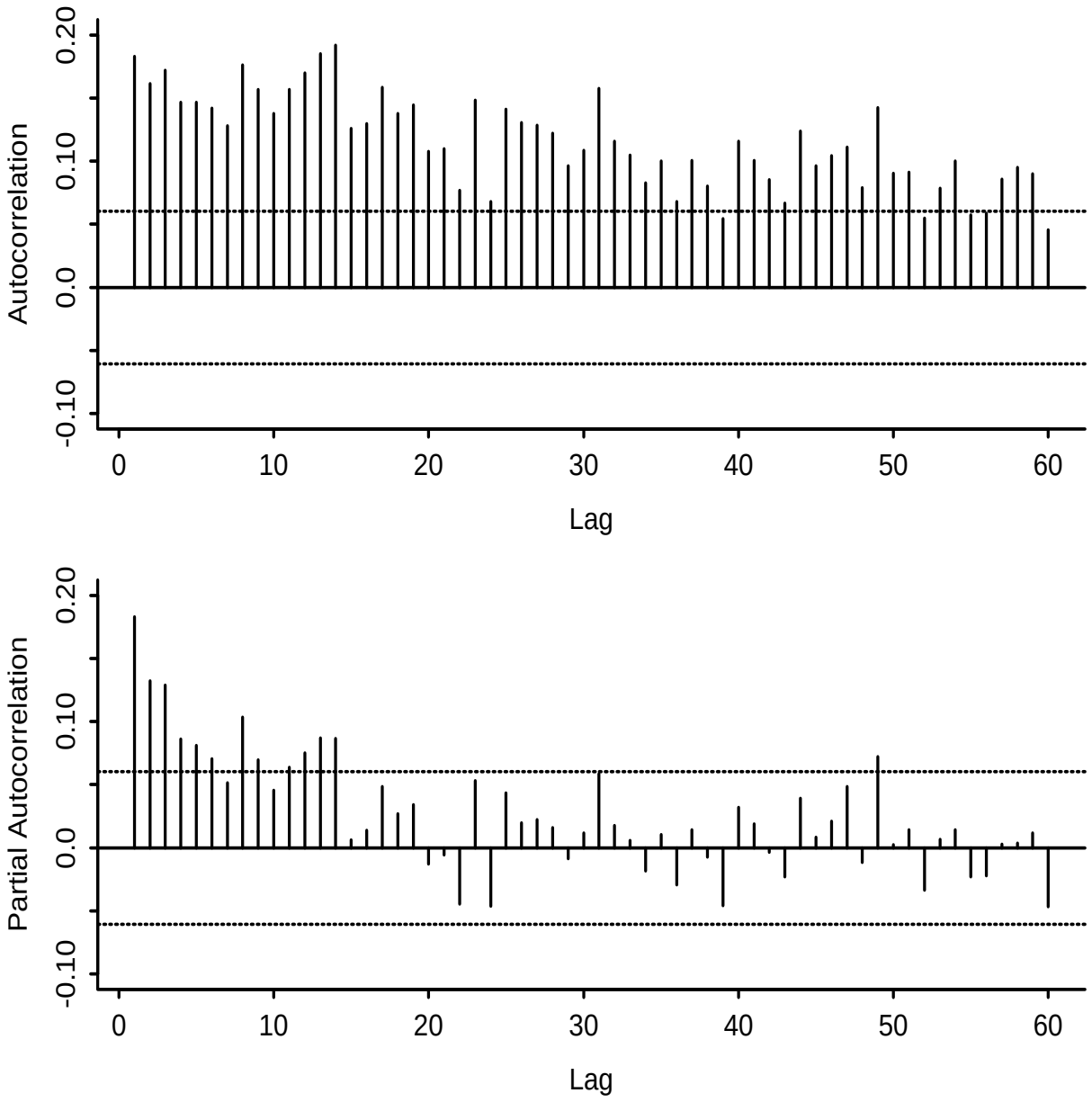

Figure 23: Autocorrelation and partial autocorrelation plots for the Vancouver log mortality counts.