Time Series Smoother For Effect Detection

Cheng You*[1], Dennis K.J. Lin†[1], and S. Stanley Young‡[2]

[1]Department of Statistics, Pennsylvania State University

[2]CGStat, LLC

October 1, 2016

Abstract

In environmental epidemiology, the dilemma that multiple time series data with inner trend cannot be fully adjusted by the observed covariates is often encountered. Inner trend is complicated to determine and separate from abnormal signals of interest. This article addresses how to de-trend time series in order to recover abnormal signals. It is found that the current spline or kernel smoothing methods can produce either positive or negative significant cross-correlations, depending on how the smoothing parameters are chosen. To circumvent this problem, three classes of robust and stable smoothers are proposed to de-trend time series data. Their properties are demonstrated by both case study, using design of experiment on scenarios, and simulation study, using indistinguishable datasets generated from the original dataset. Finally, general guidelines are provided on how to reveal abnormal signals from time series data with inner trend.

Keywords: acute effect, air quality, de-trending, design of experiment, human morality, indistinguishable simulation, robustness, stability, spline smoothing, time series smoother

*czy112@psu.edu
†DKL5@psu.edu
‡stan.young@omicsoft.com

1 Introduction

In environmental epidemiology, how the current environment change would affect human health is of great interest. In particular, do daily fluctuations in air quality induce fluctuations in human morality? In the United States, Environmental Protection Agency claims that the small particulate $PM_{2.5}$ can cause acute death and would enforce cleaner energy production by new protocols and facilities. Although their claim remains to be evaluated with some caution, the air quality problem not only affects long-term human welfare but also poses a major economical challenge to the United States and many other countries in the world.

While the air quality problem is an undergoing investigation, whether higher amount of Ozone or $PM_{2.5}$ in the air would associate with higher mortality is of important concern. The available mortality data are usually the aggregated counts at different time scales. These count data display a unique pattern with certain inner trend over time at any spacial location that one collects data from, whether it is near the equator or the poles. More importantly, the inner trend cannot be well adjusted by the observed covariates. Besides, there are other issues such as lag effect discussed in Bhaskaran et al. (2013). To further elaborate, one can do comparisons as follows. With time-dependent data, one can have the central location at each time, and the central tendency over time is called the inner trend. For example, consider a sine curve with a period of one year with signals riding on the sine curve. The inner trend is the sine curve and the actual environmental changes would be the deviations from the inner trend.

The use of nonparametric smoothing in time series models of air quality and acute death was first suggested in Schwartz (1994), where generalized additive Poisson models were used with LOESS smooths of covariates including time. The smooth function of time serves as a linear filter on the mortality and removes the inner trend in the mortality data. Several alternatives for representing the smooth functions have been applied including smoothing splines, penalized splines and parametric natural splines in Dominici et al. (2002), Ramsay, Burnett, and Krewski (2003), Schwartz, Zanobetti, and Bateson (2003) and Touloumi et al. (2004). The conclusions are drawn based on these embedded time series models and positive correlation between the air quality variable and human mortality is often found. However, this is subject to how well the de-trending works. Therefore, we intend to investigate the de-trending process systematically and provide good guidance and practice on smoothing time series with inner trends.

Human mortality excluding unnatural causes can still be attributed to various reasons, either from the latent factors or from the air quality fluctuations. It is crucial that one can extract the abnormal signals from the inner trend. However, in general, the complexity of the seasonal and long-term trends in the mortality time series or in the air quality time series is not precisely known. Oversmoothing the time series (thereby undersmoothing the deviations) can leave temporal cycles in the deviations that can produce confounding bias of inner trends; undersmoothing the time series (thereby oversmoothing the residuals) can remove too much temporal variability and potentially attenuate the acute health effect, if any. The current automatic embedded spline or kernel smoothing can be either too lenient or too harsh in de-trending the time series. By varying their smoothing flexibility, any significant result (from negative to positive) can be calibrated. This excessive flexibility could undermine the whole investigation of air quality study.

In Peng, Dominici, and Louis (2006), natural spline and penalized spline smoothing was studied as a model choice problem, and model-based simulations showed that under moderate concurvity, smoother spline of air quality variable can lead to less confounding bias. Nonetheless, different model selection criteria can still lead to different smoothness; how smooth the spline of any air quality variable is also vaguely defined. To circumvent this problem, we propose stable and robust nonparametric smoothing methods that can closely capture the inner trend and extract the abnormal signals. After smoothing or de-trending, we focus on the deviations of the raw time series subtracting its estimated inner trend. These deviations contain the most helpful information in investigating the association between air quality fluctuations and acute effects. Along the way, the general properties of these smoothers are examined to ensure the robustness and stability.

For case illustration, the research dataset is basically identical as the one investigated in Lopiano, Smith, and Young (2015). It contains human mortality, air quality and weather covariate data in Los Angeles county, California from 2000 to 2012. The daily mortality count data are obtained from California Department of Public Health. The daily air quality data $PM_{2.5}$ and Ozone are downloaded from California Environmental Protection Agency. The daily temperature data were downloaded from Carbon Dioxide Information Analysis Center of United States Historical Climatology Network and the daily humidity data were from United States Environmental Protection Agency.

Based on the whole study, there are three main points to be stated. First of all, it is found that the current spline or kernel smoothing methods can result in different (either positive or negative) effects depending on how the smoothing parameter is varied. Second, the proposed time series smoothers are robust and stable as shown by the compact correlation and partial correlation confidence interval and analyzed by factorial design on scenarios. Third, indistinguishable simulated datasets can be generated to examine their capability of detecting cross-correlations among multiple time series.

The rest of the article is organized as follows. In Section 2, the detailed problem formulation is elaborated. In Section 3, three classes of time series smoothers are proposed and their properties are summarized. In Section 4, the air quality and mortality data in Los Angeles are investigated and the factorial sensitivity analysis is conducted along with comparison to other methods. In Section 5, simulation study is conducted and it is shown that the proposed methods can detect even small effects correctly with small variance. In Section 6, the conclusion is drawn and general recommendation is made.

2 Problem Formulation

In air quality study, aggregated count data over time are usually encountered. The basic model formulation is generalized linear model with log link, as used in Schwartz (1999).

$$\begin{aligned} Y_t &\sim Poisson(\mu_t) \\ log(\mu_t) &= \alpha + \beta s_1(x_t) + \eta' \mathbf{s_2}(\mathbf{z_t}) \end{aligned} \tag{1}$$

Y_t is the mortality count variable at certain aggregation level at time t. x_t is the air quality variable of interest at time t. $\mathbf{z_t}$ is a vector of measured covariates at time t. $s_1(\cdot)$ and $\mathbf{s_2}(\cdot)$ are a spline or kernel smoothing estimator and vector of such estimators, respectively, that address the lag effect of a vector of covariates before and after the time point of interest.

However, there are two major problems. First of all, in the raw data, long-term pattern including seasonality is dominating. Due to the universal inner trend of mortality, any short-term positive association between mortality and the air component exposure of interest cannot be found. Second, Poisson regression assumes that the observations are independent, but the observations close in time are likely to be more similar than those distant in time.

To account for the inner trend, kernel or spline smoothing methods have been proposed to de-trend the original time series so the deviations can be retained for further consideration. In Peng, Dominici, and Louis (2006), the model formulation evolves as below.

$$\begin{aligned} Y_t &\sim Poisson(\mu_t) \\ log(\mu_t) &= \alpha + m_1(t, \lambda_1) + \beta x_t + \eta' \mathbf{s}(\mathbf{z_t}) \\ x_t &= m_2(t, \lambda_2) + \xi_t \end{aligned} \tag{2}$$

Here, ξ_t is the de-trended air quality variable of interest at time t. $m_1(\cdot)$, $m_2(\cdot)$ and $\mathbf{s}(\cdot)$ are embedded spline or kernel smoothing estimators over time. λ_1 and λ_2 are the smoothing parameters that control the flexibility of the spline or kernel estimator during the model fitting process. The de-trended time series observations are more likely to be independent and identically distributed. The choice of λ_1 and λ_2 depends on the degree of the inner trend that the entire model wants to remove and results in smoother or rougher splines $m_1(\cdot)$ and $m_2(\cdot)$. This arbitrariness makes the result ambiguous for evaluation.

Hence, how to de-trend and extract these abnormal signals becomes an important subject of study. We propose stable and robust estimators with systematic evaluation. Therefore, our model formulation is the following.

$$\begin{aligned} D_t^y &= y_t - rm_1(y_t^*) \\ D_t^x &= x_t - rm_2(x_t^*) \end{aligned} \tag{3.1}$$

where y_t^* and x_t^* are sets of before and after points to estimate the central tendency at time t; $rm_1(\cdot)$ and $rm_2(\cdot)$ are stable and robust estimators of the inner trend to be defined. Then, we utilize certain dependence measures or models on the de-trended time series. For instance,

$$E(D_t^y | x_t, \mathbf{z_t}) = \alpha + \beta D_t^x + \eta' \mathbf{s}(\mathbf{z_t}) \tag{3.2}$$

Here, applying logarithm transformation on mortality count is not recommended, since the logarithm transformation can shrink all the point into a narrow band and this may weaken the association. This can also be viewed as a two-stage model so that one would first de-trend the time series and then analyze using the resulting deviations. Both y_t and x_t are de-trended and thus the dominating long-term patterns including seasonality would be controlled. Meanwhile, the concurvity, analogue of multicollinearity, between x_t and $\mathbf{z_t}$ would also be controlled.

More specifically, our problem formulation in rewinding the acute effects is to simply define time series smoothers such that the inner trend can be well estimated without involving the excessive modeling and other observed covariates at the same time. The mathematical statement should simply be the first stage of model formulation. Note that in Equation 3.1, D_t^y and D_t^x are changeable due to different $rm_1(\cdot)$ and $rm_2(\cdot)$. Robustness and stability mean that no matter how the setup parameters change, the association between mortality and air quality under certain dependence measure is stable and consistent.

$$|max(\mu(D_t^y, D_t^x)) - min(\mu(D_t^y, D_t^x))| < c$$

where $\mu(\cdot)$ is a certain dependence measure and c is a relatively small constant.

The resulting deviations should not display any systematic pattern except heterogeneity during the shock of acute events. In other words, when the observations are in the non-volatile

period, $Var(D_t^y)$ and $Var(D_t^x)$ are constants. The ultimate purpose is that the de-trended time series contain little information about the inner trend but all the information of the shock events with a low level of noises. To elaborate further, the raw time series are non-stationary; after de-trending, the deviations are still non-stationary due to the abnormal signals such as forest fires or human interventions that cause the air quality changing drastically. Based on these mild assumptions, the time series decomposition can be written below.

$$X_t = S_t + \xi_t + \epsilon_t \tag{4}$$

where S_t is the long-term pattern with seasonality, ξ_t is the abnormal signal driven by an event and ϵ_t is the independently and normally distributed random error with mean 0 and variance σ^2. During the non-volatile period, ξ_t is close to 0; during the volatile period, ξ_t displays a certain pattern. The ultimate target is to closely estimate and then eliminate S_t while retaining ξ_t.

3 Proposed Methodology

To solve the formulated problem in Equation 3.1, three classes of methods utilizing a centered moving window with a centered gap of varying width are hereby proposed.

The moving window includes all the necessary information before and after the time point of interest. The gap at the center can be helpful in retaining any abnormality in the time series data. If there is a strong signal at the time point of interest, the gap reduces the local signal in the smoothed estimate and thus leaves it in the deviation; if not, this window with centered gap should be approximately the same as the ordinary moving window. This type of window includes the points on both edges to estimate the central trend at the time point of interest. The resulting estimate resembles an experimental control. After time series smoothing, the detrended observations can behave as if independently with much less serial correlation and no or little long-term including seasonality patterns left.

To elaborate further, the formulation in Equation 3.1 and 4 is utilized for justification. For simplicity, one point before and after the time point of interest is considered and the smoothing measure is the mean.

$$\begin{aligned} D_t^X &= X_t - rm(X_t^*) \\ &= X_t - \frac{X_{t-1} + X_{t+1}}{2} \\ &= S_t + \xi_t + \epsilon_t - \frac{S_{t-1} + S_{t+1}}{2} - \frac{\xi_{t-1} + \xi_{t+1}}{2} - \frac{\epsilon_{t-1} + \epsilon_{t+1}}{2} \\ &= (S_t - \frac{S_{t-1} + S_{t+1}}{2}) + (\xi_t - \frac{\xi_{t-1} + \xi_{t+1}}{2}) + (\epsilon_t - \frac{\epsilon_{t-1} + \epsilon_{t+1}}{2}) \end{aligned}$$

For the first term, $\frac{S_{t-1}+S_{t+1}}{2}$ is a linear interpolation of S_t, instead of extrapolation. The long term pattern including seasonality S_t is assumed to be smooth enough. Thus, $\frac{S_{t-1}+S_{t+1}}{2}$ is close enough to S_t. The first term is close enough to 0. For the second term, there are two cases: if t is the shock time, ξ_t should display a sharp spike while $\frac{\xi_{t-1}+\xi_{t+1}}{2}$ is at regular level and therefore the difference exposes the shock signal; if t is the regular time, ξ_t and $\frac{\xi_{t-1}+\xi_{t+1}}{2}$ are both at regular level and therefore the difference is close to 0. For the third term, since ϵ_t is i.i.d normal with mean 0 and variance σ^2, $\epsilon_t - \frac{\epsilon_{t-1}+\epsilon_{t+1}}{2}$ should be close to 0 too in most cases. Hence, the resulting term is only the shock signal.

In general, the formulation is as follows.

$$\begin{aligned} D_t^X &= X_t - rm(X_t^*) \\ &= S_t + \xi_t + \epsilon_t - rm(S_t^*) - rm(\xi_t^*) - rm(\epsilon_t^*) \\ &= (S_t - rm(S_t^*)) + (\xi_t - rm(\xi_t^*)) + (\epsilon_t - rm(\epsilon_t^*)) \end{aligned}$$

where $S_t - rm(S_t^*) \to 0$, $\xi_t - rm(\xi_t^*) \to \xi_t$ and $\epsilon_t - rm(\epsilon_t^*) \to 0$. D_t^X is an unbiased and consistent estimator to ξ_t.

In this section, our major contribution is to define the robust and stable time series smoothers. These smoothers can de-trend the time series efficiently, regardless of the window size or gap size. The rest of this section is organized as follows. Three classes of time series smoothers are introduced consequentially and each has their own merits. The first class is named moving trimmed mean, where the idea of trimmed mean is used within each moving window. It can change the robustness, from mean to median, controlled by the trimming percentage upon the user's preference or certain selection criterion. The second class is named moving weighted mean, where the weight scheme can be determined upon the user's preference or certain selection criterion. It can incorporate the user's prior knowledge in trend estimation. The third class is named moving recursive weighted mean. It can re-weight the points within each moving window recursively in order to make abnormal signals revealed. The re-weighting function is non-increasing and specified by the user, based on their prior knowledge and preference on the subject of study. The performance of the three classes of smoothers are demonstrated by both case study and simulation study in Section 4 and 5.

3.1 Moving Trimmed Mean

The first class of methods is moving trimmed mean or truncated mean. A trimmed mean or truncated mean is a statistical measure of central tendency which involves the calculation of the mean after discarding given parts of a probability distribution or sample at the high and low end, and typically discarding an equal amount of both. Trimmed mean is less sensitive to outliers and gives reasonable estimate of central tendency. Therefore, it is regarded as a robust estimator.

The trim parameter, ranging from 0% to 50%, allows us to obtain a large collection of

estimators. When the trim parameter is 0%, we trim no points and the estimator is arithmetic mean; when the trim parameter is 50%, we trim 50% of high and low points and it results in the most robust estimator, median.

Moving average is defined as the average of several days' observations with a centered gap.

$$MA(x_t^*) = \frac{1}{2(k-l)}(x_{t-k} + x_{t-k+1} + \cdots + x_{t-l-1} + x_{t+l+1} + \cdots + x_{t+k-1} + x_{t+k})$$

Moving median is defined as the median of several days' observations with a centered gap.

$$MM(x_t^*) = \frac{1}{2}(X_{(k-l)} + X_{(k-l+1)})$$

where $X_{(k-l)}$ and $X_{(k-l+1)}$ are the $(k-l)$th and $(k-l+1)$th ordered statistics of the set $\{x_{t-k}, x_{t-k+1}, \cdots, x_{t-l-1}, x_{t+l+1}, \cdots, x_{t+k-1}, x_{t+k}\}$; t is the time of our estimation; k is the number of observations we use before and after the estimation time point ($k \geq 1$) hence the total number of observations is $2k+1$ before gapping; l is the number of observation we take away from the center before and after the estimation time point ($l \geq 0$) excluding x_t hence the total number of observations taken away is $2l+1$ and the total number of observations left is $2(k-l)$.

3.2 Moving Weighted Mean

The second class of methods is moving weighted mean. The weighted mean assigns different weights to the data points within each window in order to form an informed estimator. In general, it is upon the user to determine how to assign the weights. Two typical types of moving weighted mean are presented: center weighted moving average and edge weighted moving average. Center weighted moving average puts more weights around the center outside the gap than the edges symmetrically. It weighs more on the local trend near the time point of interest. Edge weighted moving average puts more weights on the edges than the center symmetrically. It weighs less on the specific local trend but more on the pattern of a longer term.

The sum of the normalized weight vector $(w_{t-k}, w_{t-k+1}, \cdots, w_{t-l-1}, w_{t-l+1}, \cdots, w_{t+k-1}, w_{t+k})$ is equal to 1. Dividing the weight vector by its sum is called normalizing the weights.

Center weighted moving average is defined as the center weighted average of several days'

observations with a centered gap. The weight vector contains $\wedge$-shaped linear weights and is normalized to 1, with more weight at the center and less weight at the edges symmetric to time t.

Edge weighted moving average is defined as the edge weighted average of several days' observations with a centered gap. The weight vector contains $\vee$-shaped linear weights and is normalized to 1, with less weight at the center and more weight at the edges symmetric to time t.

3.3 Moving Recursive Weighted Mean

The third class of methods is moving recursive weighted mean. Recursiveness is that after we obtain the deviations, we assign more weight on small deviation but less weight on large deviation to calculate weighted mean and then perform this procedure repeatedly until convergence. We would like those abnormal signals to stand out while those around the central tendency are almost zero. The detailed algorithm is as follows.

Step 1: Apply moving average to obtain the estimated trend of time series.

Step 2: Calculate deviations from the estimated trend and standardize them.

Step 3: Apply the user-defined weight function on standardized deviations, for example e^{-x^2}, normalize them and obtain moving weighted mean trend.

Step 4: Repeat Step 2 and 3 until the estimated trend converges with the difference of the estimated inner trends $||\cdot||_\infty < 10^{-6}$.

The convergence of trend estimates is very fast. Typically, it takes less than 10 iterations for each window.

There are infinitely many choices of recursive weight function. In the following section, two recursive weight functions $e^{-|x|^{1/2}}$ and $e^{-|x|^2}$ are used for demonstration, where $e^{-|x|^{1/2}}$ has less penalized weighting on deviations while $e^{-|x|^2}$ has more penalized weighting on deviations.

3.4 General Properties

In general, the moving measures should incorporate two basic considerations. If the period of seasonality is known, then the window width should be its fraction. The literature in air quality

study typically picks the window width from a few weeks up to a month. If certain shocks are known, then the gap width should be approximately the shock width. Shocks are usually some natural phenomena, for example, wildfires, that immediately raise the level of Ozone or $PM_{2.5}$; or human intervention such as a regulatory action, for example, declaring non-attainment and attainment regions in Chay and Greenstone (2003).

For the three classes of time series smoothers, they have shared and individual properties. The shared properties come from the window and gap. When the window size increases, the trend estimated is stiffer and the deviation will contain more seasonal signals; when the gap size increases, the trend is also stiffer and the deviation will contain more seasonal signals. The individual properties come from the different measures of central tendency. Moving trimmed mean is robust and stable as well as bearing a good scientific meaning. When the trimming percentage increases, the estimated trend is more robust. Moving weighted mean is more flexible, depending on the weighting scheme. It can accommodate the users' preference and prior knowledge. Moving recursive weighted mean estimates a relatively robust central tendency and makes the abnormal deviations more noticeable and the regular deviations near zero. In this sense, it is more helpful in recovering abnormal signals.

More specifically, the exemplified smoothers also have their own properties. Moving average is less robust than moving median but more efficient in the trend removal. Within the moving window, if the fluctuations around the trend are symmetric, moving average can easily recover the underlying trend of the time series. If the fluctuations around the trend are normally distributed, then moving average is statistically optimal. Moving median is the most robust moving measure that captures the overall trend of the time series. Within the moving window, no matter whether the fluctuations around the trend are symmetric or not, the median can always estimate the trend central location. If the fluctuations are Laplace distributed, then moving median is statistically optimal.

Center weighted moving average is used when one wants more weight on the observations near the time point of interest to obtain a better local or short-term trend estimate. It weighs the local observations more in forming an experimental control thus more aggressive in the short-term trend removal. Edge weighted moving average is used when one wants more weight on the observations further away from the time point of interest to obtain a better long-term estimate. It weighs the local observations less in forming an experimental control thus less aggressive in

the short-term trend removal.

Moving recursive weighted mean with $e^{-|x|^{1/2}}$ places more weight on the ordinary observations and less weight on the abnormal observations. This weight function penalizes less on large deviations and the resulting smoother follows the curve fluctuation closer. Hence, the deviations will moderately stand out. Moving recursive weighted mean with $e^{-|x|^2}$ has weighted function with much heavier weight on the ordinary observations and much lighter weight on the abnormal observations. This weight function penalizes more on large deviations and the resulting smoother follows the central trend closer. Hence, the deviations will be more noticeable.

4 Case Study

4.1 Data and Description

To explicate the proposed methods, their performances on air quality and mortality data are illustrated. The research dataset comes from Los Angeles, California. The dataset contains daily mortality counts, air quality levels for ozone and $PM_{2.5}$, minimum and maximum temperature and maximum relative humidity from Year 2000 to 2012 with 4749 data entries in total. Below is a snapshot of the dataset.

RowID	basin	year	month	day	dayofyear	AllCause75	PM25davg	o3	tmin.0	tmax.0	MAXRH.0
1	south-coast	2000	1	1	1	226	68.6	40	37	62	94.3
2	south-coast	2000	1	2	2	235	7.5	40	19	65	71.0
3	south-coast	2000	1	3	3	202	NA	39	35	69	53.6
4	south-coast	2000	1	4	4	223	27.1	38	37	72	49.8
5	south-coast	2000	1	5	5	217	30.2	41	38	72	58.0
6	south-coast	2000	1	6	6	216	24.3	40	38	74	41.1
7	south-coast	2000	1	7	7	248	35.8	36	37	69	54.3
8	south-coast	2000	1	8	8	196	49.6	39	35	71	66.3
9	south-coast	2000	1	9	9	185	65.5	34	36	69	75.4
10	south-coast	2000	1	10	10	185	73.4	40	35	69	98.0
11	south-coast	2000	1	11	11	199	59.5	28	38	70	83.1
12	south-coast	2000	1	12	12	190	61.6	33	39	68	89.3
13	south-coast	2000	1	13	13	165	53.4	37	40	79	69.3
14	south-coast	2000	1	14	14	178	15.0	39	42	82	44.2
15	south-coast	2000	1	15	15	165	30.6	42	44	83	58.0
16	south-coast	2000	1	16	16	196	38.8	36	43	75	81.7
17	south-coast	2000	1	17	17	149	49.2	32	45	72	88.8
18	south-coast	2000	1	18	18	177	17.9	28	55	78	80.8
19	south-coast	2000	1	19	19	163	18.9	31	50	78	85.3
20	south-coast	2000	1	20	20	182	18.3	23	52	72	90.5
21	south-coast	2000	1	21	21	160	20.3	31	52	65	90.1
22	south-coast	2000	1	22	22	154	36.2	34	46	65	86.5

Figure 1: Sample Data from Los Angeles California

$AllCause75$ is the daily mortality counts above Age 75, excluding unnatural deaths such as accidents. $o3$ is the daily average level of ozone. $tmin.0$ is the daily minimum temperature. $tmax.0$ is the daily maximum temperature. $MAXRH.0$ is the daily maximum relative humidity level.

In the demonstrative analysis, the association between daily ozone level and mortality is our focus. Due to the large number of levels of factors, a factorial design on different scenarios is formed and linear model is be applied on. By the factorial design and sensitivity analysis, more features of three smoothers could be understood better, along with the research question about air quality and acute death.

For the response variable, the correlation and partial correlation as well as their p-values are utilized as the dependence measures. Partial correlation controls for minimum temperature, maximum temperature and maximum relative humidity level.

4.2 Current Smoothing Methods

In this section, it is found that the current smoothing methods such as spline or kernel smoothing can produce significant correlations, from negative to positive. For demonstration, the penalized cubic smoothing splines are utilized. Cubic smoothing splines performs a regularized regression over the natural spline basis, placing knots at all points. They circumvent the problem of knot selection, as they just use the inputs as knots, and simultaneously, they control for over-fitting by shrinking the coefficients of the estimated function. It can be shown that these smoothing methods can be too flexible to be relied on.

In Figure 2, the smoothing spline is allowed to vary from a straight line to a extremely flexible curve by varying the smoothing parameter. The smoothing parameter is controlled by the option *spar* of *smooth.spline*() in R. The upper bound $spar = 1.5$ represents the stiffest while the lower bound $spar = -1.5$ represents the most flexible. For simplicity, the smoothing parameters are set the same for both $AllCause75$ and $O3$.

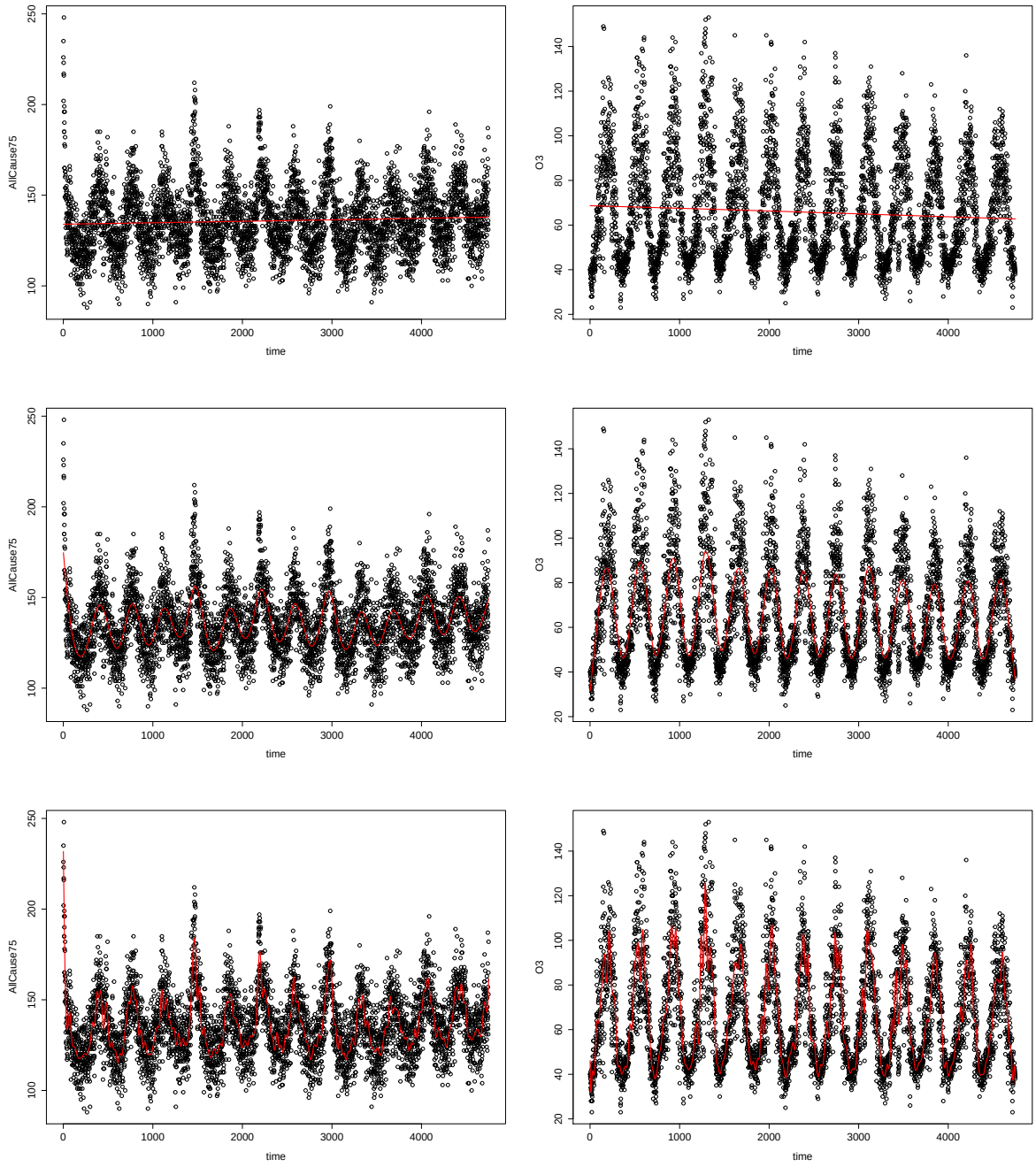

Figure 2: Raw Time Series and Smoothing Splines of Mortality and Ozone

The de-trended time series or deviations are shown correspondingly in Figure 3. The more de-trending, the more similar to *i.i.d.* the deviations.

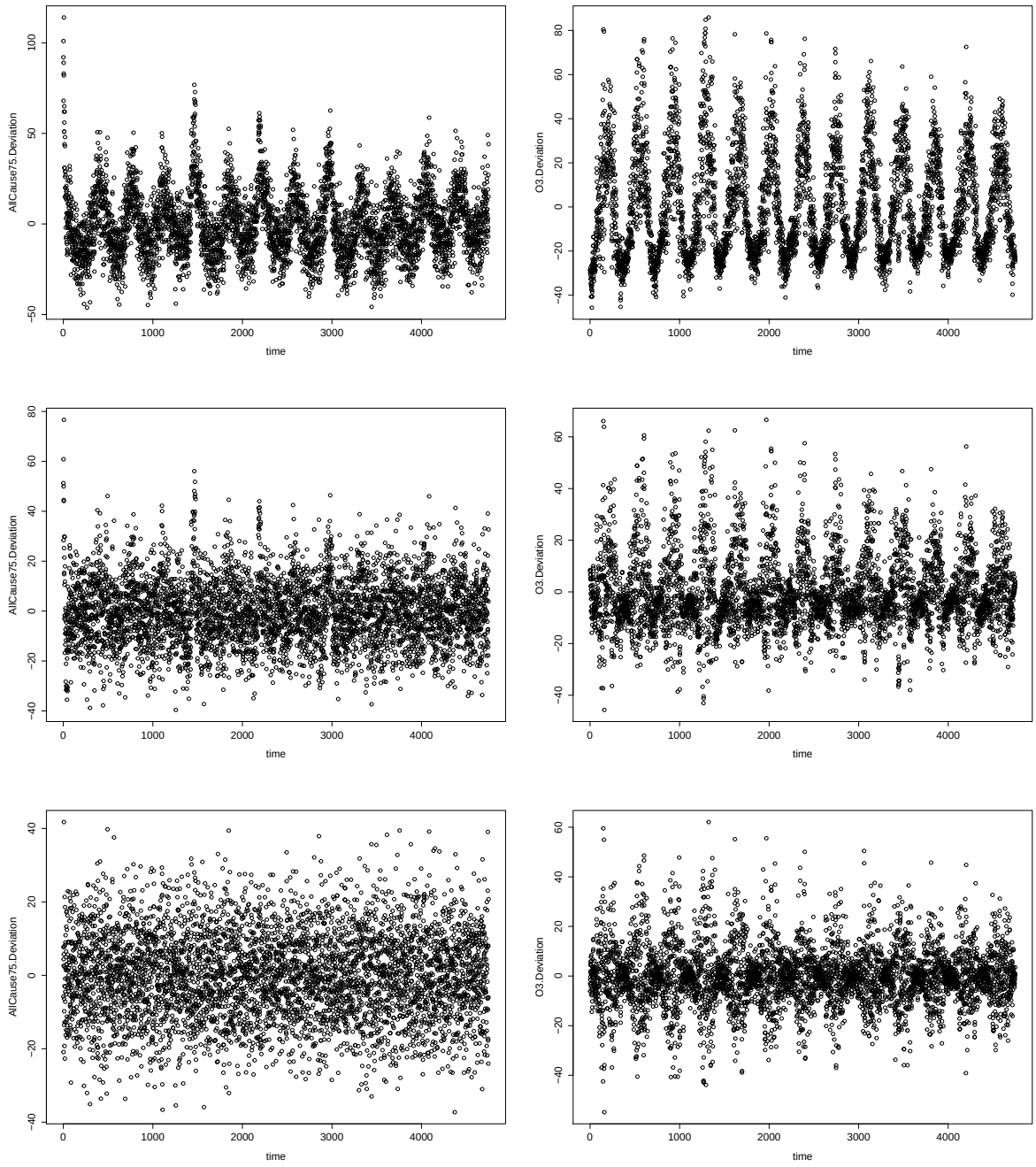

Figure 3: De-trended Time Series of Mortality and Ozone

Spar	Correlation	P-value	Partial Corrrelation	P-value
1.50	-0.4313	≈ 0	-0.08354	≈ 0
0.71	-0.2665	≈ 0	-0.001043	0.9427
0.70	-0.2429	≈ 0	0.008141	0.575
0.61	-0.005198	0.7203	0.06978	≈ 0
0.60	0.01395	0.3367	0.07174	≈ 0
-1.50	0.07248	≈ 0	0.04627	≈ 0

Table 1: Correlations and Partial Correlations with P-values between Mortality and Ozone

In Table 1, Spar is the smoothing parameter *spar* in *smooth.spline*() and its values are

chosen based on the sign changes of the resulting correlations. Correlation means the correlation between deviation of mortality and deviation of ozone while partial correlation means the partial correlation between them given the weather covariates temperature and relative humidity. As the inner trend including seasonality gets eliminated more and more, both the correlation and partial correlations between mortality and ozone changes from significantly negative, to insignificant and then to significantly positive. Indeed, one has to find a proper way to justify de-trending or at least find a stable and robust method before making a claim on any possible association.

4.3 Proposed Smoothing Methods

In order to study the robustness and stability of our proposed smoothers, a factorial design of experiment on different settings is utilized for the time series smoothers. Table 2 is a brief summary of the factorial design.

Factor	Level
Window	7, 15, 21, 29, 35, 43, 49 or 57
Gap	0, 1, 3, 5, 7, 9, 11 or 13
Scenario	Two Levels
Type	DD, DR, RD, RR

Table 2: Factors for Design of Experiments on Scenarios

For the moving window size, the variable Window is defined and assumes 7, 15, 21, 29, 35, 43, 49 or 57 days. The actual meaning would be one week, two weeks and up to one month before and after the time point of interest.

For the gap size, the variable Gap is defined and assumes 0, 1, 3, 5, 7, 9, 11 or 13 days. The actual meaning would be no removal, one-day point removal and up to one-week points removal before and after the time point of interest. Note that when the gap size is 0, the ordinary moving measures is simply used. Also, the gap should always be smaller than the window.

For each class of time series smoothers, two representative scenarios are included as described before to understand the effect of the pre-selected parameter or weighting scheme. There are infinite choices of the parameters or weighting schemes within the window. This is upon the user's preference.

To simplify the presentation, the factor Type is defined and plan to investigate whether there is any effect between deviations of mortality and deviations of ozone by assuming DD, between

deviations of mortality and raw time series of ozone by assuming DR, between raw time series of mortality and deviations of ozone by assuming RD and between raw time series of mortality on raw time series of ozone by assuming RR. Since no smoother is applied on RR, it is a special case in this design.

All scenarios combined resembles a full factorial design of experiments for us to examine all possible effects. This factorial design is of the size $8 \times 8 \times 2 \times 4 - \#\text{ missing} = 325$.

For the factorial design analysis, linear regression model is utilized and all main effects and two-way interactions are included. To summarize the effect and significance, volcano plot is utilized.

For illustration, moving trimmed mean is shown below. Moving weighted mean and moving recursive weighted mean can be used in the same way and produce similar results. Here, moving average with trimming percentage = 0% and moving median with trimming percentage = 50% are selected.

In Figure 4, it is shown that only the effects with Type have both large size and significance; all other effects have small size although highly significant.

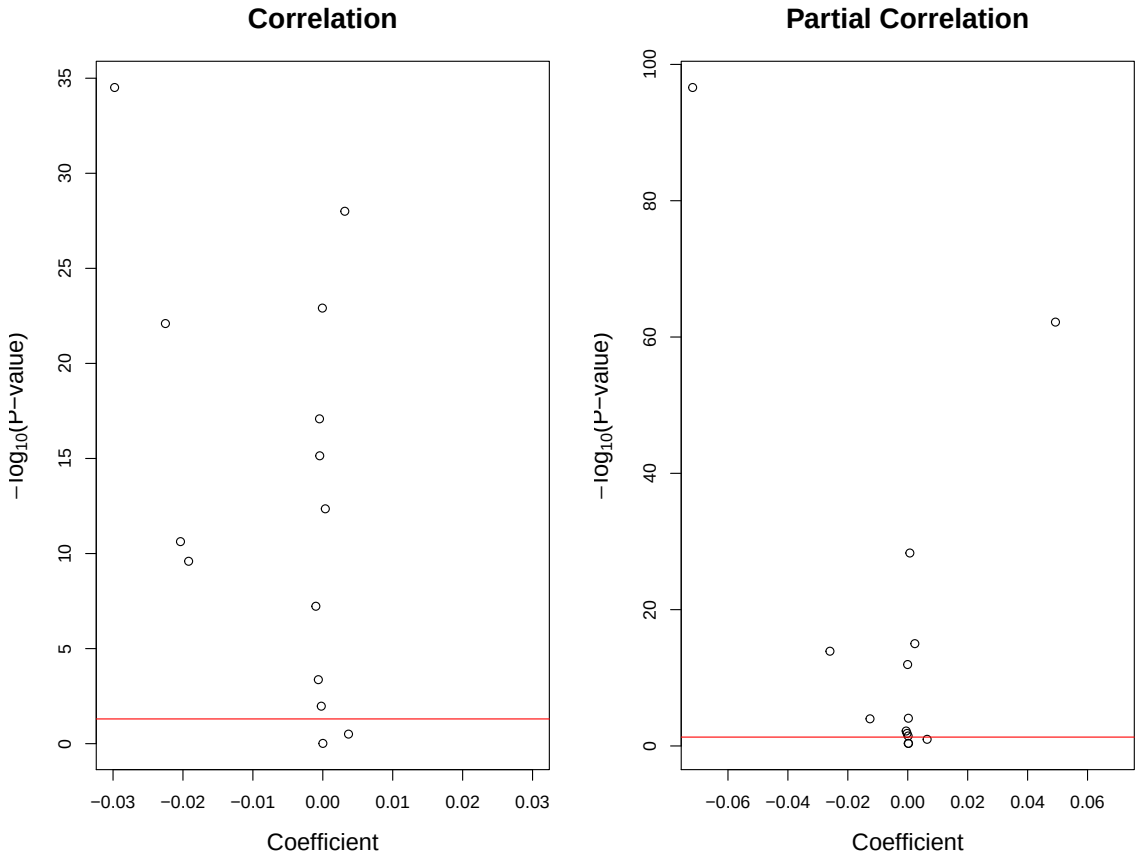

Figure 4: Volcano Plot by Moving Trimmed Mean;
Red Line = $-log_{10}0.05$

In more detail, for correlation, all main effects and two-way interaction effects are significant except the main effect Lambda and the interaction effect Gap×Lambda. However, the increments of Window or Gap are so small when Type is fixed that they would not alter the sign of the correlation. On the other hand, Type can make a significant difference on the association between mortality and ozone. The trimming percentage Lambda is insignificant. For the response measure partial correlation, we can obtain similar results. The detailed tables of coefficient estimates and significance are in Appendix A. Thus, moving trimmed mean is robust and stable with respect to different Window, Gap and Lambda.

In Figure 5, correlation and partial correlation across different types are summarized in boxplots.

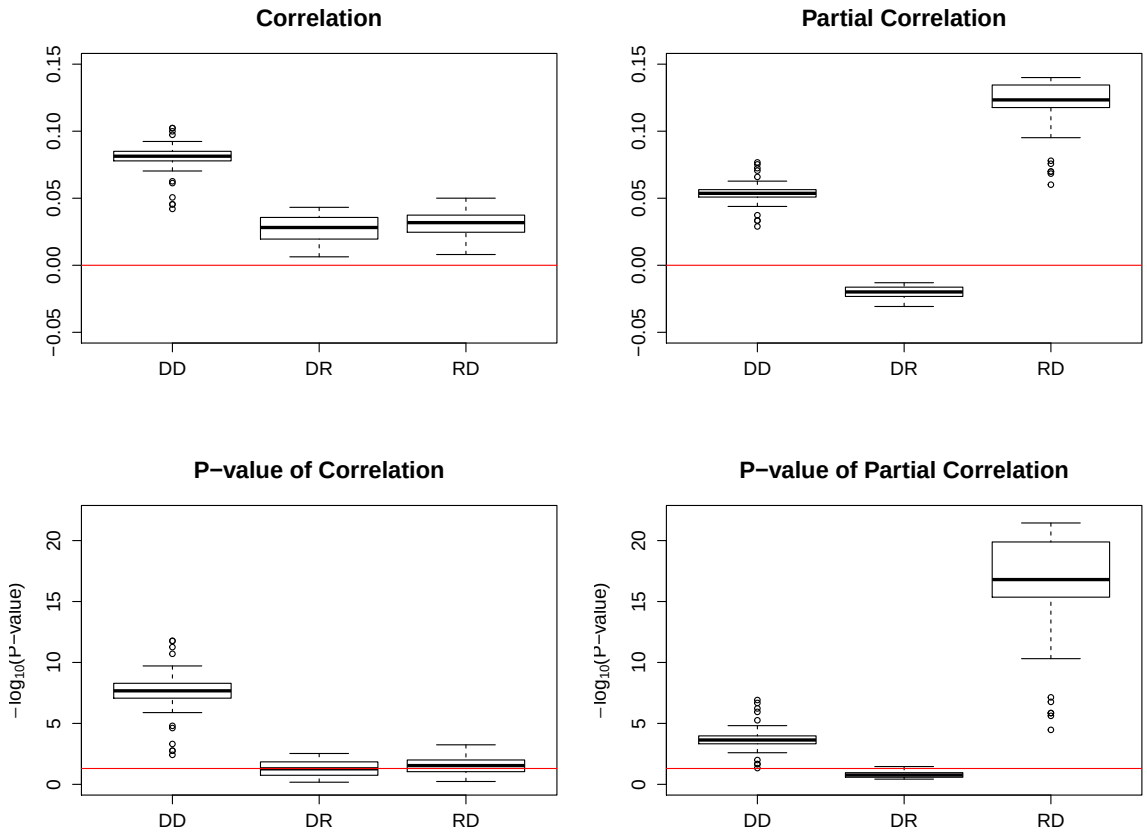

Figure 5: Boxplots of Association by Moving Trimmed Mean;
Red Line (above) = 0, Red Line (below) = $-log_{10}0.05$

Table 3 shows the correlations and partial correlations with their p-values for each type when Window=21, Gap=5 and Trimming Percentage=50%.

Type	Correlation	P-value	Partial Correlation	P-value
DD	0.08693	≈ 0	0.05683	≈ 0
DR	0.03314	0.02266	-0.02832	0.05148
RD	0.03364	0.02071	0.1140	≈ 0
RR	-0.4338	≈ 0	-0.0956	≈ 0

Table 3: Association for Type by Moving Trimmed Mean:
Window=21, Gap=5, Trimming Percentage=50%

Based on Figure 5 and Table 3, it can be seen that the correlations and partial correlations when Type=DD are positively significant. When Type=DR, half of correlations are insignificant while all the partial correlations are insignificant. When Type=RD, half of correlations are significant while all the partial correlations are highly positively significant. De-trending on both time series is greedier in finding significance.

From the factorial design of analysis, we conclude that Window and Gap can adjust the

association slightly, however, the sign of association will not flip when Type is fixed. For moving trimmed mean, the small effect of Window and Gap shows the robustness and the no-effect of Lambda also shows the stability.

4.4 Discussion & Recommendation

On one hand, it has been discovered that the current spline or kernel smoothing can produce any significant correlations, from negative to positive, depending on how the smoothing parameters are varied. On the other hand, robust and stable smoothers are proposed to circumvent the sensitiveness of smoothing parameter and evaluated on Los Angeles dataset.

The main framework of the proposed methods is to utilize a moving window with centered gap. Within each window, different central tendency measures can be selected, based on the user's preference. As the window size increases, the estimated curve is stiffer and de-trending is less severe; stronger correlation will be found between deviations. As the gap size increases, more local signal is removed, the estimated curve is also stiffer and de-trending is less severe; stronger correlation will be found between deviations. However, no matter how the window or gap size changes, within a wide range of days, the sign or significance of the association for each type remains the same. This is one important property, robustness.

Within each moving window with centered gap, three classes of time series smoothers are proposed. For each moving measure, varying the parameter or weight can adjust the correlation or partial correlation. However, the sign or significance would not change. For moving recursive weighted mean, even though the sign of correlation changes when Type=DR or RD, the correlations are around zero and mostly insignificant thus the results are still consistent. This is another important property, stability.

For moving trimmed mean, it involves a collection of classic central tendency measures with the robustness changing from mean to median. Within each window, different trimmed means can reach statistical optimality under different underlying distributions. If one does not have any prior knowledge or preference on the abnormal signals of time series data, moving trimmed mean can work.

For moving weighted mean, it involves weighted arithmetic mean. Within each window, one can decide assign different weights to the points based on their prior knowledge or preference on

the abnormal signals of time series data. If one has any prior knowledge or preference, moving weighted mean can help.

For moving recursive weighted mean, it involves re-weighting scheme in order to obtain a better trend estimate in terms of retaining central tendency conservatively and enlarging abnormal signals. It is more greedy in searching for anomaly. If one emphasizes on conserving the central tendency and revealing more anomaly, moving recursive weighted mean can handle.

5 Simulation Study

5.1 Simulation Setting

Due to the robustness of the smoothers, Window=21, meaning the most useful information are within 10 days before and after the time point of interest, and Gap=5, meaning the removal of local trend are within 2 days before and after, are chosen in the simulation.

The target correlation between deviations of mortality and ozone are set as 0, 0.01, 0.02, 0.05 and 0.10. In Section 4, the detected correlation is between 0.05 and 0.10. In literatures, it is understood that the correlation or partial correlation between air quality and mortality is relatively low. Hence, it is reasonable to set the correlation to be 0.05 and 0.10. Also, including 0.01 and 0.02 can demonstrate how sensitive the proposed smoothers are in revealing abnormal signals.

When the correlation between deviations of mortality and ozone is pre-specified, leaving the other covariate structures unchanged, the partial correlation will also be pre-specified and can be obtained from the whole correlation matrix. By R package *corpcor*, the corresponding partial correlations are computed as -0.0290, -0.0182, -0.0074, 0.0250 and 0.0791.

Therefore, the simulation times for each scenario is 100 and the total simulation times for each smoother is 1000.

In this simulation study, the idea is to generate many indistinguishable datasets from the original dataset. Indistinguishable means that the datasets are almost identical under the measures correlation and partial correlation. To maintain the original correlation structure, each time series is decomposed by LOESS into trend, seasonal and reminder terms. Only the deviations are simulated by multivariate normal distribution with the actual means and variances of the decomposed deviations, and pre-defined correlations thus covariances. The detailed algorithm is as follows.

Step 1: Decompose each time series into trend, seasonal and reminder components.

Step 2: Retain the actual means and variances of each reminder. Set the correlation between reminders of mortality and ozone to be the pre-specified value above and obtain the new covariance matrix of all the variables.

Step 3: Simulate the deviations by multivariate normal distribution with the actual means and

new covariance matrix.

Step 4: Add the simulated deviations back to the decomposed trend and seasonal components so the simulated data are formed.

5.2 Simulation Comparison

Current Smoothing Methods

Given the cubic spline smoothing, generalized cross validation can be adopted to select the tuning parameter. The leave-one-out cross validation gives almost the same result.

In Figure 6, it can be seen that the current smoothing can identify the correct correlation and partial correlation with small errors, under multivariate normality.

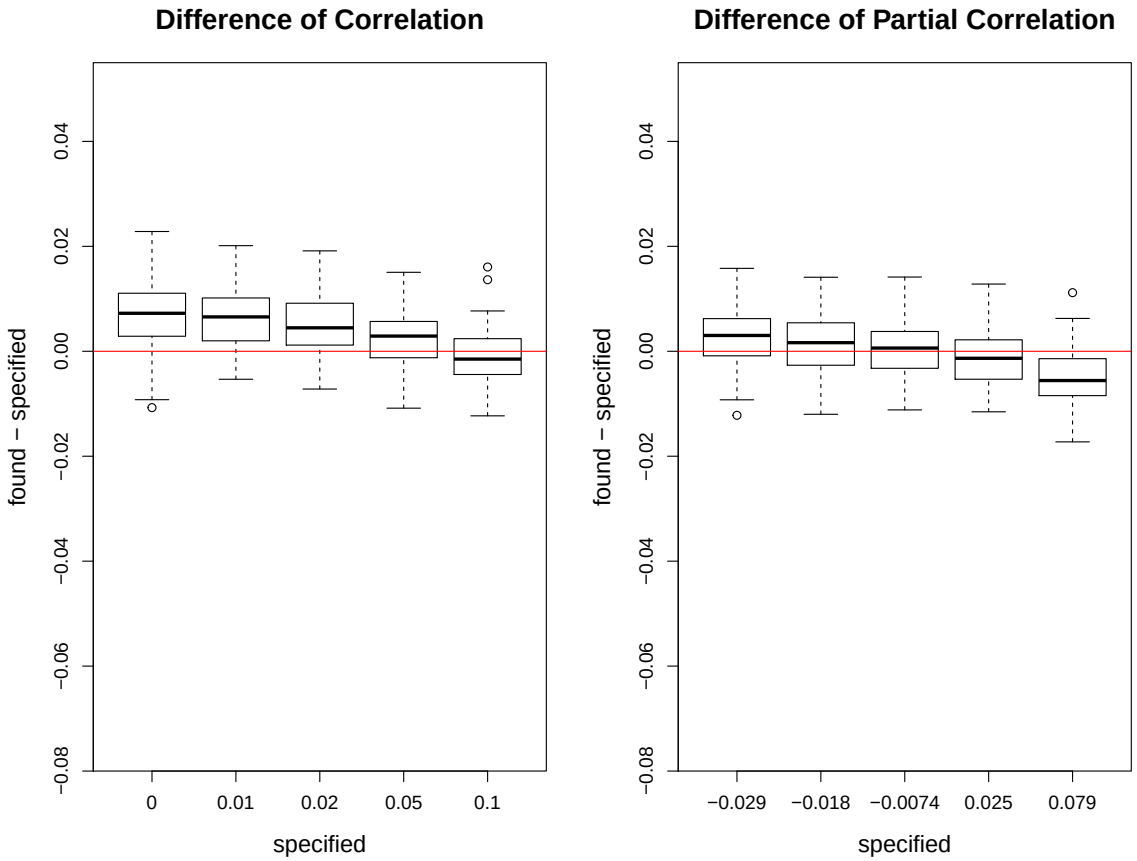

Figure 6: Plots of Found-Specified by Cubic Spline Smoothing Under GCV

In Figure 7, both pre-specified correlation and partial correlation can be found accurately; the significance can be detected when the pre-specified correlation is higher than 0.05.

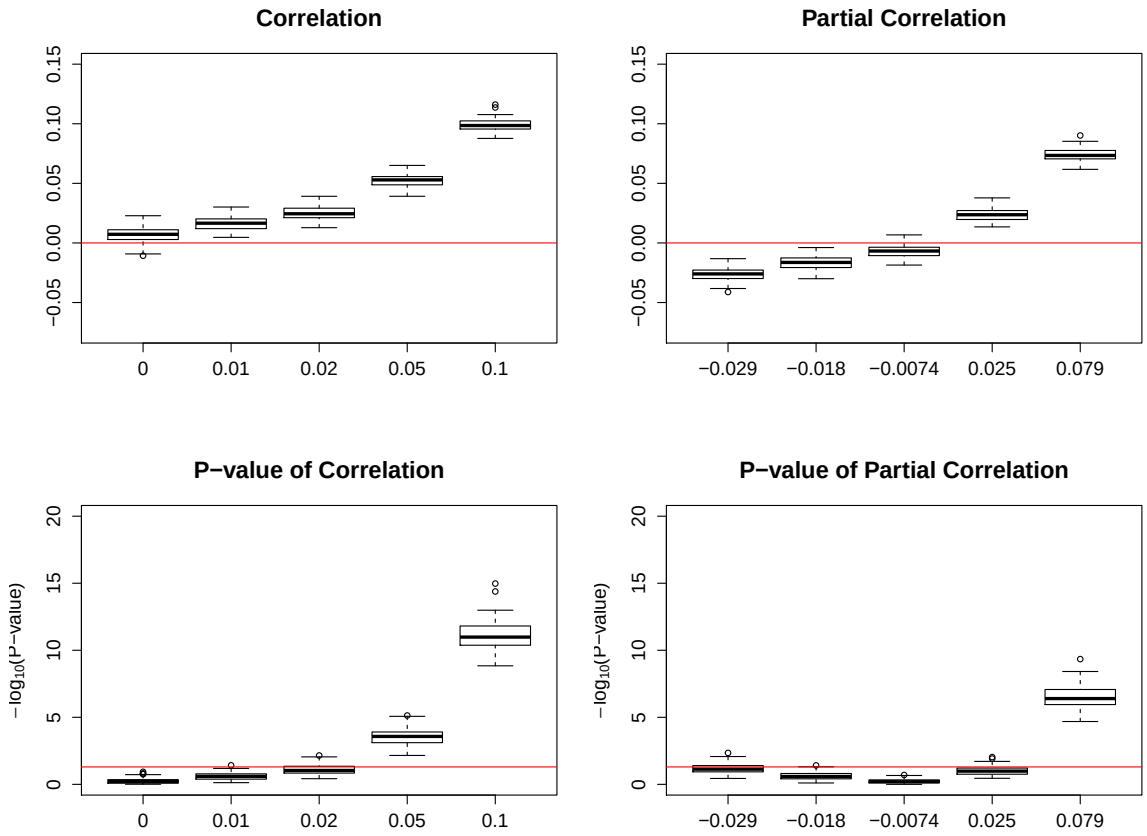

Figure 7: Boxplots of Association by Cubic Spline Smoothing Under GCV; Red Line (above) = 0, Red Line (below) = $-log_{10}0.05$

Without the loss of generality, the highly flexible $S_p = -1.5$ and the moderately flexible $S_p = 0.60$ are selected for other smoothing cases.

In Figure 8, it can be seen that the estimated correlations or partial correlations display some systematic bias,i.e., the estimated correlations are uniformly smaller while the estimated partial correlations are uniformly larger than the pre-defined values. Also, the estimation variation is large.

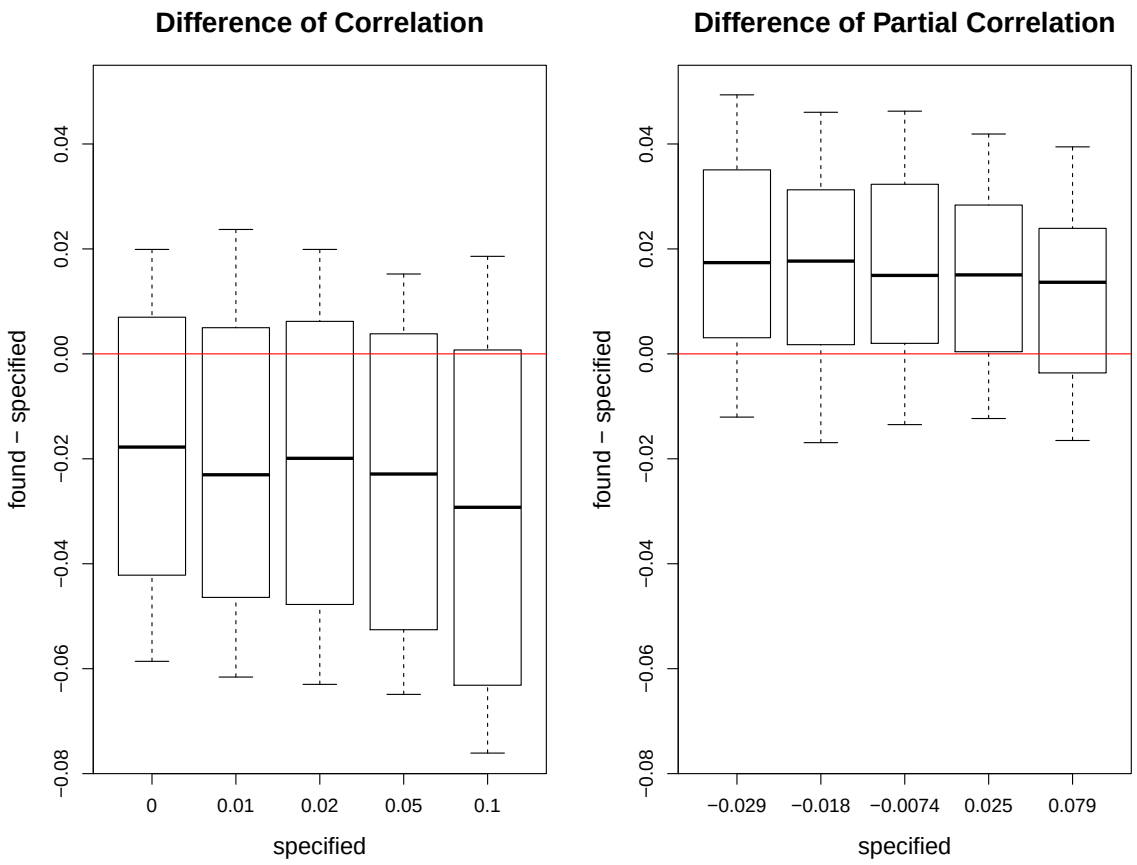

Figure 8: Plots of Found-Specified by Cubic Spline Smoothing

In Figure 9, statistical significance generally cannot be obtained except when the pre-defined correlation is 0.10, for either correlation or partial correlation.

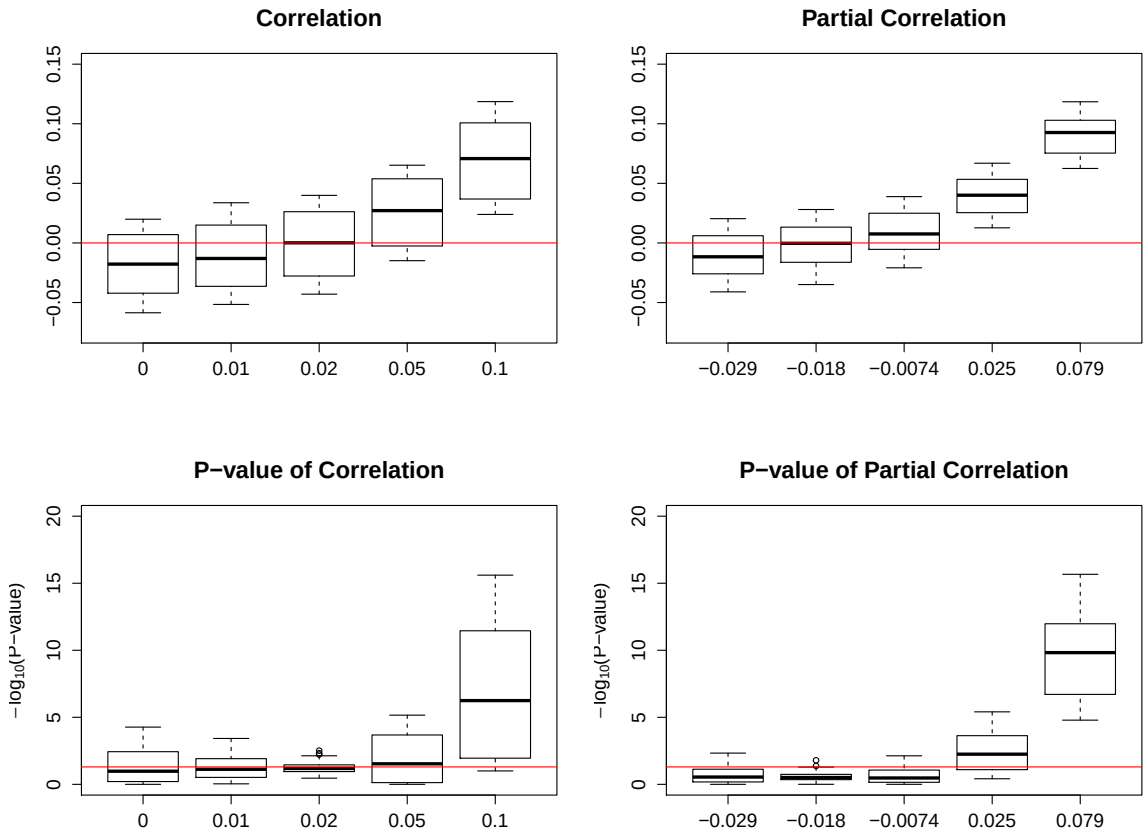

Figure 9: Boxplots of Association by Cubic Spline Smoothing;
Red Line (above) = 0, Red Line (below) = $-log_{10}0.05$

Proposed Smoothing Methods

Following the convention above, moving trimmed mean is demonstrated. Again, moving average with trimming percentage = 0% and moving median with trimming percentage = 50% are used for simulation.

In Figure 10, it can be seen that when the pre-specified correlation or partial correlation increases, the estimated correlation or partial correlation will be more accurate. When the pre-specified correlation or partial correlation is relatively small, the found correlation or partial correlation tends to be slightly larger than the pre-specified ones.

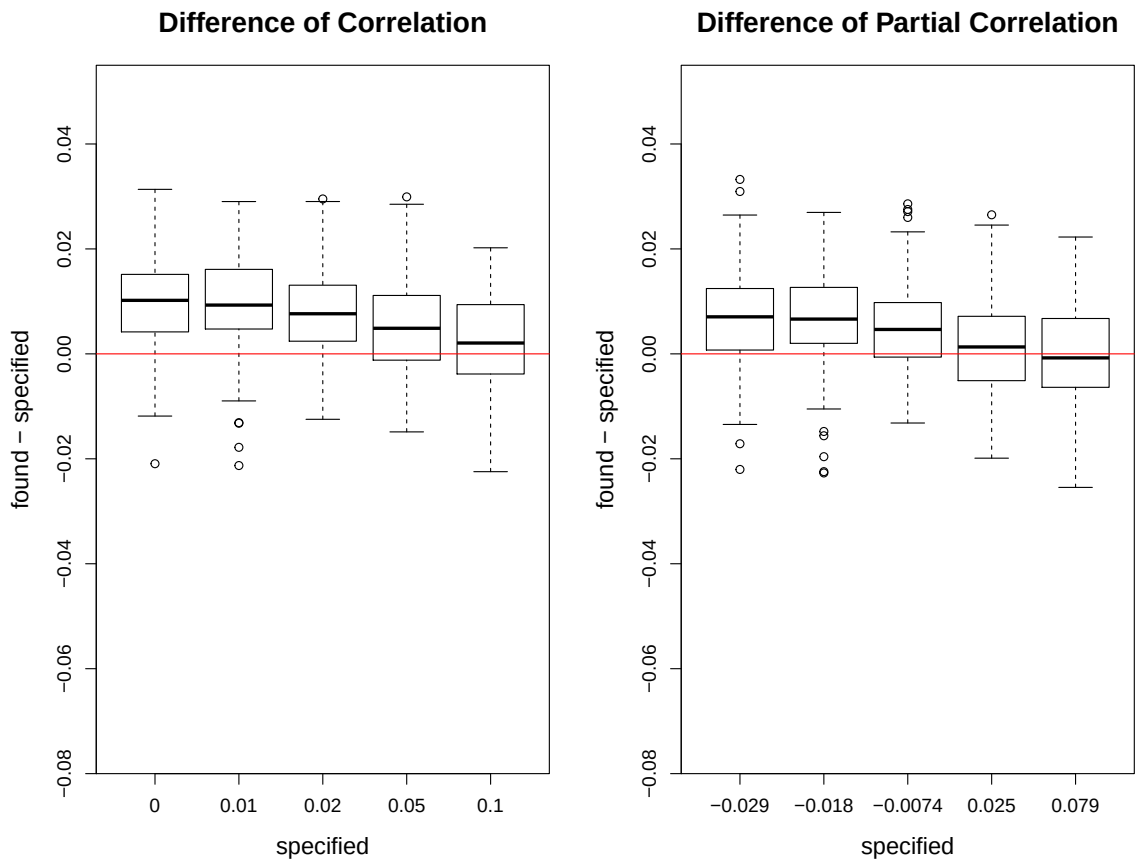

Figure 10: Plots of Found-Specified by Moving Trimmed Mean

In Figure 11, when the pre-specified correlation is 0, no statisitcal significance can be found in either correlation or partial correlation. when the pre-specified correlation is small, such as 0.01 or 0.02, even if the correlations can be correctly found, we cannot obtain statistical significance. When the pre-specified correlation becomes larger than 0.05, our estimate is more accurate. Likewise, the partial correlation behaves in the similar way except that less significance can be found when the pre-specified association is set small.

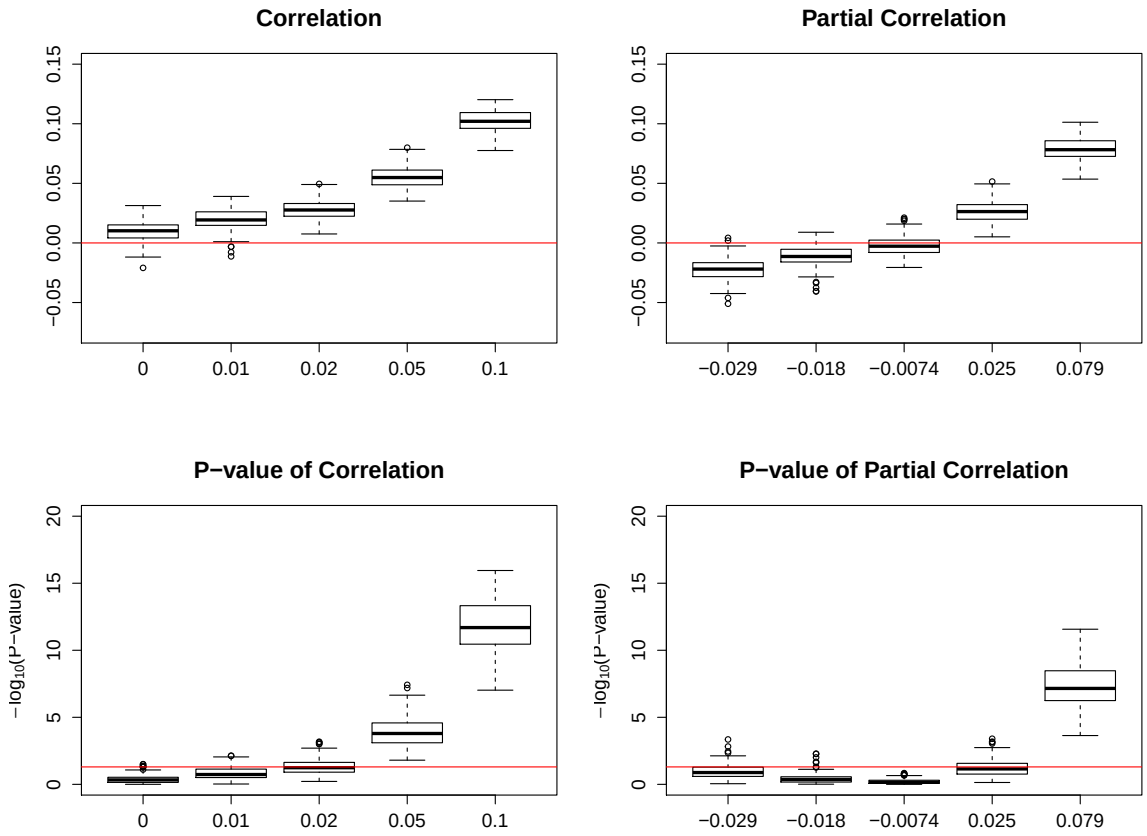

Figure 11: Boxplots of Association by Moving Trimmed Mean;
Red Line (above) = 0, Red Line (below) = $-log_{10}0.05$

5.3 Discussion & Recommendation

Without proper tuning parameter, cubic spline smoothing can alleviate the association or even bias the estimation; it may also introduce large variation in estimation. Three proposed classes of time series smoothers can correctly detect small correlations and partial correlations. The precision can be improved when the association strengthens. Meanwhile, if more accurate inner trend estimates can be obtained, abnormal signals will be better recovered. One possible remedy is to adjust the window and gap size adaptively. For instance, we can enlarge the window and gap size during the non-volatile period and shrink their sizes during the volatile period under certain criterion.

Another possible remedy is ideally to choose the most appropriate measure within each rolling window with centered gap. Assuming that the de-trended time series are weak stationary, there is a certain measure of central tendency that can achieve the optimality within the moving window. For instance, when the fluctuations around the inner trend are normally distributed,

moving average is statistically optimal for recovering the abnormal signals; when the fluctuations around the inner trend are Laplace distributed, which places higher probability on rare events than does the normal, moving median is statistically optimal. In more complicated distributions, different weight schemes in moving average can empirically separate the inner trend and abnormal signals well. Different recursive weighting schemes can also be useful in greedily searching for abnormal signals. In the simulation setting, multivariate normal datasets are generated and all the measures behave very well under multivariate normal distribution. Thus, the simulation results are quite similar.

After discussion, the general recommendation can be delivered as follows. First of all, the window and gap size should be chosen based on the scientific background or the literatures. Ideally, they should be adaptive to capture the central trend. If no other information is available, the user can simply choose any pair of window and gap described above. Due to the robustness of the proposed smoothers, different window and gap sizes should always detect the association reasonably well. Secondly, if one has no prior knowledge or preference on the moving measure, moving trimmed mean should be used; the trimming percentage can be decided by the user, depending on how robust one wants for the central tendency within each window. If one has some prior knowledge or preference on the moving measure, moving weighted mean should be used; the weight scheme is upon the user, depending on the relevance of each point to the central location. If one wants to extract abnormal signals more greedily, moving recursive weighted mean can be used; the re-weighting scheme is also up to the user's choice, depending on how robust one wants for the overall central tendency.

6 Conclusion

In this article, the major formulations of smoothing in air quality studies are studied and summarized systematically. A potential pitfall has been found in the current spline or kernel smoothing methods. From the demonstration of cubic spline smoothing, it is shown that the current spline or kernel smoothing can obtain any significant results, from negative to positive, via varying the smoothing parameter. This can undermine the authenticity in recovering the relationship between air quality and acute death.

Hence, robust and stable time series smoothers with control characteristics are proposed. Each smoothed estimate serves as an experimental control at the time point of interest. There are three classes of robust time series smoothers: moving trimmed mean, moving weighted mean and moving recursive mean. Moving trimmed mean uses the trimming percentage to control the robustness of each experimental control estimate. Moving weighted mean can incorporate the users' prior knowledge or preference to control the degree of de-trending. Moving recursive weighted mean can make the abnormal signals or observations more noticeable by redistributing the weights recursively. All of them can generate robust and stable de-trended time series while allowing reasonable flexibility.

When examining the behaviors and properties of time series smoothers, factorial design on time series smoothers is utilized. Linear model and analysis of variance are applied to find the main and interaction effects. It is shown that the proposed time series smoothers can have correct and consistent results in recovering associations. Meanwhile, users are allowed to adjust the smoothness by changing window and gap size. Although the robustness and stability of the proposed smoothers guarantee the result's consistency, it is recommended that one chooses the window and gap size by the prior literature or domain knowledge. The window size is approximately the length of the relevant information to the time point of interest and the gap size is determined by the average length of the shock period. Within each window, users can choose different class of smoothers to further suit their needs, according to the characteristics of the time series data in their subject of study. It is recommended that one should use moving trimmed mean if no extra information is available with respect to the distribution of the time series. Otherwise, one should consider using moving weighted mean with an empirical weight scheme that addresses the distribution. If one wants to estimate the central tendency conservatively and

retain the abnormality boldly, moving recursive weighted mean can be adopted with a proper recursive re-weight scheme. In a nutshell, the time series smoothers provide us a reliable and efficient collection of tools that do not require fine tuning parameters and produce unbiased and consistent results.

Appendices

A Analysis of Variance: Moving Trimmed Mean

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0640	0.0019	34.58	0.0000
Window	0.0004	0.0000	7.57	0.0000
Gap	0.0031	0.0003	12.34	0.0000
Lambda	0.0037	0.0036	1.01	0.3143
TypeDR	-0.0298	0.0021	-14.11	0.0000
TypeRD	-0.0225	0.0021	-10.67	0.0000
Window:Gap	-0.0001	0.0000	-10.90	0.0000
Window:Lambda	-0.0002	0.0001	-2.57	0.0107
Window:TypeDR	-0.0004	0.0001	-8.52	0.0000
Window:TypeRD	-0.0005	0.0001	-9.15	0.0000
Gap:Lambda	0.0000	0.0003	0.04	0.9683
Gap:TypeDR	-0.0006	0.0002	-3.56	0.0004
Gap:TypeRD	-0.0010	0.0002	-5.56	0.0000
Lambda:TypeDR	-0.0204	0.0029	-6.94	0.0000
Lambda:TypeRD	-0.0192	0.0029	-6.54	0.0000

Summary of Correlation versus Variables and Interactions

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0426	0.0020	21.10	0.0000
Window	0.0002	0.0001	3.98	0.0001
Gap	0.0024	0.0003	8.47	0.0000
Lambda	0.0065	0.0040	1.62	0.1057
TypeDR	-0.0718	0.0023	-31.15	0.0000
TypeRD	0.0493	0.0023	21.39	0.0000
Window:Gap	-0.0000	0.0000	-7.42	0.0000
Window:Lambda	-0.0002	0.0001	-2.46	0.0144
Window:TypeDR	0.0001	0.0001	2.12	0.0345
Window:TypeRD	0.0007	0.0001	12.43	0.0000
Gap:Lambda	0.0003	0.0003	0.80	0.4218
Gap:TypeDR	-0.0005	0.0002	-2.75	0.0064
Gap:TypeRD	0.0002	0.0002	0.81	0.4166
Lambda:TypeDR	-0.0126	0.0032	-3.93	0.0001
Lambda:TypeRD	-0.0260	0.0032	-8.10	0.0000

Summary of Partial Correlation versus Variables and Interactions

References

Bhaskaran, Krishnan et al. (2013). "Time series regression studies in environmental epidemiology." In: *International Journal of Epidemiology* 42.4, pp. 1187–1195.

Chay, Kenneth Y and Michael Greenstone (2003). *Air quality, infant mortality, and the Clean Air Act of 1970*. Tech. rep. National Bureau of Economic Research.

Dominici, Francesca et al. (2002). "On the use of generalized additive models in time-series studies of air pollution and health." In: *American journal of epidemiology* 156.3, pp. 193–203.

Lopiano, Kenneth K, Richard L Smith, and S Stanley Young (2015). "Air quality and acute deaths in California, 2000-2012." In: *arXiv preprint arXiv:1502.03062*.

Peng, Roger D, Francesca Dominici, and Thomas A Louis (2006). "Model choice in time series studies of air pollution and mortality." In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.2, pp. 179–203.

Ramsay, Timothy O, Richard T Burnett, and Daniel Krewski (2003). "The effect of concurvity in generalized additive models linking mortality to ambient particulate matter." In: *Epidemiology* 14.1, pp. 18–23.

Schwartz, Joel (1994). "Nonparametric smoothing in the analysis of air pollution and respiratory illness." In: *Canadian Journal of Statistics* 22.4, pp. 471–487.

— (1999). "Air pollution and hospital admissions for heart disease in eight US counties." In: *Epidemiology* 10.1, pp. 17–22.

Schwartz, Joel, Antonella Zanobetti, and Thomas Bateson (2003). "Morbidity and mortality among elderly residents in cities with daily PM measurements." In: *Revised analyses of time-series studies of air pollution and health*, pp. 25–58.

Touloumi, Giota et al. (2004). "Analysis of health outcome time series data in epidemiological studies." In: *Environmetrics* 15.2, pp. 101–117.