AIR QUALITY AND HEALTH EFFECTS:
SYSTEMATIC INVESTIGATION

A Dissertation in
Statistics
by
Cheng You

© 2019 Cheng You

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2019
The dissertation of Cheng You was reviewed and approved* by the following:

Dennis K.J. Lin  
Distinguished Professor of Statistics  
Dissertation Advisor, Chair of Committee

Ephraim M. Hanks  
Professor of Statistics  
Graduate Head

Yanyuan Ma  
Professor of Statistics

Fuqing Zhang  
Professor of Meteorology

*Signatures are on file in the Graduate School.
Abstract

To examine whether contemporary datasets support the cross-association between air quality and human health, a systematic statistical investigation is conducted. Our major contributions are proposing a novel detrending method on time series data to address the statistical instability of cross-association detection, investigating the method properties and selecting the parameters. In addition, two prevailing detection methods, time series regression and case-crossover analysis, are studied and compared side by side to demonstrate their performances and consistency. This thesis is mainly threefold. First, the problem and research objectives are elaborated. Second, the new detrending methodology is proposed along with its properties and parameter selection; the two common detection methods are studied and compared. Third, the general procedures are presented and the conclusions are made along with discussions. It is found that the current spline detrending methods may induce any cross-correlation, from negative to positive, while the proposed time series smoothers provide a consistent solution to identify the correct short-term effects. The proposed methods’ properties, parameter selection, and diagnosis can also ensure the new opportunity in studying the cross-association among multiple time series data.

In environmental epidemiology, especially the air quality studies, it is often encountered that multiple time series data with a certain long-term trend, including seasonality, cannot be fully adjusted by the observed covariates. The long-term trend is difficult to separate from abnormal short-term signals of interest. The newly proposed detrending method addresses how to estimate the long-term trend in order to recover short-term signals. Our case study demonstrates that the current spline detrending methods can result in significant positive and negative cross-correlations from the same dataset, depending on how the smoothing parameters are chosen. To circumvent this dilemma, three classes of time series smoothers are proposed to detrend time series data. These smoothers do not require fine-tuning
of parameters and can be applied to recover short-term signals. The properties of these smoothers are shown with both a case study using a fully crossed factorial design and a simulation study using datasets generated from the original dataset. General guidelines are provided on how to discover short-term signals from time series with a long-term trend. The benefit of this research is that a problem is identified and characteristics of possible solutions are determined.

Furthermore, since the London Great Smog of 1952 was estimated to have caused the acute deaths of over 4000 people, scientists have studied the relationship between air quality and human mortality. Currently, the cross-association between air quality and acute deaths is usually taken as evidence for causality. As global air quality has markedly improved since 1952, do contemporary datasets support this view? A large dataset, eight air basins in California, is assembled to examine the possible cross association of ozone and PM$_{2.5}$ with acute deaths after removing seasonal and weather effects. A regression-corrected, case-crossover analysis for all non-accidental deaths age 75 and older of different causes was conducted. A stepwise time series regression was used to examine three causes of deaths. After seasonal and weather adjustments, there was essentially no predictive power of ozone or PM$_{2.5}$ for acute deaths. The case-crossover analysis produced odds ratio very close to 1.000, which means no acute effect. The very narrow confidence limits indicated good statistical power. The recent air quality in California was investigated in both time-stratified, symmetric, bi-directional case-crossover and time series regression methods and both give consistent results. There is no statistically significant cross association between either ozone or PM$_{2.5}$ and acute mortality. In the absence of an association, the causality is in question.
# Table of Contents

List of Figures ........................................ viii
List of Tables ........................................... xii
List of Symbols .......................................... xiv
Acknowledgments ......................................... xv

## Chapter 1

**Introduction** ........................................ 1
1.1 Background ........................................... 1
1.2 Objective ............................................. 3
1.3 Contribution .......................................... 5
1.4 Organization .......................................... 6

## Chapter 2

**Literature Review** .................................... 9
2.1 Case-Crossover Design and Analysis ................. 9
2.2 Time Series Regression .................................. 10
2.3 Research Issues ........................................ 11

## Chapter 3

**The Air Quality and Mortality Dataset in California** 14
3.1 Background ............................................. 14
3.2 California Dataset ...................................... 15

## Chapter 4

**Intrinsic Detrending: Time Series Smoother for Effect Detection** 20
Chapter 7

General Procedures & Conclusions  83
7.1 Manual on Detrending and Detecting  83
7.2 Conclusion and Future Work  84

Bibliography  86
List of Figures

1.1 **Air Quality and Health Issues** BaP = benzo(a)pyrene, NO$_2$ = nitrogen dioxide, O$_3$ = ozone, PM = particulate matter, SO$_2$ = sulphur dioxide Source: European Environment Agency, 2013 .......................... 2

3.1 **Air Basins of California**: Mountain Counties, Sacramento Valley, Salton Sea, San Diego County, San Francisco Bay, San Joaquin Valley, South Central Coast, South Coast ............................................. 16

3.2 **Twenty Data Entries of Los Angeles**: Column headers are the variable names of air quality and human mortality ........................................... 17

3.3 **Time Series Plots**: Daily Mortality above Age 75, PM$_{2.5}$, O$_3$, Min Temperature, Max Temperature, Max Relative Humidity ........................................... 18

3.4 **Wildfire near Los Angeles on July 3, 2008**: Satellite Picture (Left), Mortality above Age 75 (Right Top), and PM$_{2.5}$ (Right Bottom) ............................................................... 19

3.5 **Wildfire near Los Angeles on August 31, 2009**: Satellite Picture (Left), Mortality above Age 75 (Right Top), and PM$_{2.5}$ (Right Bottom) ............................................................... 19

4.1 **Raw Time Series and Smoothing Splines of Mortality and Ozone** The left three figures are the time series of mortality and the right three figures are the time series of ozone. The black dots represent the observed values and the red curves represent the smoothing splines. The smoothing parameters from top to bottom are set to be 1.5, 0.6, -1.5, representing highly smooth, moderately smooth and highly rough splines. ................................................................. 34
4.2 **Volcano Plot of Regression Coefficient and P-values by Moving Trimmed Mean** The x-axis is the effect size measured by regression coefficient. The y-axis is the transformed p-value by $-\log_{10}()$; the larger transformed value, the more significance. The black dots represent each coefficient and the red lines are $y = -\log_{10}0.05$: any black dot above the red line means that the coefficient is nominally significant. 

4.3 **Boxplots of Association between Mortality and Ozone by Moving Trimmed Mean** The top two figures are the correlation and partial correlation between mortality and ozone; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has three levels of the factor Type. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under different Window, Gap and Lambda.

4.4 **Plots of Difference between Found and Specified in Correlation and Partial Correlation by Cubic Spline Smoothing Under GCV** The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation after applying cubic spline smoothing under GCV. The red lines are $y = 0$. Each box plot is the graphical summary under simulations.

4.5 **Boxplots of Association between Mortality Deviations and Ozone Deviations by Cubic Spline Smoothing Under GCV** The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.
4.6 **Plots of Difference between Found and Specified in Correlation and Partial Correlation by Cubic Spline Smoothing**
The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation after applying cubic spline smoothing under the two smoothing scenarios. The red lines are $y = 0$. Each box plot is the graphical summary under simulations.

4.7 **Boxplots of Association between Mortality Deviations and Ozone Deviations by Cubic Spline Smoothing**
The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.

4.8 **Plots of Difference between Found and Specified in Correlation and Partial Correlation by Moving Trimmed Mean**
The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation under the two trimming percentages. The red lines are $y = 0$. Each box plot is the graphical summary under simulations.

4.9 **Boxplots of Association between Mortality Deviations and Ozone Deviations by Moving Trimmed Mean**
The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.

5.1 **Illustrative Example: extracting the short-term signal**
Black Dotted Line = Original Data, Red Line = Spline Smoothing (GCV), Green Line = Kernel Smoothing (GCV), Blue Line = Proposed Smoothing (Mean), Blue Dotted Line = Moving Gapped Windows.
5.2 Simulation with Proposed Method and Smoothing Spline Fitting: Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=91, G=11) .......................... 57
5.3 Simulation with Proposed Method and Smoothing Spline Fitting: Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=31, G=3) .......................... 59
5.4 Simulation with Proposed Method and Smoothing Spline Fitting: Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=57, G=15) .......................... 60
5.6 The ACF and PACF of Detrended Ozone Daily Time Series: Window=365 and Gap=0 .......................... 63
5.7 The ACF and PACF of Detrended Ozone Daily Time Series: Window=183 and Gap=0 .......................... 63
5.8 The ACF and PACF of Detrended Ozone Daily Time Series: Window=91 and Gap=0 .......................... 64
5.9 The ACF and PACF of Detrended Ozone Daily Time Series: Window=43 and Gap=0 .......................... 64
5.10 The ACF and PACF of Detrended Ozone Daily Time Series: Window=21 and Gap=0 .......................... 65
5.11 The Coefficient of Variation for Detrended Ozone Daily Time Series For Different Gaps .......................... 66
5.12 Auto Correlation and Partial Auto Correlation for Detrended Ozone with Moving Gapped Mean: Window=43, Gap=7 .......................... 67
5.13 Periodogram for Detrended Ozone with Moving Gapped Mean: Window=43, Gap=7 .......................... 68

6.1 South Coast Air Basin. (a) Daily deaths versus Day of Year, DOY, for the years 2004-2007. Four years are overprinted. (b) Deviations from Day of Year time series smoother, linear and quadratic model. ........................................ 73
6.2 Histogram of p-values. If the effects are random, we should see a uniform distribution. .......................... 78
6.3 P-value plot of 96 tests of hypothesis, $-\log_{10}$ of p-values versus the expectation of p-values coming from a uniform distribution. If the points fall on a 45-degree line, then the results are consistent with randomness. .......................... 78
## List of Tables

4.1 Correlations and Partial Correlations with P-values between Mortality and Ozone ........................................ 35
4.2 Factors and Levels of Factorial Design of Scenarios .............. 35
4.3 Summary of Correlation versus Variables and Interactions ....... 39
4.4 Summary of Partial Correlation versus Variables and Interactions ................................................................. 39
4.5 Correlations and Partial Correlations with P-values by Type for Moving Trimmed Mean: Window=21, Gap=5, Trimming Percentage=50% ................................................. 40

5.1 Numerical Comparison on root mean squared error and mean absolute error: moving gapped mean(Random: W=91, G=11), smoothing spline(CV) and smoothing spline(GCV) .... 57
5.2 Numerical Comparison on root mean squared error and mean absolute error: moving gapped mean(Random: W=31, G=3), smoothing spline(CV) and smoothing spline(GCV) .... 58
5.3 Numerical Comparison on root mean squared error and mean absolute error: moving gapped mean(Random: W=57, G=15), smoothing spline(CV) and smoothing spline(GCV) .... 60

6.1 Numbers of data values imputed in the analysis dataset. PM25davg is the daily average PM$_{2.5}$ level. Tmin and tmax are the minimum and maximum temperature. MAXRH is the maximum relative humidity. .................................................. 74
6.2 Results of case-crossover analysis. Odds ratio(OR), for eight air basins, confidence limits, p-values (raw and adjusted for 16 test); PM$_{2.5}$ and ozone. There are 16 statistical tests of hypothesis. .... 76
6.3 P-values testing health effects versus air quality. All-Cause deaths; CV: cardiovascular deaths; respiratory deaths for 8 air basins. There are 96 tests of hypothesis. ......................... 77
6.4 AllCause risk ratios from cohort studies for PM$_{2.5}$ deaths in California. See Enstrom (2017) for details.
List of Symbols

$y_t$  The mortality count of certain aggregation level at time $t$

$x_t$  The air quality variable of interest at time $t$

$z_t$  The vector of measured meteorological covariates at time $t$

$\alpha$  The intercept of the regression model

$\beta$  The coefficient of air quality variable of interest

$\eta$  The vector of coefficients of measured meteorological covariates

$\mu_t$  The expected mortality count of certain aggregation level at time $t$

$\lambda_1$  The smoothing parameter for the mortality count time series

$\lambda_2$  The smoothing parameter for the air quality variable of interest time series

$D'_t$  The detrended mortality count of certain aggregation level at time $t$

$D''_t$  The detrended air quality variable of interest at time $t$

$l_t$  The long-term trend at time $t$

$\xi_t$  The short-term change at time $t$

$\epsilon_t$  The random error at time $t$

$w_t$  The weight for a specific data point at time $t$

$b_i$  The $i$th coefficient for the regression model on mortality with meteorological adjustment
First of all, I am deeply indebted to my thesis advisor, Dr. Dennis K.J. Lin, for his inspiring guidance, encouragement, and criticism throughout the entire journey of my Ph.D. program. In short, Dr. Dennis K.J. Lin took my statistical, independent and critical thinking up to a whole new level. Meanwhile, I would like to express my heartfelt gratitude to my collaborator Dr. S. Stanley Young for giving me invaluable insights on many different aspects in environmental epidemiology, especially the air quality study. Moreover, I would like to thank my committee members Dr. Ephraim Hanks, Dr. Yanyuan Ma, and Dr. Fuqing Zhang for their beneficial comments and suggestions.

More importantly, I would love to thank my parents for their unwavering love, without which I would not have made the completion of my thesis possible.

Last but certainly not least, I want to thank my friends and colleagues for their kind help and support.
Chapter

Introduction

1.1 Background

In environmental epidemiology, how environmental change would affect our lives is of great interest. Good air quality is fundamental to human well-being. On average, a person inhales about 14,000 liters of air every day and the presence of air contaminants can adversely affect people's health. People with pre-existing respiratory and heart conditions, the younger and the older people are considered more vulnerable.

Figure 1.1 illustrates the possible adverse impacts of air pollution on human health. These impacts can be a headache and anxiety; irritation of eyes, nose, and throat; cardiovascular diseases. Potential impacts can be on the respiratory system, liver, spleen and blood, and the reproductive system.

Studies have shown that poor air quality might also adversely affect the environment. In the short run, under extreme circumstances, ecological damage can occur when air pollutants come into direct contact with vegetation or animals. Air pollutants can settle onto farmland and water bodies. From the soil, they can be washed into waterways, or be taken up by plants and animals. In the long run, poor air quality might affect our climate: some pollutants are claimed to have a warming effect while others contribute to cooling, although further verification research needs to be done.

These impacts of poor air quality on human health and the environment can, in turn, have negative economic impacts. Major costs may be incurred, for example,
for hospitalization, medical treatment and lost work days. In rural areas, damage to soils, vegetation and waterways may reduce the productivity of the agriculture and forestry industries. In urban areas, air pollution can be costly when, for example, transport is disrupted, due to large-scale smog events, or corroded buildings need to be repaired. On the other hand, over excessive regulations on the air quality can harm the new development of certain industries such as oil and gas and cause unnecessary loss of jobs.

Among the adverse effects, do the fluctuations in air quality induce the fluctuations in human morbidity or even mortality? The United States Environmental Protection Agency claims that the ambient PM$_{2.5}$ and O$_3$ can cause acute death and thus would enforce cleaner energy production by new regulations. Although their claim should be evaluated with some caution, the air quality problem does

---

**Figure 1.1. Air Quality and Health Issues**

BaP = benzo(a)pyrene, NO$_2$ = nitrogen dioxide, O$_3$ = ozone, PM = particulate matter, SO$_2$ = sulphur dioxide

Source: European Environment Agency, 2013
not only affect long-term human welfare but also poses major economic challenges to the United States and many other countries, especially those developing countries such as China and India. In general, the health effects of air quality can be classified into two categories: short-term or acute effect and long-term or chronic effect. The short-term or acute effect is the adverse effect, due to exposure to a harmful substance on humans, whereby severe symptoms develop rapidly and lead quickly to a health crisis or even death. These symptoms often subside when the exposure stops. The long-term or chronic effect is the adverse effect on humans with symptoms that develop slowly, due to long and continuous exposure to low concentrations of a hazardous substance. Such symptoms do not usually subside when the exposure stops. To begin with, the acute effect of air quality on human mortality is investigated.

1.2 Objective

To study the acute effect between the contemporary air quality and human mortality, multiple time series of air quality, mortality attributed to different causes, and meteorological covariates are assembled and investigated. Will human mortality change when a certain air quality indicator changes? How large of the air quality change would possibly induce the mortality change? Is the detected cross association statistically significant? Those are the major questions to be answered. To elaborate further, there are two major objectives in finding the cross-association: detrending and detecting.

First of all, time series detrending is required to reveal the unusual signals. A trend is a general movement over time of a statistically detectable change. Usually, a trend will distort or obscure the relationships of interest. In our case, the cross-association between air quality and human mortality is largely obscured by the meteorological and unmeasured covariates; hence, intrinsic detrending becomes our very first objective. Being intrinsic requires that the method used in defining the trend be adaptive, so that the trend extracted is derived from and based on the data. In traditional time series analysis, a time series was decomposed into linear trend, seasonal or periodic components, and irregular fluctuations, and the various parts were studied separately. Modern analysis techniques frequently
treat the series without such routine decomposition, but separate consideration of trend is still often required. However, there is no precise definition of trend, not to mention long-term trend and short-term change. Various ad hoc extrinsic methods have been used to determine the trend and to facilitate a detrending operation. However, those extrinsic methods with preselected functional forms or simplified assumptions are shown to be problematic in terms of reducing the actual relationship or inducing some spurious correlation.

To avoid the ambiguity, one tentative definition of time threshold on long-term and short-term has been proposed. The long-term trend is any observable movement or frequency that occurs over a specific period of time, for example, a month. The short-term change is any observable movement or frequency that occurs under the specific period of time. For the specific time period, there is no absolute setting rule. Researchers can do sensitivity analysis or cross-validation for the time period selection.

Second, detecting the cross-association among multiple time series with or without lags is demanded. In general, dependence measures among time series need to be utilized or proposed. In the literature, there are two prevailing modeling methods: case-crossover analysis and time series regression. The case-crossover analysis is an analytical epidemiological approach and is unique in that the case serves as one’s own control and is used to investigate the transient effects of an intermittent exposure on the onset of acute outcomes. Time series regression is a conventional statistical method for predicting a future response based on the response history (known as autoregressive dynamics) and the transfer of dynamics from relevant predictors. Time series regression can help us understand and predict the behavior of dynamic systems from the experimental or observational data. Do both methods provide consistent statistical results? This question is one of our great interest. Future work could be done towards how to propose a reliable dependence measure.

In a nutshell, multiple time series data in the air quality study pose new challenges in different aspects. First, times series data violate the assumption of independence. Second, there is a long-term trend including seasonality, which needs to be pre-defined and estimated. Third, after long-term trend estimation and certain detrending, there may still be different covariance structures across different
time periods. Fourth, how to measure the cross-association and extract the short-
term signals is another challenge ahead. Hence, new statistical methodologies are
required to further the research of air quality and health effects.

1.3 Contribution

In previous work, researchers have endeavored to understand whether contempo-
rary air quality affects human health. For detrending, parametric (natural) spline,
penalized spline, and smoothing spline have been applied. They have different pros
and cons in smoothing the time series while estimating the long-term trend but
not necessarily to the accurate estimation of the air quality effect. For detecting,
there are two prevailing methods: case-crossover analysis and time series regres-
sion. Both methods will be reviewed in detail in Chapter 2. Time series regression
is becoming more popular these days, however, there are potential optimization
issues regarding the generalized additive models to estimate the coefficient accu-
rately.

In air quality studies, the main challenges can be summarized as follows. First
of all, the time series data have a certain long-term trend that includes seasonality
and other potential cycles; the exact definition of long-term trend is not clearly
proposed. This poses identifiability issues in air quality studies. Second, after the
identifiability issue is well addressed, how we accurately estimate the long-term
trend becomes the next obstacle in studying the short-term effect. In general,
parametric, semi-parametric and nonparametric methods are invented for the re-
gression purpose and have their own pros and cons. Which method would be a
good fit in the air quality study requires theoretical justification and numerical
demonstration. Third, after accounting for the long-term trend, how to measure
and infer the short-term effect is also very challenging; the short-term signals are
sparse in nature and changing over time. It is even more difficult to deduce the
causality of increased human mortality and poor air quality.

The contributions towards the field of statistics and epidemiology are mainly
in twofold, the scientific and statistical aspect. In air quality studies, the scientific
contributions are stated as follows. First of all, the characteristics of air quality and
acute effect are identified. Both the level and change of air quality are examined as
well as their cross association with human mortality. Second, it is concluded that the contemporary air quality in California does not have an acute effect on elderly individuals. Third, researchers should be very cautious about certain air quality conditions, especially when the wind or precipitation is low and the sunlight is strong.

In the field of statistics, our statistical contributions for time series detrending are stated below. First of all, new definitions of the long-term trend and short-term signal are given under the time and frequency frame to avoid an ad hoc judgment. Second, a new collection of time series smoothers are proposed and targeted to reveal the short-term signals. Third, a new recursive algorithm is proposed in the moving recursive mean to exhaustively extract the short-term signals. These works are documented in “Time series smoother for effect detection” [You et al., 2018b]. Our statistical contributions for detecting cross associations are summarized below. First of all, two prevailing methods, case-crossover analysis and time series regression, in air quality studies are utilized, compared, and cross-verified in air quality studies. Second, multiple testing is addressed and emphasized in air quality studies. These work are documented in “PM 2.5 and ozone, indicators of air quality, and acute deaths in California” [You et al., 2018a].

In a nutshell, the statistical importance and contributions in the thesis are reiterated below. Firstly, a statistical manual is proposed on detrending and detecting the cross-association among multiple time series, either the level or change. Secondly, a collection of valid and stable time series smoothers are devised to reveal the short-term signals. Last but not least, the pros and cons of different methods are compared and analyzed in air quality studies.

1.4 Organization

The organization of the thesis is outlined as follows.

In Chapter 2, the literature review was done for the two prevailing methods in air quality studies, case-crossover analysis and time series regression. Case-crossover is one of the most used designs for analyzing the health-related effects of air pollution. The case-crossover design and analysis are used to identify risk factors of acute events; it is characterized by the fact that each subject serves as
his or her own control by assessing referent exposure at a point in time prior to the event. Time series regression is the other prevailing method in understanding the air quality and acute health-related issues. Time series data on air quality and human mortality are generally analyzed using log-linear, Poisson regression models for overdispersed counts with the daily number of deaths as an outcome, the possibly lagged daily level of pollution as a linear predictor, and smooth functions of weather variables and calendar time used to adjust for time-varying confounders. In addition, the statistical research issues in air quality studies are also briefly addressed.

In Chapter 3, the air quality and mortality dataset in California is described in details. More background information has been introduced to emphasize the importance of the air quality problem in California. Different air quality variables are introduced as well as their characteristics. Plots are made to further illustrate the time series data and potential challenges.

In Chapter 4, the intrinsic detrending method, time series smoother for effect detection, is proposed. First, the major formulations of smoothing in air quality studies are reviewed and a potential pitfall has been found: the prevailing spline smoothing methods can obtain any significant result, from negative to positive, by varying the smoothing parameters. Second, three classes of valid and stable time series smoothers are proposed and each smoothed estimate can serve as an experimental control at the time point of interest. Third, the characteristics and sensitivities of time series smoothers are examined using a factorial design of scenarios in our case illustration and indistinguishable synthetic datasets in our simulation.

In Chapter 5, the properties of the proposed time series smoothers are studied. Due to the inherent identification issue, validity and reliability of the proposed methods are shown. Moreover, a graphical method on the selection of window and gap sizes is proposed. The detrending diagnosis is also proposed to examine whether a certain detrending method is acceptable.

In Chapter 6, two prevailing methods, case-crossover analysis and time series regression, are demonstrated and compared, which give consistent results. In the absence an association of air quality, as measured by ozone or PM$_{2.5}$, with acute mortality, there is no evidence supporting current air quality being causal of acute
deaths in California.

In Chapter 7, the general statistical procedures on time series detrending and detecting are proposed in air quality studies. The contribution and conclusion are presented. Limitations and potential future work are discussed.
Ever since the London Great Smog of 1952 killed over 4,000 people, scientists have studied the relationship between air quality and acute mortality. In [Dockery et al., 1993], Dockery et al. claimed to observe a statistically significant association between air pollution and mortality from six communities, after adjusting for smoking and other risk factors. They used Cox proportional-hazards models and found a higher risk ratio. In [Styer et al., 1995], Styer et al. found no evidence that particulate matter contributed to excess mortality in one county and some evidence that positive effect occurred during spring and autumn in the other county by using generalized additive regression models with the log link so the overall effect on mortality was unconfirmed. In the United States, the air quality issues, especially particulate matter, drew a lot of public attention after the two papers’ appearance and new methodologies are greatly demanded in the air quality study. The methodologies in the air quality study gradually evolved into two categories, case-crossover analysis and time series regression.

2.1 Case-Crossover Design and Analysis

Case-crossover design was first proposed in [Maclure, 1991] and then gained its popularity in accessing the acute effect. In [Navidi, 1998], bidirectional case-crossover designs were proposed to handle time trends. In [Bateson and Schwartz, 1999], symmetric bidirectional case-crossover designs were proposed to handle both time trend and seasonal variation. In [Lumley and Levy, 2000], time-stratified case-
crossover designs were proposed to provide more adequate controls and thus reduce biases. In [Maclure et al., 2000], the concurrent case-crossover designs are reviewed with suggestions and limitations clarified. In [Navidi and Weinhandl, 2002], semisymmetric bidirectional case-crossover designs were proposed to increase the method’s efficiency. In [Janes et al., 2005], a new taxonomy of referent selection strategies is proposed to reduce overlap bias and gain favorable statistical properties. In [Carracedo-Martinez et al., 2010], a systematic review of case-crossover designs was published; the pros and cons of each variation were discussed and some future research directions were identified. In [Avalos et al., 2012], case-crossover design via sparse conditional likelihood was proposed and several variable selection procedures were applied in the context of case-crossover studies. In [?], the time-stratified case-crossover design was applied on the large-scale myocardial infarction dataset of the Myocardial Ischaemia National Audit Project database and found that no clear evidence for pollution effects on myocardial infarction and stroke. In [Doerken et al., 2016], the case-crossover design via penalized regression was proposed; penalized conditional logistic regression via the lasso was claimed to yield better risk classifications and more plausible risk estimates than standard methods. In [Hallas et al., 2016], the existence of a bias by persistent-user contamination in the execution of case-crossover designs when being applied to medication exposures was revealed and further adjustment was in need. In [Maclure, 2017], the case-crossover design was evaluated and studied by using simulated data; it has inferred that estimates of risk ratios from case-crossover analyses are generally closer to true risk ratios when effects of exposure are more transient and outcome onset is more abrupt.

### 2.2 Time Series Regression

On the other hand, time series regression was also developed largely in air quality studies. In [Schwartz, 1994], time series models on air quality and acute death were suggested, where generalized additive Poisson models were used with LOESS smoothing functions of covariates including time. In [Zeger et al., 2000], Dominici et al. addressed exposure measurement error in observational studies of air pollution and health; a systematic conceptual formulation of the problem of
measurement error in epidemiologic studies of air pollution is proposed to understand the differences between true personal exposure for every individual and measured ambient concentrations. In [Dominici et al., 2002b], [Ramsay et al., 2003], [Schwartz et al., 2003], [Touloumi et al., 2004], [Schimek, 2009] and [Marra and Radice, 2010], several variations of time series regression models were proposed using parametric natural splines and penalized splines. In [Dominici, 2004], a statistical review of time series analysis of air pollution and acute mortality was carried out. In [Peng et al., 2006], model choice in time series regression was studied and natural splines were considered preferable. In [Dominici et al., 2008], more concerns on model choice and effect detection are discussed and the development of appropriate statistical tools remains an open area of investigation. In [Bhaskaran et al., 2013], time series regression was generally reviewed in environmental epidemiology studies. In [Dominici et al., 2014], the limitations of the current methodology were described and alternative approaches were introduced; the air quality problem was raised to a higher level of concern as a global health and economic issue. In [Bernal et al., 2017], interrupted time series regression is proposed for evaluating the effectiveness of population-level health interventions. Basically, segmentation regression is used to enhance the robustness while over-dispersion of time series data, adjusting for seasonal trends and controlling for time-varying confounders are carefully considered. In [Dimakopoulou et al., 2017], spatiotemporal land use regression models are proposed to account for spatial variability; they also found that the effect estimates using classical Poisson regression time series yielded consistently smaller size effects compared to the case-crossover method.

2.3 Research Issues

The analysis of time series biomedical data poses several methodological problems, which results in intense research. The main research issues are summarized below.

Model selection: Time series models are usually built with a pre-defined set of potential confounders. However, some criteria are needed to select other model parameters, such as the degree of control for long-term trend including seasonality, or the adequacy of assumptions on the shape of the exposure-response re-
relationship of predictors showing potential non-linear effects. Some investigators
have tested the comparative performance of selection criteria based on informa-
tion criteria (Akaike, Bayesian or related), minimization of the partial autocorre-
lation of residuals, (generalized) cross-validation and others. Further research is
needed to produce robust and general selection criteria. ([Dominici et al., 2008],
[Crainiceanu et al., 2008], [Baccini et al., 2007], [He et al., 2006] and
[Peng et al., 2006])

Smoothing methods: The specification of non-linear exposure-response rela-
tionship for predictors in the regression model is quite essential, both to determine
the association with the exposure of interest and to control for potential con-
founders. Smoothing techniques based on both parametric and non-parametric
methods have been proposed in time series regression. The former usually rely
on regression splines within generalized linear models (GLM), while the latter are
specified through smoothing or penalized splines within generalized additive mod-
els (GAM). ([Marra and Radice, 2010], [Schimek, 2009], [Wood, 2006],
[Dominici et al., 2002a] and [Dominici et al., 2002b])

Distributed lag (non-linear) models: Commonly the effect of an exposure is not
limited to the day it occurs, but persists for further days or weeks. This introduces
the additional problem of modeling the lag structure of the exposure-response rela-
tionship. This issue has been initially addressed by distributed lag models, which
allows the linear effect of a single exposure event to be distributed over a specific
period of time. More recently, this methodology has been generalized to non-linear
exposure-response relationships through distributed lag non-linear models, a mod-
eling framework which can flexibly describe simultaneously non-linear and delayed
associations. ([Gasparrini, 2014], [Gasparrini, 2011], [Gasparrini et al., 2010],
[Muggeo, 2008] and [Schwartz, 2000a])

Harvesting effect (mortality displacement): This phenomenon arises when ap-
plying an ecological time series analysis to grouped data, for example, mortality
counts. The conceptual framework is based on the assumption that the exposure
can affect mainly a pool of frail individuals, whose events are only brought forward
for a brief period of time by the effect of exposure. For non-recurrent outcomes,
the depletion of the pool following a high exposure event results in some reduction
of cases a few days later, thereby reducing the overall long-term impact. Specific
models are needed to account for this reduction in the overall effect and thereby produce accurate estimates. ([Rabl, 2005], [Schwartz, 2001] and [Schwartz, 2000b])

Two-stage analysis: The usual approach to time series studies of environmental factors involves the analysis of series from multiple cities or regions. The complexity of the regression models prevents the specification of a very highly parameterized hierarchical structure in a single multilevel development. The analysis is instead carried out through a two-stage process, with a common city-specific model and then a meta-analysis to pool the results. The specification of complex exposure-response relationships in the first stage requires the development of non-standard meta-analytic techniques, such as meta-smoothing and multivariate meta-analysis. ([Gasparrini et al., 2012], [Dominici et al., 2000] and [Schwartz and Zanobetti, 2000])

Time-varying confounders: While interrupted time series designs are rarely affected by normal confounders, such as differences in socioeconomic status or age composition, which typically only change relatively slowly over time, they may be affected by time-varying confounders. This is particularly an issue if the confounders are unmeasured and change over the same period as the intervention, for example, other concurrent events or policies. Design adaptations may be introduced to address this limitation such as the introduction of a control series, multiple baseline designs (where the intervention is introduced in different locations at different times) and multiple phases (where the intervention is first introduced then removed to test whether the effect is reversed). ([Wagner et al., 2002] and [Cook et al., 2002])

In a nutshell, there are many different statistical aspects that require improvement and even innovation in air quality studies. In our framework, the research issues are divided into two categories, detrending and detecting. In detrending, smoothing methods are mainly addressed, which appear in Chapter 3 and 4. In detecting, two major detecting methods are demonstrated and compared, which appear in Chapter 5.
Chapter 3

The Air Quality and Mortality Dataset in California

3.1 Background

According to the American Lung Association’s recent “State of the Air 2017” report, California is a leader in air pollution among other states, with the highest ozone levels. California has the most polluted cities in the United States as the administration seeks to force the state to weaken its vehicle emissions standards. California wildfires in the summer cast tons of smokes for weeks, which makes the air quality worse than the winter with no wind. For decades, the air quality problem in California has been discussed, due to potential impacts on the large population. For example, in Los Angeles, there is much sunshine and low rainfall; these weather characteristics contribute to high levels of ozone, fine particles, and dust. Moreover, there is often low wind speed accompanied by high air pressure. This results in little movement of the air mass and more serious problem in certain locations.

Three main factors are behind the poor levels of air quality in California: 1. Large amounts of air pollutants are generated by the activities of 39 million people; 2. Terrain or topography traps the air mass; 3. A warm, sunny climate helps form ozone and other air pollutants. During personal and business activities, Californians release thousands of tons of pollutants into the air every day. Although
each of us may only produce a small amount of air pollution, the combined pollution
from the 39 million Californians adds up to a big problem.

In addition, California is a perfect place for smog. California’s topography (the
physical shape of the land) and its warm, sunny climate are perfect for trapping
and forming air pollutants. Most California cities are built on plains or in valleys
surrounded by mountains. These areas are natural bowls that trap air pollution
and prevent the air from circulating. On some days temperature inversions (where
the air closer to the ground becomes cooler than the air above) act as lids which
trap air pollutants close to the ground. This prevents vertical mixing (the upper,
cleaner air mixing with the lower, polluted air) and the dispersion of pollutants.
On hot, sunny days, pollutants emitted by vehicles, industry, and many products
(nitrogen oxides and volatile organic compounds) react with each other to form
ozone, the main ingredient of smog. During the winter, temperature inversions
can trap tiny particles of smoke and exhaust from cars, trucks, fireplaces, and
anything else that burns fuel. This keeps the pollution close to the ground, right
where people are breathing.

### 3.2 California Dataset

To study the acute effect of air quality on human health, the following dataset was
assembled as an illustrative example. All the data start from Year 2000 to Year
2012 and come from California. The state of California contains eight populous
air basins: Mountain Counties, Sacramento Valley, Salton Sea, San Diego County,
San Francisco Bay, San Joaquin Valley, South Central Coast, South Coast. The
detailed map can be found in Figure 3.1. San Francisco Bay includes San Francisco
and its surroundings; South Coast includes Los Angeles and its satellite cities; San
Diego County includes San Diego and its suburbs.

The air quality data PM$_{2.5}$ and O$_3$ are obtained for each air basin, from Califor-
nia Environmental Protection Agency’s Air Resources Board. PM$_{2.5}$ is the daily
average of atmospheric particulate matter with a diameter of 2.5 $\mu m$ or less in
microgram per cubic meter ($\mu gm^{-3}$). O$_3$ is the daily maximum surface ozone con-
centration averaged over 8 hours in parts per billion (ppb). The minimum and
maximum temperature are obtained from Carbon Dioxide Information Analysis
Figure 3.1. **Air Basins of California**: Mountain Counties, Sacramento Valley, Salton Sea, San Diego County, San Francisco Bay, San Joaquin Valley, South Central Coast, South Coast

Center and the relative humidity is from Environmental Protection Agency. The daily minimum and maximum temperatures are recorded in Fahrenheit (°F). The daily maximum relative humidity is in the percentage of maximum saturation of the air-water mixture. The daily mortality data are obtained by aggregating daily death certificates across different causes from California Department of Public Health. After daily aggregation, there are six outcome mortality variables created: 65 AllCause, 65 HeartLung, 65-74 AllCause, 65-74 HeartLung, 75+ AllCause, and 75+ HeartLung. 65, 65-74 and 75+ mean the age of years under and including 65, between 65 and 74, and above and including 75. AllCause and HeartLung mean all different non-accidental causes and those causes attributed to “Diseases of the
Circulatory System” and “Diseases of the Respiratory System”, respectively. Since the patterns of mortality in the younger age categories are flat across time while those above and including 75 are varying across time, the daily mortality above 75 in each air basin is being investigated carefully. Older people are thought to be more susceptible to environmental changes. In the following, a snapshot of the dataset is shown in Figure 3.2. The time series of mortality above Age 75, PM$_{2.5}$, O$_3$ and weather covariates are plotted in Figure 3.3.

**Figure 3.2. Twenty Data Entries of Los Angeles:** Column headers are the variable names of air quality and human mortality

<table>
<thead>
<tr>
<th>RowID</th>
<th>basin</th>
<th>year</th>
<th>month</th>
<th>day</th>
<th>dayofyear</th>
<th>AllCause75</th>
<th>PM25davg</th>
<th>o3</th>
<th>tmin.0</th>
<th>tmax.0</th>
<th>MAXRH.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>226</td>
<td>68.8</td>
<td>40</td>
<td>37</td>
<td>62</td>
<td>94.3</td>
</tr>
<tr>
<td>2</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>235</td>
<td>7.5</td>
<td>40</td>
<td>19</td>
<td>65</td>
<td>71.0</td>
</tr>
<tr>
<td>3</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>202</td>
<td>39</td>
<td>35</td>
<td>69</td>
<td>53.6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>223</td>
<td>38</td>
<td>37</td>
<td>72</td>
<td>49.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>217</td>
<td>42</td>
<td>38</td>
<td>72</td>
<td>58.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>216</td>
<td>24.3</td>
<td>40</td>
<td>38</td>
<td>74</td>
<td>41.1</td>
</tr>
<tr>
<td>7</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>248</td>
<td>35.8</td>
<td>36</td>
<td>37</td>
<td>69</td>
<td>54.3</td>
</tr>
<tr>
<td>8</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>196</td>
<td>49.6</td>
<td>39</td>
<td>35</td>
<td>71</td>
<td>66.3</td>
</tr>
<tr>
<td>9</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>185</td>
<td>65.5</td>
<td>34</td>
<td>36</td>
<td>69</td>
<td>75.4</td>
</tr>
<tr>
<td>10</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>185</td>
<td>73.4</td>
<td>40</td>
<td>35</td>
<td>69</td>
<td>96.0</td>
</tr>
<tr>
<td>11</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>199</td>
<td>59.5</td>
<td>28</td>
<td>38</td>
<td>70</td>
<td>83.1</td>
</tr>
<tr>
<td>12</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>190</td>
<td>61.6</td>
<td>33</td>
<td>39</td>
<td>68</td>
<td>89.3</td>
</tr>
<tr>
<td>13</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>13</td>
<td>13</td>
<td>165</td>
<td>53.4</td>
<td>37</td>
<td>40</td>
<td>79</td>
<td>69.3</td>
</tr>
<tr>
<td>14</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>14</td>
<td>14</td>
<td>178</td>
<td>15.0</td>
<td>39</td>
<td>42</td>
<td>82</td>
<td>44.2</td>
</tr>
<tr>
<td>15</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>15</td>
<td>15</td>
<td>165</td>
<td>30.6</td>
<td>42</td>
<td>44</td>
<td>83</td>
<td>58.0</td>
</tr>
<tr>
<td>16</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>16</td>
<td>16</td>
<td>196</td>
<td>38.8</td>
<td>36</td>
<td>43</td>
<td>75</td>
<td>81.7</td>
</tr>
<tr>
<td>17</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>17</td>
<td>17</td>
<td>149</td>
<td>49.2</td>
<td>32</td>
<td>45</td>
<td>72</td>
<td>88.8</td>
</tr>
<tr>
<td>18</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>18</td>
<td>18</td>
<td>177</td>
<td>17.9</td>
<td>28</td>
<td>55</td>
<td>78</td>
<td>80.8</td>
</tr>
<tr>
<td>19</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>19</td>
<td>19</td>
<td>163</td>
<td>18.9</td>
<td>31</td>
<td>50</td>
<td>78</td>
<td>85.3</td>
</tr>
<tr>
<td>20</td>
<td>south-coast</td>
<td>2000</td>
<td>1</td>
<td>20</td>
<td>20</td>
<td>182</td>
<td>18.3</td>
<td>23</td>
<td>52</td>
<td>72</td>
<td>90.5</td>
</tr>
</tbody>
</table>

RowID is the number of data entry counting from Jan 1, 2000 till Dec 31, 2012. basin is the air basin south coast in Los Angeles California; year is the year of the data entry; month is the month of the data entry; day is the day of the data entry; dayofyear is the day of the year of the data entry; AllCause75 is the daily number of mortality at Age 75 and above, excluding accidental deaths; PM25davg is the daily average level of PM$_{2.5}$ in microgram per cubic meter ($\mugm^{-3}$); o3 is the daily average level of ozone in parts per billion (ppb); tmin.0 is the daily minimum temperatures are recorded in Fahrenheit ($^oF$); tmax.0 is the daily maximum
temperatures are recorded in Fahrenheit (°F); MAXRH.0 is the daily maximum relative humidity level in percentages of the air-water mixture.

**Figure 3.3. Time Series Plots:** Daily Mortality above Age 75, PM$_{2.5}$, O$_3$, Min Temperature, Max Temperature, Max Relative Humidity

Figure 3.3 shows that almost all the time series have a long-term trend with the dominating seasonality; PM$_{2.5}$ has slightly weaker seasonality and the spikes are usually caused by wildfires. For mortality and ozone, it can be seen that the two time series are out of phase: higher ozone, lower mortality. This phenomenon is caused by the common long-term trend including seasonality in both time series, which needs to be adjusted. For mortality and PM$_{2.5}$, the following plots Figure 3.4 and 3.5 are presented along with the corresponding satellite pictures. The satellite picture on the left shows the wildfire and smoke on the incident day; the two time series plots on the right are the mortality above Age 75 and PM$_{2.5}$ over certain periods that cover the incident day. It appears that there is no spike in mortality associated with the spike in PM$_{2.5}$ within the relevant period of time,
however, statistical quantification and inference are required. More work will be done in the coming chapters.

Figure 3.4. Wildfire near Los Angeles on July 3, 2008: Satellite Picture (Left), Mortality above Age 75 (Right Top), and PM$_{2.5}$ (Right Bottom)

Figure 3.5. Wildfire near Los Angeles on August 31, 2009: Satellite Picture (Left), Mortality above Age 75 (Right Top), and PM$_{2.5}$ (Right Bottom)
Chapter 4

Intrinsic Detrending: Time Series Smoother for Effect Detection

4.1 Introduction

In environmental epidemiology, how the variations in the environment affect human health is of great interest. In particular, do daily fluctuations in air quality induce fluctuations in human mortality? The United States Environmental Protection Agency claims that small particulate matter, PM$_{2.5}$, can cause acute death and would enforce cleaner energy production by new regulations. Although their claim should be evaluated with some caution, the air quality problem not only affects long-term human health but also poses a major economic challenge to the United States and many other countries.

While the study of air quality is a broad question, whether a higher amount of ozone in the ambient air would associate with higher mortality is an important concern. The mortality data available are typically aggregated counts at different time scales. These count data display certain long-term trend including seasonality over time at any spatial location, see [Douglas and Rawles, 1999]. The long-term trend cannot be completely adjusted by the observed weather covariates. Besides, there are other issues discussed in [Bhaskaran et al., 2013], such as lag effects, i.e. what happens in today’s air quality can have an impact on the mortality tomorrow or the day after tomorrow.
To estimate the long-term trend with seasonality, smoothing methods are proposed, which are very often embedded in a semi-parametric time series Poisson regression model. In [Schwartz, 1994], the LOESS smooth functions of covariates including time was suggested in generalized additive Poisson models of air quality variables and human mortality. The smooth function of time is used to remove any long-term trend including seasonality in both the air quality and mortality time series. Thereafter, several alternative smoothing methods were proposed, such as parametric natural splines and penalized splines in [Dominici et al., 2002b], [Ramsay et al., 2003], [Schwartz et al., 2003], [Touloumi et al., 2004], [von Klot et al., 2005], [Duarte and Saraiva, 2003] and [Terzi and Cengiz, 2009]. The conclusions are drawn based on these embedded time series Poisson regression models and a significantly positive association between the air quality variable of interest and human mortality is often found. However, this result is subject to the degree of detrending in multiple time series data. Therefore, we intend to conduct a comprehensive investigation of the detrending process in order to understand and suggest how multiple time series with a long-term trend, including seasonality, should be detrended.

Human mortality, excluding accidental deaths, can be attributed to various reasons, from either latent variables or air quality fluctuations. It is crucial to separate the short-term signals from the long-term trend to study any short-term effect. However, the long-term trend including seasonality in either human mortality or an air quality variable is not precisely known. Overly rough functions of time can be too harsh on removing trends, even local fluctuations, and thus leave very little short-term signals in the deviations for the detection of short-term effect; overly smooth functions of time can be too lenient on removing trends and thus leave much long-term trend in the deviations, which overwhelms any short-term effect of human mortality and air quality. The current spline smoothing without examining the variability of smooth functions can be problematic in detrending multiple time series. It will be shown that by varying the smoothing parameters, any significant result, from negative to positive, can be obtained. This excessive modeling flexibility could undermine the plausibility of any air quality study.

In [Peng et al., 2006], natural spline smoothing and penalized spline smoothing were studied and model-based simulations showed that under moderate concurvity,
an analog of multicollinearity, smoother spline of air quality variable can lead to less confounding bias. In [Marra and Radice, 2010], an overview of generalized additive models based on the penalized likelihood approach with regression splines was provided. In [Wood et al., 2015], efficient smoothing parameter estimation was claimed to be performed and reduced rank spline smoothing methods were demonstrated for large datasets. In [Fang et al., 2016], Bayesian model averaging with generalized additive mixed models was proposed to address the modeling uncertainty. Nonetheless, different model selection criteria can lead to different smoothness; the smoothness of a certain spline is usually vaguely described and hardly defined.

To circumvent this problem, we propose robust and stable nonparametric smoothing methods that can be specified before analysis without undue experimentation and separate the long-term trend and the local short-term signals. After smoothing, the deviations, which are the raw time series subtracting the estimated long-term trend, are examined. These deviations contain the local information useful for investigating the short-term association between human mortality and air quality fluctuations, namely the acute effect.

For our case study, the California dataset is assembled and analyzed. This dataset contains the time series of daily human mortality, air quality and weather covariates in Los Angeles California, from 2000 to 2012. The mortality data were obtained from California Department of Public Health. The air quality data PM$_{2.5}$ and ozone were downloaded from California Environmental Protection Agency. The temperature data were downloaded from Carbon Dioxide Information Analysis Center of United States Historical Climatology Network and the relative humidity data were from the United States Environmental Protection Agency. For our simulation study, the synthetic datasets were generated from first decomposing the original time series into the nominal trend and detrended time series and then recomposing with our pre-specified correlation structure on the detrended time series, which makes that the synthetic data are almost indistinguishable from the real data.

There are three major contributions of our research study. First, it is found that the current spline smoothing methods can obtain different, both positive and negative, effects depending on how the smoothing parameter is varied. Second,
the proposed time series smoothers are shown to be robust and stable by the tight confidence intervals of correlation as well as partial correlation, and by the insensitivity of different factors via a factorial design of scenarios. Third, the characteristics of different solutions of time series detrending and association detection are determined by the sensitivity analysis in both our case and simulation studies.

The rest of this chapter is organized as follows. In Section 4.2, the formulation of our problem is described in detail. In Section 4.3, three classes of time series smoothers are proposed with a discussion on their properties. In Section 4.4, the air quality and mortality data in Los Angeles are analyzed by the proposed smoothers with sensitivity analysis, compared to cubic spline smoothing as a representative of spline smoothing methods. In Section 4.5, synthetic datasets are generated and the proposed smoothers show their capability of detecting even small acute effects, compared to cubic spline smoothing. In Section 4.6, the overall conclusion and general recommendations are made.

4.2 Problem Formulation

In air quality studies, mortality count data over time are very often encountered. The common model formulation is a generalized linear model with log link or Poisson regression, as suggested in [Schwartz, 1999].

\[
y_t \sim \text{Poisson}(\mu_t) \\
\log(\mu_t) = \alpha + \beta s_1(x_t) + \eta' s_2(z_t)
\]  

(1)

In Eq (1), \(y_t\) is the mortality count variable of certain aggregation level at Time \(t\). \(x_t\) is an air quality variable of interest at Time \(t\). \(z_t\) is a vector of measured covariates at Time \(t\). \(s_1(\cdot)\) and \(s_2(\cdot)\) are a spline smoothing estimator and a vector of such estimators, respectively.

Regarding the above formulation, there are two major concerns. First of all, the long-term trend including seasonality is dominating the overall pattern of the original time series data. Due to the common long-term trend of human mortality and air quality, no positive association between human mortality and air quality variable of interest can be found, even if the observed weather covariates are in-
cluded for adjustment. In fact, both Type I and Type II errors raise the concerns: claimed effects that are not real but artifacts of the smoothing method, and real effects that fail to be detected because of inappropriate smoothing methods. Second, Poisson regression model assumes that the observations are independent; however, in time series data, the observations close in time are more alike than those distant in time and therefore the autocorrelated structure violates the independence assumption.

To address both concerns, several spline smoothing methods were proposed for time series Poisson regression modeling. The reasoning here is that the air quality variable of interest and human mortality have different long-term trends that cannot be fully adjusted by the meteorological covariates. In [Peng et al., 2006], the refined model formulation becomes the following.

\[
y_t \sim \text{Poisson}(\mu_t) \\
\log(\mu_t) = \alpha + s_1(t, \lambda_1) + \beta x_t + \eta^T s(z_t) \\
x_t = s_2(t, \lambda_2) + \xi_t
\] (2)

In Eq (2), \(\xi_t\) can be viewed as the detrended air quality variable of interest at Time \(t\). \(s_1(\cdot), s_2(\cdot)\) and \(s(\cdot)\) are embedded spline smoothing estimators over time; \(s_1(\cdot)\) is the spline to control the long-term trend of \(y_t\) and \(s_2(\cdot)\) is for the air quality variable of interest. \(\lambda_1\) and \(\lambda_2\) are the parameters that control the smoothness of each corresponding spline estimator during the model fitting process. The choice of \(\lambda_1\) and \(\lambda_2\) can be selected via experimentation and result in rougher or smoother splines \(s_1(\cdot)\) and \(s_2(\cdot)\). This arbitrariness makes the resulting model ambiguous for further evaluation.

Hence, how to detrend the long-term trend including seasonality and extract these short-term signals becomes an important issue. We intend to propose a new collection of estimators with robustness and stability. Our model formulation is stated below.

\[
D^y_t = y_t - m_1(y^*_t) \\
D^x_t = x_t - m_2(x^*_t)
\] (3)

where \(y^*_t\) and \(x^*_t\) are sets of time series points before and after Time \(t\) excluding the
points in the neighborhood of Time $t$, which are used for the estimation of long-term central tendency at Time $t$; $m_1(\cdot)$ and $m_2(\cdot)$ are robust and stable estimators of the long-term trend to be defined. Thereafter, certain dependence measures or models are built on the detrended time series. For instance,

$$E(D^n_t | x_t, z_t) = \alpha + \beta D^x_t + \eta' z_t$$

(4)

The mortality count is viewed as a continuous variable in the above formulation; the Poisson distribution assumption can be restored if we add a positive constant integer to every detrended observation. This can be viewed as a two-stage model so that a researcher can first detrend multiple time series and then analyze the deviations. Both $y_t$ and $x_t$ are detrended respectively so the dominating long-term trend including seasonality is removed in the model fitting. Since $m_1(\cdot)$ and $m_2(\cdot)$ are adaptive based on the real data, the concurvity between $x_t$ and $z_t$, an analog of multicollinearity, can also be eliminated. In a two-stage model, the detrending should happen before Eq (4) is fitted so that there is no iteration back and forth between detrending and fitting.

More precisely, our problem is formulated to find robust and stable time series smoothers such that the long-term trend can be estimated without manipulation of tuning parameters and other observed covariates. The mathematical statement is the first stage of our model formulation. Note that in Eq (3), different $m_1(\cdot)$ and $m_2(\cdot)$ can result in different $D^n_t$ and $D^x_t$. Robustness and stability mean that no matter how the pre-selected parameters or centrality measures change, the cross-association between human mortality deviations and air quality deviations under certain dependence measure is approximately the same.

$$|\max(\gamma(D^n_t, D^x_t)) - \min(\gamma(D^n_t, D^x_t))| < c$$

(5)

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

The remaining deviations should not display any long-term pattern except heterogeneity, which is caused by the varying covariance structure of different time series. In other words, when the observations are in the non-volatile period, both $\text{Var}(D^n_t)$ and $\text{Var}(D^x_t)$ are small constants; when they are in the volatile period, $\text{Var}(D^n_t)$ and $\text{Var}(D^x_t)$ differ largely. The main purpose of detrending time series
is that the deviations contain little information about the long-term trend but all
the information of the shock events with a low level of noise. After detrending, the
deviations are still non-stationary, but only due to the shock of short-term events,
such as forest fires or human interventions, that cause the human mortality or air
quality to change drastically. Based on these mild assumptions, the time series
decomposition can be written below.

\[ x_t = l_t + \xi_t + \epsilon_t \]  

(6)

where \( l_t \) is the long-term trend with seasonality, \( \xi_t \) is the abnormal signal driven by
a short-term event and \( \epsilon_t \) is the independently and normally distributed random
error with mean 0 and variance \( \sigma^2 \), where \( \sigma \) is assumed to be much smaller than the
abnormal signals. During a non-volatile period, \( \xi_t \) is close to 0; during a volatile
period, \( \xi_t \) displays a certain short-term signal. The ultimate research aim is to
discover \( \xi_t \) by estimating and eliminating \( l_t \).

4.3 Proposed Methodology

To solve the formulated problem in Eq (3), three classes of methods utilizing a
moving window with a centered gap of varying width are hereby proposed.

The moving window includes the most useful information before and after the
time point of interest. The gap at the center can be helpful in retaining any local
abnormality in time series data. The width of the window and gap will be discussed
later. 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 a centered gap is approximately the same as the ordinary moving
window. The points in the gap are excluded to ensure that the short-term signals
do not contaminate the trend estimate. The equal number of points inside the
center-gapped window before and after the time of interest are used to estimate
the long-term trend; this ensures that equal amount of pre and post information
is taken. This type of window includes the points on both sides of the time of
interest to estimate the central tendency at the time point of interest and the
resulting estimate resembles an experimental control. After time series smoothing,
it is assumed that the detrended observations can behave as if independently with much less serial correlation and little long-term trend including seasonality left.

To elaborate further, the formulation in Eq (3) and (6) is utilized for justification. For simplicity, the window size \( 2k + 1 \) and gap size \( 2l + 1 \) are considered, and the average is selected as the smoothing instrument.

\[
D^*_t = x_t - \frac{1}{2(k - l)} (x_{t-k} + x_{t-k+1} + \cdots + x_{t-l-1} + x_{t+l+1} + \cdots + x_{t+k-1} + x_{t+k})
\]

\[
= l_t + \xi_t + \epsilon_t - \frac{1}{2(k - l)} (l_{t-k} + l_{t-k+1} + \cdots + l_{t-l-1} + l_{t+l+1} + \cdots + l_{t+k-1} + l_{t+k})
+ \xi_{t-k} + \xi_{t-k+1} + \cdots + \xi_{t-l-1} + \xi_{t+l+1} + \cdots + \xi_{t+k-1} + \xi_{t+k}
+ \epsilon_{t-k} + \epsilon_{t-k+1} + \cdots + \epsilon_{t-l-1} + \epsilon_{t+l+1} + \cdots + \epsilon_{t+k-1} + \epsilon_{t+k})
= [l_t - \frac{1}{2(k - l)} (l_{t-k} + l_{t-k+1} + \cdots + l_{t-l-1} + l_{t+l+1} + \cdots + l_{t+k-1} + l_{t+k})]
+ [\xi_t - \frac{1}{2(k - l)} (\xi_{t-k} + \xi_{t-k+1} + \cdots + \xi_{t-l-1} + \xi_{t+l+1} + \cdots + \xi_{t+k-1} + \xi_{t+k})]
+ [\epsilon_t - \frac{1}{2(k - l)} (\epsilon_{t-k} + \epsilon_{t-k+1} + \cdots + \epsilon_{t-l-1} + \epsilon_{t+l+1} + \cdots + \epsilon_{t+k-1} + \epsilon_{t+k})]
\]

Here, \( \{x^*_t\} \) is the set of time series points outside the neighborhood of Time \( t \), excluding Time \( t \) and the closest points to Time \( t \), to estimate the long-term central tendency at Time \( t \).

For the first term in the last sum, the average is the interpolation of \( l_t \), instead of extrapolation, since both the points before and after Time \( t \) are used. Compared to the extrapolation, the interpolation should be a good estimate for the long-term trend. In addition, the curvature of the long-term trend is very small and almost as that of a straight line in the nearest neighborhood of Time \( t \). Thus, the first term is close to 0.

For the second term, there are three cases discussed as follows. If \( t \) is the time of a short-term event, \( \xi_t \) should display a sharp spike while the average is 0 and therefore the second term exposes the shock signal. If \( t \) is not the time of a short-term event and the gapped window does not cover that time, \( \xi_t \) and the average are both 0 and therefore the second term is 0. If \( t \) is not the time of a short-term event but the gapped window covers that time, \( \xi_t \) is 0; the average slightly deviates from
0 but much smaller than the spike in absolute value. Moreover, this deviation can be reduced or removed by a trimmed mean which truncates a certain percentage of the largest and smallest values. Therefore, the second term slightly deviates from 0; there is a small pattern in this case that reveals a local abnormality, if moving average is used. However, under robust central tendency measures, for instance moving median, this small pattern will become close to 0.

For the third term, assuming that $\epsilon_t$ and $\{\epsilon_t^*\}$ are independent and identically normal distributed with mean 0 and relatively small variance $\sigma^2$, $E(\epsilon_t - \tilde{\epsilon}_t^*) = 0$ and $Var(\epsilon_t - \tilde{\epsilon}_t^*) = \sigma^2 + \frac{1}{\text{Window Size}} \frac{\text{Gap Size}}{\sigma^2}$. $\epsilon_t - \tilde{\epsilon}_t^*$ should be very small, compared to the shock signal $\xi_t$.

After summing up all the three terms, $D_t^x$ is $\xi_t$ when $t$ is the time of a short-term event; $D_t^x$ is near 0 when $t$ is not the time of a short-term event. This assures that the shock signal can be recovered.

In general, $x_t^*$ is a set of points in the neighborhood of Time $t$ with a gap centered at Time $t$.

$$D_t^x = x_t - m(x_t^*)$$
$$\approx l_t + \xi_t + \epsilon_t - m(l_t^*) - m(\xi_t^*) - m(\epsilon_t^*)$$
$$= (l_t - m(l_t^*)) + (\xi_t - m(\xi_t^*)) + (\epsilon_t - m(\epsilon_t^*))$$
$$\approx \xi_t$$

since $l_t - m(l_t^*) \approx 0$, $\xi_t - m(\xi_t^*) \approx \xi_t$ and $\epsilon_t - m(\epsilon_t^*)$ is negligibly small, compared to $\xi_t$. $D_t^x$ is a robust and stable estimator to $\xi_t$.

In summary, under the definition of time series decomposition,

$$x_t = l_t + \xi_t + \epsilon_t$$

where $x_t$ is the original time series, $l_t$ is the long-term trend with seasonality, $\xi_t$ is the abnormal signal driven by an event and $\epsilon_t$ is the independently and normally distributed random error with mean 0 and small variance $\sigma^2$. If $m(x_t^*) \approx l_t$,

$$D_t^x = x_t - m(x_t^*) \approx \xi_t$$

So far, the idea of defining the time series smoothers has been elaborated. These
smoothers can estimate and remove the long-term trend of multiple time series efficiently, regardless of the window or gap size. In the following, three classes of smoothers are proposed and each has its own merits. The first class is moving trimmed mean, where the trimmed mean is used within each window. It can change its local robustness, controlled by the trimming percentage, up to a researcher’s subject knowledge or preference. The second class is moving weighted mean, where the weighting scheme can also be determined by a researcher’s subject knowledge or preference. The third class is moving recursive weighted mean. It can re-weight the points within each moving window recursively in order to greedily detect abnormal signals. The re-weighting function is non-increasing and specified by a researcher, based on one’s subject knowledge or preference.

4.3.1 Moving Trimmed Mean

A trimmed mean is a statistical measure of central tendency which equals to the mean after discarding given parts of a sample at the upper and lower tail of the observed distribution, and typically discarding an equal amount of both. A trimmed mean with nonzero trimming percentage is less sensitive to outliers and can give a reasonable estimate of central tendency. In this regard, it is referred to as a robust estimator.

The trimming percentage, ranging from 0% to 50%, allows us to obtain a collection of estimators with different robustness. When the trimming percentage is 0%, no points are trimmed and the estimator is the arithmetic mean; when the trimming percentage is 50%, 50% of high and low points are trimmed and this results in the median.

Moving average is defined as the average of several days’ observations with a centered gap. Here, we assume that the window size is $2k + 1$ and the gap size is $2l + 1$.

$$MA(x^*_t) = \frac{1}{2(k - l)}(x_{t-k} + x_{t-k+1} + \cdots + x_{t-l+1} + x_{t+l+1} + \cdots + x_{t+k-1} + x_{t+k}) \quad (10)$$

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

$$MM(x^*_t) = \frac{1}{2}(x_{(k-l)} + x_{(k-l+1)}) \quad (11)$$
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 observations we take away from the center before and after the estimation time point \((l \geq 0)\) excluding \(x_t\). Thus, the total number of observations taken away is \(2l+1\) and the total number of observations left is \(2(k-l)\).

### 4.3.2 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 up to a researcher to determine how to design the weighting scheme. Two typical types of moving weighted mean are presented here, center-weighted mean and edge-weighted mean. Center-weighted mean puts more weight on the center outside the gap than on the edges symmetrically. It weighs more on the local region near the time point of interest. Edge-weighted mean puts more weight on the edges than on the center symmetrically. It weighs less on the local region but more on the long-term trend. Moving weighted mean is defined as below.

\[
WM(x_t^*) = w_{t-k}x_{t-k} + w_{t-k+1}x_{t-k+1} + \cdots + w_{t-l-1}x_{t-l-1} + w_{t-l+1}x_{t-l+1} \\
+ \cdots + w_{t+k-1}x_{t+k-1} + w_{t+k}x_{t+k}
\] (12)

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 each element of 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 \(t\).
4.3.3 Moving Recursive Weighted Mean

Recursiveness means that after the first set of deviations are obtained, more weight is assigned on small deviations and less weight is assigned on large deviations to form a new weighted mean and then we perform this procedure repeatedly until convergence. The idea is that those abnormal signals would stand out while those around the central tendency are almost zero. The detailed algorithm can be described as follows.

Step 1: Apply moving average, meaning equal weights, in the center-gapped window to obtain the estimated trend of time series.

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

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

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

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

There are infinitely many choices of the 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.

4.3.4 General Properties

In general, choosing the width of window and gap should incorporate two basic considerations. First, if the period of seasonality is roughly known, then the window size should be a fraction of the period. The literature in the case-crossover design, see [Carracedo-Martinez et al., 2010], of air quality study, typically use a window size from a few weeks up to a month. If certain shock events are roughly known, then the gap size should be approximately the shock width. Shocks are
usually certain extreme phenomena that immediately raise or reduce the level of ozone.

For the three classes of time series smoothers, there are both shared and individual properties. The shared properties come from the window and gap. When the window size increases, the trend estimated is smoother and the deviation will contain more long-term signal; when the gap size increases, the trend is also smoother and the deviation will contain more long-term signal. The individual properties come from the different measures of central tendency. Moving trimmed mean has widespread use. When the trimming percentage increases, the estimated trend becomes more robust. Moving weighted mean is more flexible, depending on the weighting scheme. It can accommodate a researcher’s subject knowledge and preference. Moving recursive weighted mean estimates a relatively robust central tendency and makes the abnormal deviations more prominent and the ordinary deviations near zero. In this sense, it should be 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 the moving average is statistically optimal ([Arce, 2005]). 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 estimate the central tendency with little influence by outliers. If the fluctuations are Laplace distributed, then moving median is statistically optimal ([Arce, 2005]).

The center-weighted moving average can be used when a researcher wants more weight on the observations near the time point of interest to obtain a better local or short-term fluctuation estimate. It weighs heavily on the local observations in forming a long-term estimate thus is more aggressive on the short-term signal removal. The edge-weighted moving average is used when a researcher wants more weight on the observations further away from the time point of interest to obtain a longer-term estimate. It weighs lightly on the local observations in forming a
long-term estimate thus less aggressive on the short-term signal 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 smoothed estimate follows the fluctuated curve closer. Hence, the deviations would moderately stand out. Moving recursive weighted mean with $e^{-|x|^2}$ has a weight 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 smoothed estimate follows the long-term trend closer. Hence, the deviations would be more prominent.

4.4 Case Study

4.4.1 Current Smoothing Methods

The current smoothing methods, such as spline smoothing, require the smoothing parameter tuning. Depending on how to choose the smoothing parameters, significant correlations can be induced, from negative to positive. For demonstration, cubic smoothing splines are utilized. Cubic smoothing splines perform a regularized regression over the natural spline basis, placing knots at all the points. They circumvent the problem of knot selection and simultaneously control for over-fitting by shrinking the coefficients of the estimated smooth function. More details can be seen in [Hastie and Tibshirani, 1990].

In Fig 4.1, the cubic smoothing spline changes from a straight line to the roughest curve by varying the smoothing parameter. The smoothing parameter is controlled by the option spar of smooth.spline() in R. The upper bound spar = 1.50 represents the smoothest while the lower bound spar = −1.50 represents the roughest. For simplicity, the smoothing parameters are set to be equal for both AllCause75 and O3.

In Fig 4.1, the red line in each plot is a cubic smoothing spline with a certain spar; spar changes from the highly smooth spar = 1.50, the moderately smooth spar = 0.60 to the highly rough spar = −1.50, from top to bottom. The deviations are obtained by using the observed values subtracting the spline values at each
time.

Figure 4.1. Raw Time Series and Smoothing Splines of Mortality and Ozone
The left three figures are the time series of mortality and the right three figures are the time series of ozone. The black dots represent the observed values and the red curves represent the smoothing splines. The smoothing parameters from top to bottom are set to be 1.5, 0.6, -1.5, representing highly smooth, moderately smooth and highly rough splines.
Table 4.1. Correlations and Partial Correlations with P-values between Mortality and Ozone

<table>
<thead>
<tr>
<th>Spar</th>
<th>Correlation</th>
<th>P-value</th>
<th>Partial Correlation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.50</td>
<td>-0.4313</td>
<td>≈ 0</td>
<td>-0.0835</td>
<td>≈ 0</td>
</tr>
<tr>
<td>0.71</td>
<td>-0.2665</td>
<td>≈ 0</td>
<td>-0.0010</td>
<td>0.9427</td>
</tr>
<tr>
<td>0.70</td>
<td>-0.2429</td>
<td>≈ 0</td>
<td>0.0081</td>
<td>0.5750</td>
</tr>
<tr>
<td>0.61</td>
<td>-0.0052</td>
<td>0.7203</td>
<td>0.0698</td>
<td>≈ 0</td>
</tr>
<tr>
<td>0.60</td>
<td>0.0140</td>
<td>0.3367</td>
<td>0.0717</td>
<td>≈ 0</td>
</tr>
<tr>
<td>-1.50</td>
<td>0.0725</td>
<td>≈ 0</td>
<td>0.0463</td>
<td>≈ 0</td>
</tr>
</tbody>
</table>

In Table 4.1, Spar is the smoothing parameter \( \text{spar} \) in \textit{smooth.spline()} and its values are chosen based on the sign changes of the resulting correlations. Correlation means Pearson correlation between deviations of mortality and deviations of ozone, while partial correlation means the partial correlation between deviations of mortality and deviations of ozone given the weather covariates temperature and relative humidity. As the long-term trend including seasonality gets eliminated more and more, both the correlation and partial correlations between mortality and ozone changes from significantly negative, insignificant and then to significantly positive. Hence, researchers have to find a proper way to justify their detrending method or find a method with better accuracy and precision before making a claim on any possible association between air quality and acute mortality.

4.4.2 Proposed Smoothing Methods

To study how the proposed smoothers change when the factors change, a factorial design of different scenarios is utilized. By experimenting with all representative combinations of analysis choices, the factorial design can give us a thorough sensitivity analysis. Table 4.2 gives a summary of the factorial design of scenarios; detailed explanations follow.

Table 4.2. Factors and Levels of Factorial Design of Scenarios

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>7, 15, 21, 29, 35, 43, 49 or 57</td>
</tr>
<tr>
<td>Gap</td>
<td>0, 1, 3, 5, 7, 9, 11 or 13</td>
</tr>
<tr>
<td>Measure</td>
<td>e.g. Trimming Percentage=0% or 50%</td>
</tr>
<tr>
<td>Type</td>
<td>DD, DR, RD, RR</td>
</tr>
</tbody>
</table>
For the moving window size, the factor Window is defined as 7, 15, 21, 29, 35, 43, 49 or 57 days. The actual meanings are half a week, one week and up to one month before and after the time of interest. For the gap size, the factor Gap is defined as 0, 1, 3, 5, 7, 9, 11 or 13 days. The actual meanings are 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 are simply used. The gap is always smaller than the window. For the measure, the factor Measure uses two distinctive levels on central tendency to study the effect of the pre-selected parameter or weighting scheme. For moving trimmed mean, the trimming percentages are selected to be 0% the average or 50% the median. There are infinite choices of the parameters or weighting schemes within the window. This can be decided by a researcher's subject knowledge or preference.

The factor Type is defined, where D and R stands for deviations and raw data, respectively, to investigate whether there is any effect between deviations of mortality and deviations of ozone by the abbreviation DD, between deviations of mortality and raw time series of ozone by the abbreviation DR, between raw time series of mortality and deviations of ozone by the abbreviation RD and between raw time series of mortality on raw time series of ozone by the abbreviation 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 analysis, a linear regression model is utilized and all main effects and two-way interactions are chosen. Volcano plots are utilized to summarize the effect size and significance.

For illustration, the results of moving trimmed mean are demonstrated. Moving weighted mean and moving recursive weighted mean can be analyzed in a similar fashion. For the factor Measure, moving average with trimming percentage = 0% and moving median with trimming percentage = 50% are selected.

Table 4.3 and 4.4 contain the detailed information by moving trimmed mean for Fig 4.2. The rows represent the predictors and the columns represent the coefficient estimates, standard errors, t statistics and 2-sided p-values of the linear model with main effects and interactions. The responses are correlation and par-
Figure 4.2. Volcano Plot of Regression Coefficient and P-values by Moving Trimmed Mean

The x-axis is the effect size measured by regression coefficient. The y-axis is the transformed p-value by $-\log_{10}()$; the larger transformed value, the more significance. The black dots represent each coefficient and the red lines are $y = -\log_{10}0.05$: any black dot above the red line means that the coefficient is nominally significant.

The predictors are window size, gap size, type and trimming percentage lambda specified in the factorial design of scenarios.

In Fig 4.2, Table 4.3 and 4.4, it is shown that only the effects with Type have both large size and significance; all other effects have small size although they are significant. 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 in the cross-association between mortality and ozone. The trimming percentage Lambda is insignificant. For the response measure partial correlation, respectively.
Figure 4.3. Boxplots of Association between Mortality and Ozone by Moving Trimmed Mean

The top two figures are the correlation and partial correlation between mortality and ozone; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}(\cdot)$: the larger transformed value, the more significance. The x-axis has three levels of the factor Type. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under different Window, Gap and Lambda.

correlation, we can obtain similar results. Thus, moving trimmed mean is robust and stable with respect to different Window, Gap and Lambda for a fixed Type.

In Fig 4.3, correlation and partial correlation across different types are summarized in boxplots. Table 4.5 shows the correlations and partial correlations with their p-values for each type when Window=21, Gap=5 and Trimming Percentage=50%.

Based on Fig 4.3 and Table 4.5, it can be seen that the correlations and partial correlations when Type=DD are positively significant. This means that the acute mortality has a nominally significant positive linear association with the sudden
Table 4.3. Summary of Correlation versus Variables and Interactions

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

Table 4.4. Summary of Partial Correlation versus Variables and Interactions

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

change in ozone, with or without meteorological covariates. When Type=DR, half of the correlations are insignificant while all the partial correlations are insignificant. This means that without meteorological covariates, the linear relationship between the acute mortality and ozone level is unclear; with meteorological co-
Table 4.5. Correlations and Partial Correlations with P-values by Type for Moving Trimmed Mean: Window=21, Gap=5, Trimming Percentage=50%

<table>
<thead>
<tr>
<th>Type</th>
<th>Correlation</th>
<th>P-value</th>
<th>Partial Correlation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>0.0869</td>
<td>≈ 0</td>
<td>0.0568</td>
<td>≈ 0</td>
</tr>
<tr>
<td>DR</td>
<td>0.0331</td>
<td>0.0227</td>
<td>-0.0283</td>
<td>0.0515</td>
</tr>
<tr>
<td>RD</td>
<td>0.0336</td>
<td>0.0207</td>
<td>0.1140</td>
<td>≈ 0</td>
</tr>
<tr>
<td>RR</td>
<td>-0.4338</td>
<td>≈ 0</td>
<td>-0.0956</td>
<td>≈ 0</td>
</tr>
</tbody>
</table>

variates, there is no significant linear relationship between the acute mortality and ozone level. When Type=RD, half of the correlations are significant while all the partial correlations are positively significant. This means that without meteorological covariates, the linear relationship between the mortality level and ozone change is unclear; with meteorological covariates, there is no significant linear relationship between the mortality level and ozone change. In addition, it is a researcher’s decision to examine the raw data or the deviations. Is it the jump in the air quality variable that has an acute effect on mortality or is it the absolute level? Depending on the researcher’s decision on Type, either positive significant effect or null effect can be deduced.

From the sensitivity analysis, Window and Gap can adjust the cross-association slightly; however, the sign of cross-association will not flip when Type is fixed. For moving trimmed mean, the small effect of Window and Gap implies the robustness to the different window and gap sizes, and the null effect of Lambda implies the stability.

### 4.4.3 Discussion & Recommendations

In this case illustration, it is discovered that the current spline smoothing methods can induce significant correlations, both negative and positive, depending on how the smoothing parameters are varied, see Table 4.1.

The framework of the proposed methods is to utilize a moving window with a central gap. Within each window, different central tendency measures can be selected, based on a researcher’s preference. As the window size increases, the estimated curve is smoother and the detrending is less severe; a stronger association will be found between deviations. As the gap size increases, more local signals are
removed from the window, the estimated curve is also smoother and the detrending is less severe; stronger association will be found between deviations. However, no matter how the window or gap size changes, within a wide range of days used by subject experts, the sign or significance of the cross-association for each type remains the same. This is the first important property, robustness to different window and gap sizes.

Within each moving window with a central 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 does not change. This is the second important property, stability to different central tendency measures.

For moving trimmed mean, a collection of trimmed means is utilized, changing from mean to median. Within each moving window, different trimmed means can reach statistical optimality by a goodness of fit criterion, depending on the underlying distributions. If a researcher does not have any subject knowledge or preference on the abnormal signals of time series data, moving average or median should work well.

For moving weighted mean, weighted arithmetic mean is utilized. Within each moving window, a researcher can decide how to assign different weights to the points based on one’s subject knowledge or preference on the abnormal signals of multiple time series data. If a researcher has insight into the nature of the data, moving weighted mean can be quite useful.

For moving recursive weighted mean, a certain re-weighting scheme is utilized in order to retain central tendency conservatively and enhance abnormal short-term signals. It is more greedy in searching for an anomaly. If a researcher emphasizes conserving the long-term central tendency and revealing more anomalies, moving recursive weighted mean can be a good choice.
4.5 Simulation Study

4.5.1 Simulation Setting

For our setup, the target cross-correlations between deviations of mortality and deviations of ozone are selected to be 0, 0.01, 0.02, 0.05 and 0.10, because it is generally thought that the acute effect between air quality and mortality is relatively small. In Section 4.4, the detected correlation is between 0.05 and 0.10 and therefore it is reasonable to set the correlation to be under 0.10. Including 0.01 and 0.02 can demonstrate how sensitive the proposed smoothers are in discovering abnormal short-term effect. Including 0 can demonstrate whether the proposed smoothers cause large spurious correlation. When the cross-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. Using the R package corpcor, the corresponding partial correlations are computed as -0.0290, -0.0182, -0.0074, 0.0250 and 0.0791.

Since the proposed smoothers give consistent results, the choice of Window and Gap size can be flexible. In our case, Window=21, meaning the extracted information are within 10 days before and after the time point of interest, and Gap=5, meaning the removal of the local signal are within 2 days before and after, are chosen, based on the literature of air quality studies. The number of simulations for each scenario is 100 and the total number of simulations is 1000.

In this simulation study, the idea is to generate synthetic datasets very close to the original dataset, namely indistinguishable. In our case, indistinguishable means that the measures correlation and partial correlation are identical on both the original and synthetic datasets. To maintain the original correlation structure, each time series is decomposed by LOESS into the linear, seasonal and remainder terms. LOESS is locally weighted regression developed by [Cleveland, 1979] and the seasonal-trend decomposition is further developed in [Cleveland et al., 1990]. 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 linear, seasonal and deviation components.

Step 2: Retain the actual means and variances of each time series of deviations. Set the correlation between deviations 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 linear and seasonal components so that the simulated data are generated.

In the following, the boxplots of the differences between the found correlation (or partial correlation) and the specified correlation (or partial correlation) across different pre-specified values of correlation (or partial correlation) are displayed and interpreted. These boxplots are expected to be narrowly centered at 0 so that the estimation is accurate and precise. Additionally, the boxplots of correlation (or partial correlation) and \(-log_{10}(p\text{-value})\) across different pre-specified values of correlation (or partial correlation) are displayed and interpreted. The correlations (or partial correlations) are expected to be consistently larger or smaller than 0 to show stability and the \(-log_{10}(p\text{-value})\) plot should be above the threshold \(-log(0.05)\) to show significance.

4.5.2 Simulation Comparison

4.5.2.1 Current Smoothing Methods

Given the cubic spline smoothing, generalized cross-validation can be adopted to select the smoothing parameter.

In Fig 4.4, it is seen that these boxplots are mostly around 0 and becomes closer to 0 when the pre-specified correlations value gets larger. Under multivariate normality, the current smoothing with generalized cross-validation (GCV) can identify the correct correlation and partial correlation with small errors. Generalized cross-validation is a common criterion for model selection, in this case,
spline tuning parameter selection, which is a computationally efficient alternative to cross-validation, see [Hastie and Tibshirani, 1990].

**Figure 4.4. Plots of Difference between Found and Specified in Correlation and Partial Correlation by Cubic Spline Smoothing Under GCV**

The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation after applying cubic spline smoothing under GCV. The red lines are \( y = 0 \). Each box plot is the graphical summary under simulations.

![Box plots showing the difference between found and specified correlations or partial correlations](image)

In Fig 4.5, both pre-specified correlation and partial correlation can be found consistently on the same side of 0; the significance can be detected when the pre-specified correlation is higher than 0.05.

Without the loss of generality, the highly rough \( S_p = -1.5 \) and the moderately smooth \( S_p = 0.60 \) are selected as the other smoothing scenarios.

In Fig 4.6, it can be seen that the estimated correlations or partial correlations display some systematic bias, i.e., the estimated correlations are uniformly smaller
Figure 4.5. Boxplots of Association between Mortality Deviations and Ozone Deviations by Cubic Spline Smoothing Under GCV
The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.

while the estimated partial correlations are uniformly larger than the pre-defined values. Also, the estimation variance is large.

In Fig 4.7, both correlation and partial correlation switch sides, from negative to positive. Statistical significance generally cannot be obtained except when the pre-defined correlation is 0.10, for either correlation or partial correlation.

4.5.2.2 Proposed Smoothing Methods
Following the above convention, moving trimmed mean is demonstrated. Moving average with trimming percentage = 0% and moving median with trimming
Figure 4.6. Plots of Difference between Found and Specified in Correlation and Partial Correlation by Cubic Spline Smoothing

The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation after applying cubic spline smoothing under the two smoothing scenarios. The red lines are $y = 0$. Each box plot is the graphical summary under simulations.

Percentage = 50% are used for simulation illustration.

In Fig 4.8, it can be seen that when the pre-specified correlation or partial correlation increases, the estimated correlation or partial correlation will be less biased. 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. In Fig 4.9, when the pre-specified correlation is 0, no statistical 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 did not obtain statistical significance. When the pre-specified correlation gets larger than 0.05, our estimate becomes more accurate.
Figure 4.7. Boxplots of Association between Mortality Deviations and Ozone Deviations by Cubic Spline Smoothing

The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}()$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.

Likewise, the partial correlation behaves similarly except that less significance can be found when the pre-specified cross-correlation is set small.

4.5.3 Discussion & Recommendations

Without a proper tuning parameter, the conventional approach of spline smoothing can change 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, typically over 0.05, and
Figure 4.8. Plots of Difference between Found and Specified in Correlation and Partial Correlation by Moving Trimmed Mean

The x-axis has the five levels of specified correlations or partial correlations between the deviations of mortality and the deviations of ozone. The y-axis is the found - specified correlation or partial correlation under the two trimming percentages. The red lines are \( y = 0 \). Each box plot is the graphical summary under simulations.

The accuracy and precision of detection are improved when the actual association increases, given that the sample size is fixed.

If the long-term trend can be estimated more accurately, abnormal short-term signals are better recovered as well. One possible improvement is to adjust the window and gap size adaptively. For instance, we can enlarge the window and gap size during a non-volatile period and shrink their sizes during a volatile period. Another possible improvement is ideally to choose the most appropriate measure within each moving window with a central gap. Assuming that the detrended time series are weakly stationary, there is a certain measure of central tendency that can achieve the optimality within the moving window. For instance, when the
Figure 4.9. Boxplots of Association between Mortality Deviations and Ozone Deviations by Moving Trimmed Mean

The top two figures are the correlation and partial correlation between mortality deviations and ozone deviations; the bottom two figures are their corresponding transformed p-value plots by $-\log_{10}(\cdot)$: the larger transformed value, the more significance. The x-axis has the five levels of specified correlations and partial correlations. The top two red lines are $y = 0$; the bottom two red lines are $y = -\log_{10}0.05$. Each box plot is the graphical summary under simulations.

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

The general recommendations can be stated as follows. First of all, the window and gap size should be chosen based on the scientific background or the literature. Ideally, they should be adaptive to capture the long-term trend. If no other information is available, a researcher can simply choose any pair of window and gap that fits the subject knowledge. Due to the robustness of the proposed smoothers, different window and gap sizes should always detect the cross-association reasonably well. Secondly, if a researcher has no subject knowledge or preference on the moving measure, moving trimmed mean should be used; the trimming percentage can be decided by the researcher, depending on how one wants to truncate for the central tendency within each window. If a researcher has some subject knowledge or preference on the moving measure, moving weighted mean should be used; the weighting scheme is up to the researcher, depending on the relevance of each data point to the time point of interest. If a researcher wants to focus on finding abnormal short-term signals, moving recursive weighted mean can be used; the re-weighting scheme is also up to the researcher’s choice, depending on how flexible a researcher wants for the long-term central tendency.

4.6 Conclusions and Discussion

First of all, the major formulations of smoothing in air quality studies are reviewed. A potential pitfall has been found: the prevailing spline smoothing methods can obtain any significant result, from negative to positive, by varying the smoothing parameters. This can undermine the validity of the discovered relationship between air quality and acute death. In general, the de-trending method is devised to estimate the deviation of the daily value from the trend by refitting the long-term trend value. As it turns out, the researchers in air quality studies have been quite versatile in how they fit the long-term trend, with little explanation or justification. That flexibility can lead to different answers, which is one of the novel results of our research.

Second, robust and stable time series smoothers are proposed. Each smoothed estimate can serve as an experimental control at the time point of interest. There are three classes of robust time series smoothers: moving trimmed mean, mov-
ing weighted mean and moving recursive mean. The strengths of the proposed methods are the robustness to different window/gap sizes and the stability to the different tuning of the central tendency measure. One limitation is the inherent difficulty of knowing the correct separation between long-term trend and short-term fluctuation. Spline smoothing can obtain a wider range of smoothed curves, at the expense of moving the fluctuations into the trend or vice versa, while the proposed method results in less variation by the different window and gap sizes. Another limitation is that the proposed method could be computationally heavy. The moving window with a central gap needs to slide point by point to calculate the long-term trend estimate; it is time-consuming to go through all the points from the beginning to the end. It would be more computationally expensive if we attempt to optimize the window and gap sizes. All of these moving window methods can generate robust and stable detrended time series.

Third, the characteristics and sensitivities of time series smoothers are examined using a factorial design of scenarios in our case illustration and indistinguishable synthetic datasets in our simulation. Choice of data type, raw data or deviations, has a dramatic effect on results. Within each data type, the proposed smoothers have the capacity of obtaining the correct cross-association with little bias and small variation.

Last but not least, although the robustness and stability of our proposed smoothers can assure the result’s consistency, it is still recommended to choose the window and gap size using domain knowledge, rather than experimenting with the dataset at issue. The window size should be a fraction of the period of the long-term trend with seasonality and the gap size should be the average length of the shock period. Within each moving window, a researcher can tailor-make the time series smoother, according to the data characteristics. A researcher should use moving trimmed mean if no extra information is available about the time series distribution. Otherwise, one should consider using moving weighted mean with an empirical weighting scheme that addresses the specific distribution. If one wants to estimate the central tendency conservatively and retain more short-term abnormalities, moving recursive weighted mean can be adopted. In a nutshell, the proposed time series smoothers provide us with a reliable and efficient collection of statistical tools that do not require fine-tuning of smoothing parameters and
produce reliable results with little bias and small variation.
5.1 Introduction

There are two major steps in finding the cross-association: detrending and detecting. Our main focus is on how to detrend a single time series intrinsically without any ad hoc judgment. The main property that we would like to ensure is that the detrended time series contain no or little information about the long-term trend but most information of the shock events with a low level of noise. After detrending, the deviations are still non-stationary, but only due to the shock of short-term events, such as forest fires or human interventions, that potentially cause the human mortality or air quality to change. Based on these mild assumptions, the time series decomposition can be written below.

\[ x_t = l_t + \xi_t + \epsilon_t \]  

(6)

where \( l_t \) is the long-term trend with seasonality, \( \xi_t \) is the unusual signal driven by a short-term event and \( \epsilon_t \) is the independently and normally distributed random error with mean 0 and variance \( \sigma^2 \), where \( \sigma \) is assumed to be much smaller than the unusual signals. During a non-volatile period, \( \xi_t \) is close to 0; during a volatile period, \( \xi_t \) displays a certain short-term signal or spike. The ultimate research aim is to recover \( \xi_t \) by estimating and eliminating \( l_t \).
One of the major research issues is that $l_t$ and $\xi_t$ are unidentifiable without an exact definition of the long-term trend and short-term changes. In other words, based on the parametrization of the model, $l_t$ and $s_t$ can essentially be collapsed to one term if the long-term trend and short-term changes are not mathematically defined. Our definition uses the time period as a threshold for the long-term and short-term in both time and frequency domain; however, the exact form of the long-term trend is still unspecified. In general, the long-term trend should be adaptive over time so it is very challenging to find a proper definition. Therefore, validity and stability are examined in the following sections, instead of bias and variability.

5.2 Validity

Validity is the extent to which a method or model is well-founded and likely corresponds accurately to the real world based on probability. The validity of a statistical tool is considered to be the degree of probability to which the tool measures what it claims to measure; in this case, the validity is an equivalent to a percent of how accurately the model corresponds to reality. In our case, the proposed method is valid since it is directly targeted to the short-term changes; additionally, the illustrative example shows that the proposed methods work better than smoothing splines or kernel smoothing under the general criterion.

Our research objective is to extract the short-term signals while eliminating the long-term trend. It is not trying to minimize the errors of certain form but to reveal large deviations from the long-term trend, unlike the typical spline or kernel smoothing. Here is an illustrative example to show the validity and novelty.

In Figure 5.1, the short-term signal is expressed as a spike in the middle of the time series. The long-term trend is assumed to be a horizontal line. The ideal situation is to fit a horizontal line so that after taking the deviations, the short-term signals are completely recovered. The red line is a fitted cubic smoothing spline whose smoothing parameter is chosen under generalized cross-validation. This curve follows the spike so well that after taking the deviations, the detrended time series are close to zero. Similarly, the green line is a fitted Nadaraya-Watson kernel regression line with the Gaussian kernel whose bandwidth is chosen under
generalized cross-validation. This curve fits even better so that the detrended time series are almost zero. The blue line is the moving gapped window smoothing using the mean with Window=27 and Gap=7. It is clear that the smoothed line follows the long-term trend so that the spike is retained.

In summary, the detrending objective is to retain the short-term fluctuations but not to find a trade-off between minimizing the errors of certain form and the number of modeling parameters. In this demonstration, it has been shown that the proposed method is designed to retain the short-term signals and it is valid rather than the methods designed for the smoothing purpose. More characteristics of the proposed methods have been discussed in Chapter 4, especially for the effect of different factors such as window and gap.
5.3 Stability

A measure is said to have a high stability if it produces similar results under consistent conditions. The proposed method, under different simulation scenarios, can identify the long-term trend reasonably well while the smoothing spline gets much more varied results; even for the same time series, if we cut them into several different pieces, the resulted smoothing splines will be different. The proposed method is more stable than the smoothing spline.

The long-term trend is defined as any observable movement or frequency that occurs over a significant period of time, for example, a month, usually in cycles. The short-term change is defined as any observable movement or frequency that occurs under that a significant period of time, typically driven by a sudden event. For demonstration, the long-term trend is set as a cosine curve from 0 to $8\pi$. The random error is set as identically normally distributed with mean 0 and standard deviation 0.5. The time of short-term changes follows a Poisson process with $\lambda = 10$ and each fluctuation last for 7 days. The comparison criteria are the root mean square error (RMSE) and mean absolute error (MAE) between the estimated and the actual short-term changes. For the selection criteria of smoothing splines, leave-one-out CV and Generalized CV are adopted; there is no specification on the errors’ distribution so both AIC and BIC cannot be used.

The stability to the shape and height of short-term signals is examined via numerical simulations. For the shape, the re-ordered uniform distribution is constructed to be right-skewed spikes. This type of shape is typical because, for either air quality indicators or human mortality, there is a tendency that they increase to a high level quickly and decrease slowly. For the height, the lognormal distribution is utilized. This type of height is considered because the more heavy-tailed distributions can produce synthetic data that is much less like the air quality or human mortality data we have seen. The detailed specifics and results are shown below.

In the normal distribution, the time series with long-term trends and short-term changes are simulated as follows. The long-term trend is set as a cosine curve from 0 to $8\pi$. The time of short-term fluctuations follows a Poisson process with $\lambda = 10$ and each fluctuation last for 7 days with a normal distribution. Note
that the 7-day short-term changes are on both sides of the long-term trend so symmetric.

**Figure 5.2. Simulation with Proposed Method and Smoothing Spline Fitting:** Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=91, G=11)

In Figure 5.2, the red line represents the fitted smoothing spline under generalized cross-validation. It can be seen that it fits the cosine curve well and smoothly. The blue line represents the proposed fitting method with randomly chosen Window=91 and Gap=11. It is clear that the proposed method is slightly wiggly. In this case, without a principled parameter selection method, we cannot find a good proper window and gap size that can outperform the smoothing spline.

**Table 5.1. Numerical Comparison on root mean squared error and mean absolute error:** moving gapped mean(Random: W=91, G=11), smoothing spline(CV) and smoothing spline(GCV)

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving Gapped Mean</td>
<td>0.5155(0.0685)</td>
<td>0.4121(0.0699)</td>
</tr>
<tr>
<td>Smoothing Spline(CV)</td>
<td>0.5060(0.0688)</td>
<td>0.4045(0.0702)</td>
</tr>
<tr>
<td>Smoothing Spline(GCV)</td>
<td>0.5060(0.0688)</td>
<td>0.4045(0.0702)</td>
</tr>
</tbody>
</table>
In Table 5.1, the root mean square error 

\[ RSME = \sqrt{\frac{1}{n} \sum_{t=1}^{T} (s_t - \hat{s}_t)^2} \]

and mean absolute error 

\[ MAE = \frac{1}{n} \sum_{t=1}^{T} |s_t - \hat{s}_t| \]

of the proposed method are uniformly less than those of the smoothing spline. This shows that the proposed method, given a moderately good case, can recover short-term signals better than the smoothing spline, given its best case under generalized cross validation.

In the re-ordered uniform distribution, the time series with long-term trends and short-term changes are simulated as follows. The long-term trend is set as a cosine curve from 0 to 8π. The time of short-term fluctuations follows a Poisson process with \( \lambda = 10 \) and each fluctuation last for 7 days with a re-ordered uniform distribution, which forms a right-skewed spike. Note that the spike is on one side of the long-term trend so asymmetric.

In Figure 5.3, the red line represents the fitted smoothing spline under generalized cross-validation. It can be seen that it is so wiggly that it does not fit the cosine curve well. The blue line represents the proposed fitting method with randomly chosen Window=31 and Gap=3. It is not so clear that the proposed method fits the cosine curve better. However, as a matter of fact, most scenarios of the combinations of window and gap result in smaller root mean square error and mean absolute error.

Table 5.2. Numerical Comparison on root mean squared error and mean absolute error: moving gapped mean(Random: W=31, G=3), smoothing spline(CV) and smoothing spline(GCV)

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving Gapped Mean</td>
<td>0.6508(0.0526)</td>
<td>0.5215(0.0528)</td>
</tr>
<tr>
<td>Smoothing Spline(CV)</td>
<td>0.7406(0.0519)</td>
<td>0.5993(0.0516)</td>
</tr>
<tr>
<td>Smoothing Spline(GCV)</td>
<td>0.7394(0.0519)</td>
<td>0.5983(0.0516)</td>
</tr>
</tbody>
</table>

In Table 5.2, the root mean square error and mean absolute error of the proposed method are uniformly less than those of the smoothing spline. This shows
Figure 5.3. Simulation with Proposed Method and Smoothing Spline Fitting: Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=31, G=3)

that the proposed method, given a moderately good case, can recover short-term signals better than the smoothing spline, given its best case.

In the log-normal distribution, the time series with long-term trends and short-term changes are simulated as follows. The long-term trend is set as a cosine curve from 0 to $8\pi$. The time of short-term fluctuations follows a Poisson process with $\lambda = 10$ and each fluctuation last for 7 days with a lognormal distribution, either upward or downward. Due to severe asymmetry, the spline smoothing performs even worse, in terms of root mean square error and mean absolute error. The following is the fitting when the spikes are lognormal(0, 0.5^2).

In Figure 5.4, the proposed method with randomly chosen Window=57 and Gap=15 fits closer to the cosine curve so it can recover more short-term signals.

In Table 5.3, the root mean square error and mean absolute error of the proposed method with randomly chosen Window=57 and Gap=15 are uniformly less than those of the smoothing spline under generalized cross validation. This shows that the proposed method, given a moderately good case, can recover short-term
Figure 5.4. Simulation with Proposed Method and Smoothing Spline Fitting: Red = Smoothing Spline(GCV), Blue = Moving Gapped Mean(Random: W=57, G=15)

Table 5.3. Numerical Comparison on root mean squared error and mean absolute error: moving gapped mean(Random: W=57, G=15), smoothing spline(CV) and smoothing spline(GCV)

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moving Gapped Mean</td>
<td>0.6273(0.0466)</td>
<td>0.5017(0.0468)</td>
</tr>
<tr>
<td>Smoothing Spline(CV)</td>
<td>0.7896(0.0458)</td>
<td>0.6405(0.0451)</td>
</tr>
<tr>
<td>Smoothing Spline(GCV)</td>
<td>0.7883(0.0458)</td>
<td>0.6394(0.0451)</td>
</tr>
</tbody>
</table>

signals better than the smoothing spline, given its best case. Note that compared to the less heavy-tailed distribution, both methods have a larger root mean square error and mean absolute error. However, the proposed methods are less affected, in terms of the change of root mean square error and mean absolute error.

5.4 Window-Gap Selection

For effective detrending, the following method is proposed to find a proper window and gap size. This procedure is based on the necessary conditions of the detrended
time series: there are no significant long-term cycles of a certain frequency and no long-term autocorrelation/partial autocorrelations.

The properties that can be ensured are stated below. First, there are no significant long-term trend or cycles, which is examined by ACF (autocorrelation function). Second, the detrended time series does not display any observable pattern of ACF and PACF (partial autocorrelation function), which is examined by ACF and PACF plots. Third, if the window size is further reduced, the detrended time series will display a more observable pattern in both ACF and PACF plots.

Note that the central tendency measure is decided by the literature of the subject of study, before applying the moving center-gapped window. If there is no previous literature, the researcher can choose the mean as the default and then median as the alternative.

1. Plot the time series and find the most observable cycle. To be more precise, use the periodogram of the time series to identify the most dominant period (or frequency). Use the period of the found cycle as the initial window size. The recommended initial window size is one-fourth of the most dominant period. The initial gap size is set to be 0.

2. Detrend the time series by the proposed method. Plot ACF and PACF with a sufficiently large lag to display the long-term patterns and examine the detrended time series.

3. If no observable pattern of a sine or cosine curve is found in the ACF plot and no alternating positive and negative lines in the PACF plot, this window size can be accepted. If not, reduce the window size by half and then repeat the previous steps until this happens.

4. After window size, the coefficient of variation is used to select the gap size. The preferred gap size would be the maximum duration of the short-term change. The larger the coefficient of variation is for the detrended time series, the more signals the detrended time series retain. Thus, the gap size that yields the largest coefficient of variation should be selected.

For illustration, the time series of ozone is used as an example of how to implement this graphical selection procedure.
Step 1: Figure 5.5 is the plot of daily ozone time series from Year 2000 to 2012. The largest cycle’s period is approximately a year, which is the seasonal cycle of ozone oscillation. Therefore, 365 is chosen as the initial window size, although a fourth of this period is preferred to be used as a rule of thumb.

![Ozone Time Series](image)


Step 2: Detrend the ozone time series with Window=365 and Gap=0. Examine the ACF and PACF of the detrended time series.

In Figure 5.6, it can be seen that the ACF plot shows an observable pattern. In addition, the pattern is cyclical of being positive, negative, positive and then negative again. Therefore, a smaller window size is desirable.

Step 3: Detrend the ozone time series with Window=183 and Gap=0. Examine the ACF and PACF of the detrended time series.

In Figure 5.7, the cyclical pattern is weakened but still quite dominant. Further window size reduction is needed.

In Figure 5.8, The ACF plot has a cyclical pattern within 100 days. The PACF plot also displays a cyclical pattern of being positive, negative, positive and then negative again. Further reduction is needed.
Figure 5.6. The ACF and PACF of Detrended Ozone Daily Time Series: Window=365 and Gap=0

Figure 5.7. The ACF and PACF of Detrended Ozone Daily Time Series: Window=183 and Gap=0

In Figure 5.9 and 5.10, based on the minimal pattern of ACF and PACF, the window size is set to be 43. The set of rules for the minimal pattern are as follows.
Figure 5.8. The ACF and PACF of Detrended Ozone Daily Time Series: Window=91 and Gap=0

Figure 5.9. The ACF and PACF of Detrended Ozone Daily Time Series: Window=43 and Gap=0

- For the ACF plot, the sine or cosine pattern is largely reduced or even removed. If the window size is further reduced, the sine or cosine pattern
becomes strengthened and more observable with a smaller cycle, compared to the original one.

- For the PACF plot, the alternating pattern of being positive and negative is largely reduced or removed; the exponential pattern is relatively small. If the window size is further reduced, the exponential pattern becomes quite observable.

Step 4: Define the coefficient of variation as

$$CV = \frac{\sigma(|d(y_t)|)}{\mu(|d(y_t)|)},$$

where $d(y_t)$ is the detrended time series. Figure 5.11 is the coefficient of variation for detrended ozone daily time series for different gaps.

In Figure 5.11, Gap seems not to provide much help in preserving the short-term changes. One probable reason is that the time series is so long and short-term spikes are so sparse; the effect of the gap only applies to those very sparse cases. However, Gap=7 gives us the largest CV and preserve more variation in the
detrended time series. If Gap is selected, Gap should be 7 days, which means that 3 days before and after the day of interest are removed.

5.5 Detrending Diagnosis

For detrending diagnosis, our research objective is to ensure that there is no long-term pattern in the detrended time series. After detrending, it is desirable to know whether the long-term trend and short-term changes are indeed separated; thus, the diagnostic procedures are proposed. In the time domain, we look at autocorrelation and partial autocorrelation function plots; in the frequency domain, we look at periodogram. A periodogram is used to identify the dominant periods (or frequencies) of a time series. This can be a helpful tool for identifying the dominant cyclical behavior in a time series, particularly when the cycles are not related to the commonly encountered monthly or quarterly seasonality.

To refresh the definitions, the long-term trend is defined as any observable movement or frequency that occurs over a significant period of time, for example, a month. Short-term fluctuation is defined as any observable movement or frequency
that occurs under that a significant period of time. For the time periods, there is no absolute setting rule. Researchers can do sensitivity analysis or cross-validation to select an appropriate time period.

The following is the illustration of the detrending diagnosis on ozone.

![Graph showing auto correlation and partial auto correlation for detrended ozone with moving gapped mean: Window=43, Gap=7](image)

**Figure 5.12. Auto Correlation and Partial Auto Correlation for Detrended Ozone with Moving Gapped Mean: Window=43, Gap=7**

In Figure 5.12, the patterns of ACF and PACF are minimized; in Figure 5.13, the periods of spectrums of detrended ozone time series (> 2000) are less than 50. This diagnosis shows that this detrending gives us an acceptable trade-off.

In conclusion of the ozone detrending, the window size is selected to be 43 days and the gap size is 7 days. The window ensures that the detrended time series do not have any long-term trend or incur any new pattern. The gap ensures that within this window, the detrended time series has the largest variation. In the detrending diagnosis, no long-term patterns or cycles are shown so that this detrending is acceptable.
5.6 Conclusion and Discussion

In this chapter, the properties of the proposed methods are elaborated. Firstly, the proposed method is designed to recover the short-term changes; the validity was illustrated, compared to spline and kernel methods. Second, the stability is carefully demonstrated. Three representative scenarios indicate its stability. In the normal case of short-term changes, the performance of the proposed method is acceptable but not as good as smoothing spline. However, when the shape and height of the short-term changes vary, the proposed method outperforms smoothing splines and can stably estimate the long-term trend. The numerical results of RMSE and MAE are improved, even without the proper selection of window and gap sizes. In addition, because of the construction of the moving gapped window, the numerical results will not change when the time series is shortened or lengthened, which is not the same case for smoothing splines via shrinkage.

The window-gap selection method is also proposed and illustrated step-by-step. Basically, the selection procedure provides us a trade-off between under-smoothing and over-smoothing. Although the graphical selection procedure requires manual iterations, it can be extended to automatic selection by defining a proper criterion.
However, due to the complexity of time series data, such a criterion or loss function is not easy to find; moreover, it can vary drastically when the covariance structure of multiple time series changes over time.

Finally, the detrending diagnosis is derived, based on the window-gap selection procedure. It is very useful in helping us evaluate if a certain detrending result is acceptable or not.
Chapter 6

Association Detecting: Indicators of Air Quality and Acute Deaths in California

6.1 Introduction

A recent paper by [Schwartz et al., 2017] points to many time series studies that support an association of air quality and acute mortality. In a meta-analysis of myocardial infarct triggers, 14 studies support of the claim that “air pollution is an important trigger of myocardial infarction” ([Nawrot et al., 2011]), and in another meta-analysis ([Mustafić et al., 2012]), another six studies are given, again making the same claim. In contrast, other studies make the case that when potential biases are taken into account, there is no association between air quality and deaths ([Chay et al., 2003]; [Cox Jr et al., 2013]; [Enstrom, 2005]; [Greven et al., 2011]; [Janes et al., 2007]; [Wang et al., 2015]; [Yang et al., 2004]; [Zu et al., 2016]). There are studies on both sides of the question, “Is air quality causal of acute human deaths?” The weight of evidence is on the side of a positive association, but for any claim to be considered causal, it takes only one valid, negative association study to negate the causality claim.

It is becoming clear that many scientific claims are failing to replicate ([Begley and Ellis, 2012]; [Ioannidis, 2005]; [Young and Karr, 2011]).
Failure to replicate appears to be over many different scientific areas: psychology, epidemiology, experimental biology, physics, astrophysics, etc. The replication problem is old ([Mayes et al., 1988]); it has recently gained prominence as people wonder about its extent and what might be done to solve it. A survey of scientists ([Baker, 2016]) reports that 90% of scientist think there is a crisis in reproducibility (52% a significant crisis and 38% a slight crisis). Environmental epidemiology is likely subject to the same problems in other areas of science.

One factor that exacerbates the examination of reproducibility is the general lack of access to data. [Cecil and Griffin, 1985] note: As an abstract principle, the sharing of research data is a noble goal and meets with little opposition. However, when data sharing is attempted in a particular circumstance, the conflicting interests of the parties can thwart the exchange. A glance at the benefits and obstacles to data sharing reveals the reason: few of the benefits and most of the burdens fall to the possessor of a dataset. For example, the owner of the data might consider it burdensome to defend a claim. Finally, many authors do not attempt to report negative results as editors, typically, are much less likely to accept a negative study, so publication bias is an expected result.

We have three goals with this research. Our first objective is to analyze a dataset for California relating mortality to levels of ozone and PM$_{2.5}$. The American Lung Association regards Ozone and PM$_{2.5}$ as the most serious air quality health risks and notes that California has 7/10 US cities with poorest air quality (http://www.lung.org/assets/documents/healthy-air/state-of-the-air/state-of-the-air-2017.pdf). We use two independent analysis approaches, a case-crossover analysis, CCO ([Figueiras et al., 2005]), and a standard time series regression analysis, TSR. We give odds ratios along with confidence limits and p-values, raw and adjusted, for each method. Our second objective is to present a straightforward, stepwise regression analysis of the dataset. Here, we analyze three categories of mortality, two air quality variables, with time lags of 0 or 1 day, and eight air basin subsets (defined by geographical air basins in California, see in Chapter 3), https://www.arb.ca.gov/ei/maps/statemap/abmap.htm, for a total of 96 separate analyses. Our third objective is to provide the analysis dataset so that others can replicate our results as well as try different analysis strategies.
6.2 Methods

6.2.1 Description

For mortality, the state of California provides access to “death public use files” for research. The cause of deaths is indicated by an ICD 10 code. We coded three mortality categories: AllCause, Cardiovascular, and Respiratory (AllCause65, CV65, and Resp65 respectively). In all cases, accidental deaths were excluded. Deaths of individuals above Age 65 and older were included in the regression analysis; AllCause75 was used in the case-crossover analysis. Each type of deaths was aggregated to the day and year within each air basin. A plot of daily deaths versus DayOfYear is given in Fig. 6.1 (a); the four years are overprinted. The residuals from a model including DOY and DOY2 are displayed in Fig. 6.1 (b).

\[
\text{Mortality} = b_0 + b_1 \times \text{DOY} + b_2 \times \text{DOY2}
\]

Deviations from this model are considered seasonally adjusted since the original time series have a quadratic trend every year.

The analysis dataset used in this study, years 2004 - 2007, eight most populous air basins, can be obtained from [Young, 2017].

6.2.2 Statistical Methods

For all air basins, the daily mortality data was complete. For three of the air basins, all data was present, Sacramento Valley, San Joaquin, and South Coast. To facilitate using the using same analysis on each of the eight air basins, we imputed missing data using JMP SVD method (JMP, 2016a). Table 6.1 gives the number of imputations for the other air basins.

6.2.2.1 Case-crossover analysis

Case-crossover is a standard way to evaluate the health-related effects of air quality when the data is in the form of a time series ([Figueiras et al., 2005]; [Bateson and Schwartz, 1999]). For each time point, time stratified symmetric bidirectional case-crossover is proposed as a new way of selecting control periods,
Figure 6.1. South Coast Air Basin.
(a) Daily deaths versus Day of Year, DOY, for the years 2004-2007. Four years are overprinted.
(b) Deviations from Day of Year time series smoother, linear and quadratic model.
Table 6.1. Numbers of data values imputed in the analysis dataset.
PM25davg is the daily average PM$_{2.5}$ level. Tmin and tmax are the minimum and maximum temperature. MAXRH is the maximum relative humidity.

<table>
<thead>
<tr>
<th>Air Basin</th>
<th>Mountain</th>
<th>Salton</th>
<th>San Diego</th>
<th>San Francisco</th>
<th>South Central</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM25davg</td>
<td>26</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tmin</td>
<td>0</td>
<td>93</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tmax</td>
<td>0</td>
<td>94</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MAXRH</td>
<td>325</td>
<td>3</td>
<td>0</td>
<td>267</td>
<td>0</td>
</tr>
</tbody>
</table>

with a multiple day control period being chosen a week, two weeks, and up to four weeks before and after the event day of interest. We use eight days as controls in total. As the comparison is within a narrow time window, other factors such as age distribution, gender distribution, etc. are also controlled by the nature of the design. In our case, the outcomes are daily mortality, and the predictors are air quality measures (PM$_{2.5}$ and ozone), and weather variables. We use a Cox Proportional Hazard model:

\[
\text{logit}(p) = b_0 + b_1 \times \text{tmin}.0 + b_2 \times \text{tmax}.0 + b_3 \times \text{MAXRH}.0 + b_4 \times (\text{PM}_{2.5} \text{ or ozone})
\]

The regression coefficients, $b_1, b_2, b_3$ and $b_4$, measure the odds, $p/(1 - p)$, for each of the factors in the model. An odds ratio of 1.0 indicates the factor has no effect. Either PM$_{2.5}$ or ozone is the last variable to be fit into the model so that any effect they have after removing the weather effects is tested. The R software package “survival” was used for the computations (R survival, 2016). This model was fit for all eight air basins.

6.2.2.2 Time series regression

A second analysis, time series regression, was computed for each air basin. In our case, we used stepwise regression. First, to remove seasonal effects, Day of Year and Day of Year squared, DOY and DOY2, were fit into the model for each air basin. The following additional variables were available for selection: Air quality variables (ozone and PM$_{2.5}$); weather variables (tmin, tmax, relative humidity) for the day at issue and for the previous day; mortality variable for the day previous to the current day; day before current day for air quality and weather variables. For each air basin (8) and cause of deaths (3), we recorded four marker p-values
(4): PM$_{2.5}$, Ozone, PM$_{2.5}$-1, and Ozone-1, considering that a change in the same day’s air quality or the previous day’s air quality might increase mortality. A total of 96 p-values were computed. All the predictor p-values for an air basin were examined. Terms were inserted or taken out of the analysis if they were ($p < 0.01$) or not ($> 0.01$) significant and affected the four marker p-values.

We examined the 96 resulting p-values in four ways. First, we looked for a consistent effect, either for one of the marker p-values, air basins or mortality types. Second, we provided a histogram of the p-values. Third, we provided summary statistics of the p-values. Finally, the 96 p-values were examined using a p-value plot: p-values plotted against their expectations under the assumption of a uniform distribution using JMP Add-In (JMP, 2016b). If p-values fall on a 45-degree line in a p-value plot, then they are consistent with there being no effect.

Both case-crossover and time series regression are standard methods for evaluating time series air quality/health effect associations ([Nawrot et al., 2011]). In the case-crossover analysis, we chose to look at the same day’s values only and not to include the previous day’s values. Lags are often introduced into the modeling process, but this approach using a large dataset does not support lags for heart attacks or stroke ([Milojevic et al., 2014]). We did use 0 and 1-day lags in the stepwise regression analysis.

### 6.3 Results

#### 6.3.1 Case-crossover analysis

The results of the case-crossover analysis odds ratios are given in Table 6.2 for PM$_{2.5}$ and ozone. There are 16 odds ratios. They range in value from 0.998 to 1.001. All these odds are very close to one, the no-effect value, for each air basin and air quality measure. The average odds ratio for the eight air basins is 1.00040 for PM$_{2.5}$ and 0.99977 for ozone. For each odds ratio, we give lower and upper confidence limits: CLL and CLH. That these confidence limits are very narrow indicates there is high statistical power. We give two p-values for each combination of air basin and air quality measure, the unadjusted p-value, $p$-val, and the false discovery rate p-value FDR ([Benjamini and Hochberg, 1995]). The unadjusted
p-value treats each combination without regard to the other statistical tests. The FDR adjusts the p-values to reflect the fact that multiple questions are at issue. With multiple tests, one expects occasional nominal statistical significance. Here, the smallest unadjusted p-value is 0.008 for ozone/San Francisco. The adjusted p-value, FDR, indicates that we would expect a p-value as small as 0.008, in about 12.8% of experiments where there are 16 statistical tests, and thus such a finding is not unexpected.

Table 6.2. Results of case-crossover analysis.
Odds ratio(OR), for eight air basins, confidence limits, p-values (raw and adjusted for 16 test); PM$_{2.5}$ and ozone. There are 16 statistical tests of hypothesis.

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>CLL</th>
<th>CLH</th>
<th>p-val</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PM$_{2.5}$/Air Basin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountain-counties</td>
<td>0.99947</td>
<td>0.99780</td>
<td>1.00100</td>
<td>0.536</td>
<td>0.673</td>
</tr>
<tr>
<td>sacramento-valley</td>
<td>1.00100</td>
<td>0.99990</td>
<td>1.00200</td>
<td>0.092</td>
<td>0.288</td>
</tr>
<tr>
<td>salton-sea</td>
<td>1.00000</td>
<td>0.99800</td>
<td>1.00200</td>
<td>0.956</td>
<td>0.989</td>
</tr>
<tr>
<td>san-diego</td>
<td>1.00056</td>
<td>0.99940</td>
<td>1.00200</td>
<td>0.341</td>
<td>0.673</td>
</tr>
<tr>
<td>san-francisco</td>
<td>1.00098</td>
<td>1.00010</td>
<td>1.00200</td>
<td>0.024</td>
<td>0.192</td>
</tr>
<tr>
<td>san-joaquin</td>
<td>1.00067</td>
<td>1.00000</td>
<td>1.00100</td>
<td>0.044</td>
<td>0.235</td>
</tr>
<tr>
<td>south-central-coast</td>
<td>1.00052</td>
<td>0.99900</td>
<td>1.00200</td>
<td>0.501</td>
<td>0.673</td>
</tr>
<tr>
<td>south-coast</td>
<td>1.00000</td>
<td>0.99990</td>
<td>1.00060</td>
<td>0.223</td>
<td>0.510</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>1.00040</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ozone/Air Basin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mountain-counties</td>
<td>1.00017</td>
<td>0.99720</td>
<td>1.00300</td>
<td>0.911</td>
<td>0.989</td>
</tr>
<tr>
<td>sacramento-valley</td>
<td>1.00000</td>
<td>0.99930</td>
<td>1.00200</td>
<td>0.437</td>
<td>0.673</td>
</tr>
<tr>
<td>salton-sea</td>
<td>0.99795</td>
<td>0.99570</td>
<td>1.00000</td>
<td>0.080</td>
<td>0.288</td>
</tr>
<tr>
<td>san-diego</td>
<td>1.00048</td>
<td>0.99930</td>
<td>1.00200</td>
<td>0.407</td>
<td>0.673</td>
</tr>
<tr>
<td>san-francisco</td>
<td>1.00111</td>
<td>1.00030</td>
<td>1.00200</td>
<td>0.008</td>
<td>0.128</td>
</tr>
<tr>
<td>san-joaquin</td>
<td>1.00000</td>
<td>0.99890</td>
<td>1.00100</td>
<td>0.989</td>
<td>0.989</td>
</tr>
<tr>
<td>south-central-coast</td>
<td>0.99855</td>
<td>0.99680</td>
<td>1.00000</td>
<td>0.108</td>
<td>0.288</td>
</tr>
<tr>
<td>south-coast</td>
<td>0.99990</td>
<td>0.99940</td>
<td>1.00030</td>
<td>0.547</td>
<td>0.673</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.99977</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.3.2 Time series regression

For each mode of deaths, air basin, and air quality measure, we undertook a stepwise regression analysis where we first put DOY and DOY2 into the model to correct for seasonal effects. Fig. 6.1 (a) displayed daily deaths versus DOY, overlaying all four years in the same figure. A seasonal effect is apparent. Once
DOY and DOY2 are used to detrend the time series, we see that most of the seasonal effects are removed, Fig. 6.1 (b). We give the four, marker p-values for each type of deaths and air basin in Table 6.3. There are four p-values < 0.05, but there is no consistent pattern of small p-values. The p-values are uniformly distributed over the interval 0 to 1 in the histogram of Fig. 6.2, indicating pure randomness. The mean and median p-values are 0.4965 and 0.4788, respectively; they indicate no effect of the same day’s or the previous day’s air quality on acute mortality. The p-value plot for these data is consistent with pure randomness, Fig. 6.3.

Table 6.3. P-values testing health effects versus air quality.
AllCause deaths; CV: cardiovascular deaths; respiratory deaths for 8 air basins. There are 96 tests of hypothesis.

<table>
<thead>
<tr>
<th>Mortality</th>
<th>Air Basin</th>
<th>PM$_{2.5}$</th>
<th>Ozone</th>
<th>PM$_{2.5}$-1</th>
<th>Ozone-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllCause</td>
<td>Mountain</td>
<td>0.6195</td>
<td>0.7955</td>
<td>0.2183</td>
<td>0.6973</td>
</tr>
<tr>
<td>AllCause</td>
<td>Sacramento</td>
<td>0.2602</td>
<td>0.9916</td>
<td>0.3532</td>
<td>0.8088</td>
</tr>
<tr>
<td>AllCause</td>
<td>Salton Sea</td>
<td>0.4203</td>
<td>0.9666</td>
<td>0.0504</td>
<td>0.2702</td>
</tr>
<tr>
<td>AllCause</td>
<td>San Diego</td>
<td>0.5767</td>
<td>0.1055</td>
<td>0.4064</td>
<td>0.3587</td>
</tr>
<tr>
<td>AllCause</td>
<td>San Francisco</td>
<td>0.7555</td>
<td>0.0101</td>
<td>0.9943</td>
<td>0.9598</td>
</tr>
<tr>
<td>AllCause</td>
<td>San Joaquin</td>
<td>0.1465</td>
<td>0.2895</td>
<td>0.0457</td>
<td>0.2283</td>
</tr>
<tr>
<td>AllCause</td>
<td>South Central</td>
<td>0.3723</td>
<td>0.6815</td>
<td>0.4472</td>
<td>0.9577</td>
</tr>
<tr>
<td>AllCause</td>
<td>South Coast</td>
<td>0.5209</td>
<td>0.1728</td>
<td>0.1480</td>
<td>0.9722</td>
</tr>
<tr>
<td>CV</td>
<td>Mountain</td>
<td>0.2661</td>
<td>0.7684</td>
<td>0.3831</td>
<td>0.5008</td>
</tr>
<tr>
<td>CV</td>
<td>Sacramento</td>
<td>0.1650</td>
<td>0.8144</td>
<td>0.7806</td>
<td>0.3712</td>
</tr>
<tr>
<td>CV</td>
<td>Salton Sea</td>
<td>0.8177</td>
<td>0.1488</td>
<td>0.2897</td>
<td>0.6428</td>
</tr>
<tr>
<td>CV</td>
<td>San Diego</td>
<td>0.3004</td>
<td>0.4052</td>
<td>0.0313</td>
<td>0.9563</td>
</tr>
<tr>
<td>CV</td>
<td>San Francisco</td>
<td>0.5823</td>
<td>0.5642</td>
<td>0.7499</td>
<td>0.3862</td>
</tr>
<tr>
<td>CV</td>
<td>San Joaquin</td>
<td>0.1535</td>
<td>0.4022</td>
<td>0.8076</td>
<td>0.3879</td>
</tr>
<tr>
<td>CV</td>
<td>South Central</td>
<td>0.7077</td>
<td>0.1666</td>
<td>0.3564</td>
<td>0.3569</td>
</tr>
<tr>
<td>CV</td>
<td>South Coast</td>
<td>0.7189</td>
<td>0.1142</td>
<td>0.7544</td>
<td>0.3380</td>
</tr>
<tr>
<td>Resp</td>
<td>Mountain</td>
<td>0.1804</td>
<td>0.9537</td>
<td>0.9665</td>
<td>0.7769</td>
</tr>
<tr>
<td>Resp</td>
<td>Sacramento</td>
<td>0.4111</td>
<td>0.5675</td>
<td>0.3990</td>
<td>0.7982</td>
</tr>
<tr>
<td>Resp</td>
<td>Salton Sea</td>
<td>0.9185</td>
<td>0.9624</td>
<td>0.4824</td>
<td>0.6192</td>
</tr>
<tr>
<td>Resp</td>
<td>San Diego</td>
<td>0.5025</td>
<td>0.6570</td>
<td>0.7591</td>
<td>0.6939</td>
</tr>
<tr>
<td>Resp</td>
<td>San Francisco</td>
<td>0.1539</td>
<td>0.6546</td>
<td>0.0344</td>
<td>0.1809</td>
</tr>
<tr>
<td>Resp</td>
<td>San Joaquin</td>
<td>0.6757</td>
<td>0.0538</td>
<td>0.5801</td>
<td>0.0716</td>
</tr>
<tr>
<td>Resp</td>
<td>South Central</td>
<td>0.4753</td>
<td>0.5464</td>
<td>0.0710</td>
<td>0.7504</td>
</tr>
<tr>
<td>Resp</td>
<td>South Coast</td>
<td>0.0923</td>
<td>0.4710</td>
<td>0.5538</td>
<td>0.8612</td>
</tr>
</tbody>
</table>
Figure 6.2. Histogram of p-values.
If the effects are random, we should see a uniform distribution.

<table>
<thead>
<tr>
<th>Distributions</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>0.0% maximum</td>
<td>0.99430</td>
</tr>
<tr>
<td>95%</td>
<td>0.99430</td>
</tr>
<tr>
<td>97.5%</td>
<td>0.98336</td>
</tr>
<tr>
<td>90.0%</td>
<td>0.95448</td>
</tr>
<tr>
<td>75.0% quartile</td>
<td>0.75340</td>
</tr>
<tr>
<td>50.0% median</td>
<td>0.47883</td>
</tr>
<tr>
<td>25.0% quartile</td>
<td>0.26168</td>
</tr>
<tr>
<td>10.0%</td>
<td>0.10154</td>
</tr>
<tr>
<td>2.5%</td>
<td>0.03262</td>
</tr>
<tr>
<td>0.5% minimum</td>
<td>0.01010</td>
</tr>
<tr>
<td>0.0%</td>
<td>0.01010</td>
</tr>
</tbody>
</table>

Summary Statistics:
- Mean: 0.40651
- Std Dev: 0.28888
- Std Err Mean: 0.02948
- Upper 95% Mean: 0.55504
- Lower 95% Mean: 0.43798
- N: 96.00000

Figure 6.3. P-value plot of 96 tests of hypothesis, $-\log_{10}$ of p-values versus the expectation of p-values coming from a uniform distribution.
If the points fall on a 45-degree line, then the results are consistent with randomness.


6.4 Conclusions and Discussion

In the absence an association of air quality, as measured by ozone or PM$_{2.5}$, with acute mortality (AllCause, Cardiovascular or Respiratory), there is no evidence supporting current air quality being causal of acute deaths in California.

Neither the case-crossover or stepwise regression analyses support a PM$_{2.5}$ or ozone association with acute deaths in the eight California air basins over the period 2004 - 2007. A common assumption today is that “air pollution”, no matter what level or what component is under consideration, may be detrimental to health. During the Great Smog of London, there was a dramatic increase in statistical deaths. It is fair to say the air was polluted as there was demonstrable harm. Air quality has improved dramatically since 1952 ([Schwartz et al., 2007]). The current paradigm, based on many epidemiological studies, is that air quality is causal of acute deaths. However, an association does not imply causation. If there is real causation, then well-conducted association studies using large datasets should almost always find an association. Multiple studies going back to at least 2000 ([Krewski et al., 2000]), indicate air quality geographic heterogeneity. If there is real causality, and one size fits all, there should be effects everywhere. In addition to geographic heterogeneity, a number of studies find no association between air quality and acute deaths, for example, ([Cox Jr et al., 2013]; [Wang et al., 2015]; [Yang et al., 2004]). Meta-analysis studies suggest health effects of air quality ([Nawrot et al., 2011]; [Mustafić et al., 2012]; [Shah et al., 2015]). The primary or base studies for these meta-analysis studies are only observational. These types of studies are not free of bias as none of them correct for multiple testing or multiple modeling. Careful counting of many of these studies shows that the analysis search space, the number of possible claims at issue, is large for each paper, for example, ([Young, 2017]), so their reliability is questionable, and verification is needed.

In epidemiology, quasi-experiments/natural experiments are considered more reliable. In a natural experiment, there is some event that is largely independent of the outcome, and there are other covariates that can be used to examine the relationship between outcomes and predictors, human mortality and air quality. The process looks like a real experiment. For example, some counties in the US were designated as out of compliance with respect to air quality and special efforts
were made to improve air quality in those counties ([Chay et al., 2003]). Air quality did improve, but elderly mortality did not. There was no experimental verification that improving air quality improved mortality. A re-analysis of the dataset in ([Chay et al., 2003]) reached the same conclusion ([Obenchain and Young, 2017]). In another study, an increase in PM$_{2.5}$ happened in New York City and Boston, due to forest fires in Canada ([Zu et al., 2016]). They found that PM$_{2.5}$ increased, but the mortality did not. These studies carefully controlled for confounding variables.

It is worth pointing to early evidence on etiology and atmospheric chemistry. [Nemery et al., 2001] examined the technical report of the Meuse Valley event of 1930. A thick fog formed, and 60 people died, of whom ten were necropsied. The necropsies showed no cardiovascular involvement, which supports [Milojevic et al., 2014], discussed later. The lung effects were consistent with acid injuries. They proposed sulfuric acid carried deep into the lungs adsorbed onto small particles. [Wang et al., 2016] present evidence on the atmospheric chemistry of sulfur compounds. To get to sulfuric acid you need a combination of conditions, “The sulfate formation was greatly facilitated by high RH (relative humidity), low temperature, and the presence of large fog droplets (45), yielding elevated sulfuric acid levels that persisted throughout the event.” These conditions held for the London Smog and the Muse Valley fog. This combination of conditions rarely occurs in the Los Angeles air basin. It appears that a complex interaction is needed for acute deaths. Over time, sulfur compounds have been dramatically reduced, so this complex interaction is much less likely in California or indeed the US.

The level of precision exhibited in our case-crossover analysis is very high; the confidence limits are very narrow. Confidence limits reflect the statistical precision of an analysis process but do not necessarily correct for bias. In this case, any small bias could lead to what looks like a significant effect. The fact that we see no effect suggests that there is little or no bias in this dataset and analysis. The few nominally statistically significant results could be due to small biases or be due to chance.

We define acute death in this study as death due to some immediate change of weather or air quality. The hypothesis is that something happened with these variables on the same day or the previous day that is associated with mortality. If one concedes that air quality is not causal of acute deaths, then there
still might be a chronic causal effect. No chronic effect of fine particles in California was found using a large cohort database ([Enstrom, 2005]). Since then Enstrom has accumulated other estimates of the chronic effect of PM$_{2.5}$ on All-Cause deaths for California. Results of relative risk, computed from several studies by [Enstrom, 2017], are given in Table 6.4. The average of the 20 given results is 1.010 with a standard error of the mean of 0.010. There is no apparent difference between the observed value and 1.000, the no-effect value. Positive association

### Table 6.4. All-Cause risk ratios from cohort studies for PM$_{2.5}$ deaths in California.

See Enstrom (2017) for details.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Years</th>
<th>RR</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>McDonnell et al., 2000</td>
<td>1976-1992</td>
<td>1.030</td>
<td>0.950-1.120</td>
</tr>
<tr>
<td>Krewski, 2000</td>
<td>1982-1989</td>
<td>0.872</td>
<td>0.805-0.944</td>
</tr>
<tr>
<td>Enstrom, 2005</td>
<td>1973-1982</td>
<td>1.039</td>
<td>1.010-1.069</td>
</tr>
<tr>
<td>Enstrom, 2005</td>
<td>1983-2002</td>
<td>0.997</td>
<td>0.978-1.016</td>
</tr>
<tr>
<td>Jerrett et al., 2005</td>
<td>1982-2000</td>
<td>1.110</td>
<td>0.990-1.250</td>
</tr>
<tr>
<td>Enstrom, 2006</td>
<td>1973-1982</td>
<td>1.061</td>
<td>1.017-1.106</td>
</tr>
<tr>
<td>Enstrom, 2006</td>
<td>1983-2002</td>
<td>0.995</td>
<td>0.968-1.024</td>
</tr>
<tr>
<td>Zeger et al., 2008</td>
<td>2000-2005</td>
<td>0.989</td>
<td>0.970-1.008</td>
</tr>
<tr>
<td>Jerrett, 2010</td>
<td>1982-2000</td>
<td>0.994</td>
<td>0.965-1.025</td>
</tr>
<tr>
<td>Krewski, 2010</td>
<td>1982-2000</td>
<td>0.960</td>
<td>0.920-1.002</td>
</tr>
<tr>
<td>Krewski, 2010</td>
<td>1982-2000</td>
<td>0.968</td>
<td>0.916-1.022</td>
</tr>
<tr>
<td>Jerrett, 2011</td>
<td>1982-2000</td>
<td>0.994</td>
<td>0.965-1.024</td>
</tr>
<tr>
<td>Jerrett, 2011</td>
<td>1982-2000</td>
<td>1.002</td>
<td>0.992-1.012</td>
</tr>
<tr>
<td>Lipsett et al., 2011</td>
<td>2000-2005</td>
<td>1.010</td>
<td>0.950-1.090</td>
</tr>
<tr>
<td>Ostro et al., 2010</td>
<td>2002-2007</td>
<td>1.060</td>
<td>0.960-1.160</td>
</tr>
<tr>
<td>Jerrett et al., 2013</td>
<td>1982-2000</td>
<td>1.060</td>
<td>1.003-1.120</td>
</tr>
<tr>
<td>Jerrett et al., 2013</td>
<td>1982-2000</td>
<td>1.028</td>
<td>0.957-1.104</td>
</tr>
<tr>
<td>Ostro et al., 2015</td>
<td>2001-2007</td>
<td>1.010</td>
<td>0.980-1.050</td>
</tr>
<tr>
<td>Thurston et al., 2016</td>
<td>2000-2009</td>
<td>1.020</td>
<td>0.990-1.040</td>
</tr>
<tr>
<td>Enstrom, 2016 (unpub)</td>
<td>2000-2009</td>
<td>1.001</td>
<td>0.949-1.055</td>
</tr>
</tbody>
</table>

studies on air quality and human mortality often point to cardiovascular effects as a possible etiology. Heart attacks and stroke were studied in a large UK dataset and the time of the event, heart attack or stroke, determined down to the hour ([Milojevic et al., 2014]). They found no lag effects for CO, NO$_2$, Ozone, PM$_{10}$, PM$_{2.5}$, or SO$_2$. Tellingly, they found no association between ozone and PM$_{2.5}$ and heart attacks or stroke. The association of hospital heart attack admissions
for CO, NO, NO$_2$, Ozone or PM$_{2.5}$ was investigated; no associations were found ([Wang et al., 2015]). The reliability of cause of death on death certificate is poor ([Ravakhah, 2006]), so it makes sense that attention should focus on AllCause deaths as the primary endpoint of analysis. We present an analysis of three death endpoints so that our results can be matched against the literature. We find no association between PM$_{2.5}$ and ozone and acute deaths in California.
General Procedures & Conclusions

7.1 Manual on Detrending and Detecting

According to the systematic investigation of air quality studies, a statistical manual on how to analyze this type of time series data is formed and detailed as follows. In general, there are two essential components: detrending and detecting.

For detrending, there are spline smoothing, kernel smoothing, and the proposed time series smoothers. The former two methods are well established; however, their purpose is to regress toward the mean of the nearest neighborhood and may commit over-fitting problems under the previous selection criteria such as CV, GCV and etc. One probable remedy is to use the proposed detrending diagnosis to ensure that the detrended time series do not contain the long-term trend but do retain the short-term signals. The proposed time series smoothers are specifically devised to retain the short-term signals. The validity and stability are well demonstrated. After detrending, if the short-term signals are large enough, the cross-association can be detected by any dependence measure. If they are subtle, statistical modeling can surely provide more insight into the short-term effect.

For detecting, two prevailing methods are demonstrated and compared via analyzing the California dataset. Basically, both methods should draw the same conclusion. However, when there are large samples and no missing data at different periods, time series regression with proper consideration for over-dispersion should be preferred, due to the computational advantage. When the sample size is small and missing data exist, case-crossover design and analysis can be more applicable.
In summary, the two-stage approach, first detrending and then detecting, can provide more consistent results in environmental epidemiology, especially the air quality studies.

7.2 Conclusion and Future Work

Our contributions towards the field of statistics and epidemiology are mainly in twofold, the scientific and statistical aspects. In the air quality studies, firstly, the characteristics of air quality and acute effect are identified. Both the level and change of air quality are examined as well as their cross association with human mortality. Secondly, it is concluded that the contemporary air quality in California does not have an acute effect on elderly individuals. Thirdly, researchers should be very cautious about certain air quality conditions, especially when the wind or precipitation is low and the sunlight is strong.

In the field of statistics, our contributions for time series detrending are summarized below. First of all, a new collection of time series smoothers are proposed and targeted to reveal the short-term signals. Second, a new recursive algorithm is proposed in the moving recursive mean to exhaustively extract the short-term signals. Third, the method’s properties validity and stability are shown; the window-gap selection procedure is proposed and detrending diagnosis is also devised. These statistical tools help us in selecting parameters and evaluating different detrending results. Part of this work is published in “Time series smoother for effect detection” [You et al., 2018b]. More work will be documented and submitted for publication.

Our statistical contributions for detecting the cross-association are summarized as follows. First of all, two prevailing methods, case-crossover analysis and time series regression, in the air quality studies are utilized, compared, and cross-verified in the air quality studies. Second, multiple testing is carefully considered and analyzed in air quality studies. Third, some preliminary meta-analysis work has been done by addressing the selection bias. These works are documented in “PM 2.5 and ozone, indicators of air quality, and acute deaths in California” [You et al., 2018a].

In addition, my environmental research horizon is being expanded. Besides the air quality studies, the water quality in Pennsylvania is being investigated; new statistical and data mining methods are proposed and utilized. So far, two
publications [Li et al., 2016] and [Wendt et al.,] have come out. More works are in the pipeline and new dependence measure on censored data is being considered.

In future, on environmental epidemiology, there are two major directions that are quite prominent and promising. First, the practice of statistical inference on the cross-association should be revisited and improved. The current methodology does not seem to take into consideration the varying covariance structure of multiple time series and assure that the long-trend trend is properly removed. Although the detrending can make the data more or less independent, the regime switch of the covariance structure, due to seasonality or other cycles, should be incorporated. Second, there are multiple regions and cities that may pose air quality issues. How to evaluate the complex exposure-response relationships all together is quite challenging. Hierarchical modeling may be a solution; however, all the parametrization details need to be worked out. In a word, more interesting questions await to be discovered and answered.
Bibliography


Wagner et al., 2002. Segmented regression analysis of interrupted time series


Vita
Cheng You

EDUCATION

• Doctor of Philosophy in Statistics  Dec 2018
  The Pennsylvania State University, GPA: 3.96/4.00

• Master of Science in Statistics  Dec 2012
  The Pennsylvania State University, GPA: 3.95/4.00

• Bachelor of Science in Statistics and Operations Research  Jun 2009
  Hong Kong Baptist University, First Honor, Minor in Finance

PUBLICATIONS


HONORS AND AWARDS

• Biopharm-Deming Student Scholar Award  Dec 2017
  Deming Conference

• Quality and Productivity Award  Jun 2016
  Joint Statistical Meetings, ASA

• American Petroleum Institute Research Travel Grant  Jun 2015
  American Petroleum Institute

• General Electric Research Assistantship  Jan 2015 - Aug 2015
  Pennsylvania State University

• Hong Kong Jockey Club Full Scholarship for Outstanding Mainland Students (Tuition & Stipend)  Aug 2005 - Jun 2009
  Hong Kong Baptist University