AIR POLLUTION AND HEALTH: TIME SERIES TOOLS
AND ANALYSIS

by

WESLEY S. BURR

A thesis submitted to the
Department of Mathematics & Statistics
in conformity with the requirements
for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
October 2012

Copyright © Wesley S. Burr, 2012
Abstract

This thesis is concerned, loosely, with time series analysis. It is also, loosely, concerned with smoothers and Generalized Additive Models. And, finally, it is also concerned with the estimation of health risk due to air pollution.

In the field of time series analysis, we develop two data-driven interpolation algorithms for interpolation of mixed time series data; that is, data which has a stationary or “almost” stationary background with embedded deterministic trend and sinusoidal components. These interpolators are developed to deal with the problem of estimating power spectra under the condition that some observations of the series are unavailable.

We examine the structure of time-based cubic regression spline smoothers in Generalized Additive Models and demonstrate several interpretation problems with the resultant models. We propose, implement, and test a replacement smoother and show dramatic improvement. We further demonstrate a new, spectrally motivated way of examining residuals in Generalized Additive Models which drives many of the findings of this thesis.

Finally, we create and analyze a large-scale Canadian air pollution and mortality database. In the course of analyzing the data we rebuild the standard risk estimation model and demonstrate several improvements. We conclude with a comparison of the original model and the updated model and show that the new model gives consistently more positive risk estimates.
Acknowledgements

To ...

David J. Thomson: for freedom to explore, encouragement to fail, and willingness to listen.

Aaron Springford, Charlotte Haley, Dave Riegert, Josh Pohlkamp-Hartt and Karim Rahim: for many, many hours of discussion and encouragement. And editing, too!

Glen Takahara: for thinking of me when you needed a “time series guy”, and giving me inspiration for a whole new way of thinking.

And, most of all ...

To Belinda

This is for you. Sorry it took so long!
# Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iii

List of Tables vii

List of Figures ix

Chapter 1: Introduction 1

Chapter 2: Background and Literature Review 3

2.1 A Brief History of Spectral Estimation 3

2.2 Interpolating Stationary Stochastic Processes 12

2.2.1 Conclusion and Transition 22

Chapter 3: Univariate Interpolation of Time Series 24

3.1 Hilbert Space Background 24

3.1.1 Prediction Equations for Stationary Processes 27

3.1.2 Extending the Algorithm to Interpolation 28
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.3 Estimating Error in the Gaussian Case</td>
<td>29</td>
</tr>
<tr>
<td>3.1.4 Kolmogorov’s Formula</td>
<td>30</td>
</tr>
<tr>
<td>3.1.5 Notes</td>
<td>32</td>
</tr>
<tr>
<td>3.2 Interpolation Algorithm</td>
<td>32</td>
</tr>
<tr>
<td>3.2.1 E Step</td>
<td>33</td>
</tr>
<tr>
<td>3.2.2 M Step</td>
<td>34</td>
</tr>
<tr>
<td>3.2.3 $M_t$: The Mean Term</td>
<td>35</td>
</tr>
<tr>
<td>3.2.4 $T_t$: The Trend Term</td>
<td>39</td>
</tr>
<tr>
<td>3.2.5 Sub-Iteration</td>
<td>43</td>
</tr>
<tr>
<td>3.2.6 Best Least-Squares Prediction (Interpolation)</td>
<td>44</td>
</tr>
<tr>
<td>3.2.7 Global M Step Iteration</td>
<td>44</td>
</tr>
<tr>
<td>3.2.8 Global Algorithm</td>
<td>46</td>
</tr>
<tr>
<td>3.3 Examples Using Simulation</td>
<td>46</td>
</tr>
<tr>
<td>3.3.1 Extending Eqn. 3.35</td>
<td>48</td>
</tr>
<tr>
<td>3.4 An ARMA(4,4) Example</td>
<td>53</td>
</tr>
<tr>
<td>3.4.1 Conclusions from the Simulations</td>
<td>60</td>
</tr>
<tr>
<td>3.5 Observational Example: Noon Solar Flux</td>
<td>60</td>
</tr>
<tr>
<td>3.6 Summary and Conclusions</td>
<td>64</td>
</tr>
<tr>
<td><strong>Chapter 4:</strong></td>
<td></td>
</tr>
<tr>
<td>Bivariate Interpolation of Time Series</td>
<td>67</td>
</tr>
<tr>
<td>4.1 Bivariate Interpolation Algorithm</td>
<td>71</td>
</tr>
<tr>
<td>4.2 Simulated Data</td>
<td>73</td>
</tr>
<tr>
<td>4.2.1 Bivariate Realization of Eqn. 3.35</td>
<td>73</td>
</tr>
<tr>
<td>4.2.2 Bivariate Realization of Eqn. 3.56</td>
<td>80</td>
</tr>
<tr>
<td>4.3 Real Data: Revisiting the Solar Flux</td>
<td>84</td>
</tr>
<tr>
<td>4.4 Conclusions</td>
<td>85</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Estimates of four sinusoids from a realization of Eqn. 3.35.</td>
<td>42</td>
</tr>
<tr>
<td>3.2</td>
<td>Error results for Eqn. 3.35 and randomly scattered gaps.</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Error results for model of Eqn. 3.35 and block gaps.</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Continuation of Table 3.3.</td>
<td>51</td>
</tr>
<tr>
<td>3.5</td>
<td>Error results for model of Eqn. 3.56 and randomly scattered gaps.</td>
<td>56</td>
</tr>
<tr>
<td>3.6</td>
<td>Errors from model of Eqn. 3.56 and block gaps.</td>
<td>57</td>
</tr>
<tr>
<td>3.7</td>
<td>Continuation of Table 3.6.</td>
<td>58</td>
</tr>
<tr>
<td>3.8</td>
<td>Moments of interpolation error, Penticton daily noon solar flux.</td>
<td>65</td>
</tr>
<tr>
<td>4.1</td>
<td>Error results for Eqn. 4.12 and randomly scattered gaps.</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Error results for Eqn. 4.12 and block gaps.</td>
<td>79</td>
</tr>
<tr>
<td>4.3</td>
<td>Error results for Eqn. 4.16 and randomly scattered gaps.</td>
<td>81</td>
</tr>
<tr>
<td>4.4</td>
<td>Error results for Eqn. 4.16 and block gaps.</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>Examination of coefficients for Chicago using (5.14).</td>
<td>113</td>
</tr>
<tr>
<td>5.2</td>
<td>Coefficients for Chicago using Eqn. 5.14.</td>
<td>122</td>
</tr>
<tr>
<td>5.3</td>
<td>Estimation of power below frequency cutoff.</td>
<td>124</td>
</tr>
<tr>
<td>5.4</td>
<td>Coefficients for Chicago using Eqn. 5.14.</td>
<td>131</td>
</tr>
<tr>
<td>5.5</td>
<td>Percentage of power below frequency cut-off.</td>
<td>131</td>
</tr>
<tr>
<td>5.6</td>
<td>Coefficients for Chicago using Eqn. 5.14.</td>
<td>134</td>
</tr>
<tr>
<td>5.7</td>
<td>Percentage of power below frequency cut-off, Eqns. 5.16 and 5.23.</td>
<td>134</td>
</tr>
<tr>
<td>6.1</td>
<td>Coefficients and standard errors for Eqn. 6.2.</td>
<td>146</td>
</tr>
</tbody>
</table>
List of Figures

3.1 Realization of example series given in Eqn. 3.35. . . . . . . . . . . . . . . . . 35
3.2 Polynomial fit to example of Eqn. 3.35. . . . . . . . . . . . . . . . . . . . . . . 38
3.3 Harmonic F-test Statistic for residual series, Eqn. 3.35 minus $M_t$. . . . 43
3.4 One realization of Eqn. 3.35. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
3.5 Histogram of errors for $g_d = 1$ interpolations. . . . . . . . . . . . . . . . . . . 52
3.6 Realization of Eqn. 3.56. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
3.7 Histogram of errors for $g_d = 2$ interpolations. . . . . . . . . . . . . . . . . . . 55
3.8 Errors from the 20% missing case for $g_d = 5$ using Eqn. 3.56. . . . . . 59
3.9 Daily noon solar flux. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61
3.10 Interpolation example, daily noon solar flux, Sagamore Hill. . . . . . . 62
3.11 Smoothed pointwise variance, Sagamore Hill daily noon solar flux. . 63

4.1 Realization of $X_t$ for Eqn. 4.12 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.2 Realization of $Y_t$ for Eqn. 4.12 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.3 Example of bivariate interpolation for Eqn. 4.12, 25% missing . . . . . . 75
4.4 Second example of bivariate interpolation for Eqn. 4.12, 25% missing 75
4.5 Variance of errors in bivariate interpolation using Eqn. 4.12. . . . . . . . 78
4.6 Variance of errors in bivariate interpolation for Eqn. 4.16. . . . . . . . . . . . 83
4.7 Bivariate interpolation of daily noon solar flux from Sagamore Hill. . . 84
4.8 Multitaper power spectra: Penticton and Sagamore - I. . . . . . . . . . . . . 86
4.9 Multitaper power spectra: Penticton and Sagamore - II. . . . . . . . . . . . . 87
C.2 Models 1,2,5 and 6 from Table C.1 – AC Mortality, CDs 13-24 . . . . . 213
C.3 Models 3,4,7 and 8 from Table C.1 – CP Mortality, CDs 1-24 . . . . . 214
C.4 Models 3,4,7 and 8 from Table C.1 – CP Mortality, CDs 1-12 . . . . . 215
C.5 Models 1 and 2 from Table C.1 – AC Mortality, CDs 1-24 . . . . . . 216
C.6 Models 3 and 4 from Table C.1 – CP Mortality, CDs 1-24 . . . . . . 217
C.7 Models 5 and 6 from Table C.1 – AC Mortality, CDs 1-24 . . . . . . 218
C.8 Models 7 and 8 from Table C.1 – CP Mortality, CDs 1-24 . . . . . . 219

D.1 Realization of Eqn. D.1. . . . . . . . . . . . . . . . . . . . . . . . . . 221
D.2 Coefficients from linear model fitting lagged copies of predictor . . . 222
D.3 Realization of Eqn. D.3. . . . . . . . . . . . . . . . . . . . . . . . . . 223
D.4 Coefficients from regressed lagged copies of prediction, Eqn. D.3 . . . 224
D.5 Demonstration of synthetic lags using Toronto, CD 3520 . . . . . . . 228
Chapter 1

Introduction

This thesis is composed of five chapters which can be roughly divided into two parts. The first part is concerned with time series, specifically those with gaps in the observational record. In Chapter 2 we examine the history of spectrum estimation and interpolation of time series, with numerous references and citations. This sets the stage for Chapter 3 in which we develop, demonstrate, and test a univariate multi-component interpolator for stationary time series with embedded deterministic components. The univariate interpolator in turn inspires the bivariate interpolator of Chapter 4. The real-world test case for both of these interpolators is the daily noon solar flux record, jointly observed at Sagamore Hill, Massachusetts and Penticton, British Columbia. This flux record was interpolated to provide a potential covariate for the models of the second part.

The second part of this thesis is concerned with Generalized Additive Models as used to estimate population health risk due to air pollution. These models, while a superset of linear models and not explicitly time series models, contain numerous elements that can be mapped to the time series world. In Chapter 5, by appealing to linear filter design, we demonstrate that the cubic regression spline smoother typically used in population health risk estimation is not an effective filter, and leaves
considerable long timescale variation in the residual as well as spurious peaks due to
the nonlinearity of the filter. As this directly contradicts the interpretation of the
results, we develop a new smoothing methodology that corrects the problems of the
original and gives risk estimates which can properly be interpreted in the fashion of
the original.

The final chapter of the thesis, Chapter 6, details the work required to create a
national Canadian database of mortality, air pollution and environmental covariates.
The work of Chapter 3 is used for interpolation of a number of pollution series. The
database is then analyzed and a new GAM structure developed, taking into account
all that was learned from Chapter 5 and appealing further to the worlds of time series
analysis and signal processing for several additional tools. Chapter 6 concludes with
a comparison of the original and updated models and provides some suggestions for
future work.

This thesis also contains four appendices. The first two give short introductions
to two software packages (spsmooth and tsinterp) that were written to support this
thesis. The first of these packages is available on CRAN, the Comprehensive R Archive
Network, and the second is still under development and will be posted there in the
near future. The third appendix contains some large plots showing coefficients for the
results of Chapter 6 and should be consulted when reading Section 6.10. The final
appendix is a tutorial presentation on the topic of phase for researchers interested in
the implications of Chapter 6 who may not have a thorough grounding in time series
analysis or signal processing.
Chapter 2

Background and Literature Review

When working with time series, we frequently desire estimation of the spectrum (or power spectrum, power spectral density) as part of our analysis. In many cases, the accurate estimation of the spectrum is, in fact, our primary goal. Unfortunately, when working with real-world data with all of its problems, we often encounter data sets with missing values. To date, there has been little success in developing satisfactory estimation procedures for estimation of the spectrum under this case, as all rigorous techniques require complete, contiguous data across the span of interest.

2.1 A Brief History of Spectral Estimation

The material covered in this synopsis is taken partially from Enders Robinson’s excellent paper [161], partially from the introduction to Larry Marple’s book [110], and partially from a further literature review.

The word spectrum was coined by Newton in 1671 [123, pg. 3076] to refer to the light emitted by a prism. The root of the word traces to specter, Latin for an “image” or “ghostly apparition” [110, pg. 3] and to the same word in Greek, for “to see”. As the adjective associated with spectrum is spectral, we shall follow Marple’s convention
and refer to *spectral estimation* rather than *spectrum estimation*, although the object remains a “spectrum”.

Spectral estimation for time-series can be tracked as far back as the mechanical harmonic analyzers of Sir William Thomson (Lord Kelvin) and to similar, manual, computations of Fourier series by researchers such as Hornstein [79]. This practical scientific work was continued by Stokes, Schuster and Rayleigh in the papers [179] (solar physics periodicities), [97] (numerous examples), [166] (26-day period in meteorological data), [153] (optics), [154] (radiation), [155] (spectrum of irregular disturbances, i.e., early stochastic processes), and [156] (comments on Fourier series, inconsistency). In these works the authors used an improving series of mechanical harmonic analyzers in the analysis of scientific measurements. This work can be thought of an extension of Fourier series analysis to non-harmonically related periods (such as tides), with the authors interested in determining which sinusoidal “signals” in the data were most significant. Formally, the estimates being made were of the decomposition of the power of a process into a function of frequency that is characteristic of the process; as at this point in time the fully realized concept of a statistical estimator was unknown\(^1\) the language of these papers does not match modern verbiage, but the work was nonetheless accurate.

Formally, the estimator being used at the turn of the century was known as the *periodogram*, as named by Schuster [166]. Unfortunately, as was also known by Rayleigh [156] (and others, including both Wiener and Tukey) this estimator is quite a poor one, and the results published using it as the primary analysis tool gave spectral estimation quite a bad name for decades to follow. We note here that a careful reading of [97, pg. 204] makes it clear that Stokes was grappling with the issues of missing data and that a main contributing factor for the “Cambridge gang” being able to complete

\(^1\)Although D. Bernoulli was working on the idea of Maximum Likelihood Estimators in 1777 [129].
their spectral estimation work was the existence of Kelvin’s harmonic analyzer, without which the computational burden would have been untenable. As briefly described in [110], the use of mechanical harmonic analyzers was only useful for time series with obvious periodicities; numerical methods for fitting Fourier series were still required for evaluating noisy data for nonobvious periodic signals. As one can imagine, manual computation of Fourier series in the world before calculators and computers was an intensive and time-consuming task. Schuster was also fully aware of the issues with the sidelobes around the mainlobe response in the periodogram that are inherent in all Fourier analyses of finite record lengths [110], and introduced averaging arguments to deal with this problem, a solution that would not be theoretically justified until 1930 with the work of Wiener [219].

The period between 1905 and 1930 was also interesting, with tremendous ground-breaking research being completed. In 1914 [229], Einstein published a short paper on “irregularly fluctuating series” which foreshadowed the work that would be done by both Wiener and Khinchin\(^2\). There were two complementary schools of research during this time: the theoretical school, led by Wiener and Kolmogorov, with contributions by Wold, Cramér and others, and the empirical school, led by Yule and Walker, further explored by Bartlett, and eventually taken up to great success by John Tukey. We briefly examine the contributions of these researchers in the following.

G.U. Yule, having studied under Karl Pearson and contributed to the improvements of Kelvin’s harmonic analyzers mentioned above, contributed several pioneering works: a epidemiological study of disease using frequency [68], a study of spurious correlation between separate series [232], and the first use of AR models as proposed replacements for Schuster’s periodogram [231]. Although there is no evidence that Yule and Sir Gilbert Walker ever worked together explicitly, it is clear [209] that

\(^2\)This is the reason that the Wiener-Khinchin theorem is occasionally referred to as the Einstein-Wiener-Khinchin theorem, as we shall refer to it in this thesis. Yaglom [229] expresses the opinion that Einstein’s derivation is actually cleaner than either Wiener or Khinchin.
they were aware of each other’s work. Walker [208, 209] applied similar techniques to
analysis of a damped sinusoidal time series and the combination of Yule and Walker’s
work led to the normal equations that arise out of this technique being named the
Yule-Walker equations. Further work that led to the moving-average process (named
by Herman Wold, below) was performed by E. Slutsky [174].

We now move to the work done by the Swedish school during the formative years
between 1920 and 1950. Harald Cramér [40, 41] developed theory on stationary ran-
dom processes, and presented the first formal organization of the spectral representa-
tion theorem, being honored for this development by the subsequent attachment of his
name. His 1940 paper [40] demonstrated the representation of correlation functions
by Fourier-Stieltjes integrals of the form
\[
R_{\mu \nu}(t) = \int_{-\infty}^{\infty} e^{itx} dF_{\mu \nu}(x),
\]
where the \( F_{\mu \nu}(x) \) were functions of bounded variation. More importantly, this paper
demonstrated that it is possible to find stationary random processes having the given
correlation functions. This is not quite the standard spectral representation we are
familiar with, but the bulk of the work has been done. Perhaps more interestingly,
A. Kolmogorov had also found the same theorem. This work is also known as Her-
glotz’s Theorem [21]3, and the extension of it the Einstein-Wiener-Khinchin theorem
discussed below. Cramér’s 1942 paper is more important. Reproduced in the 1992
book “Breakthroughs in Statistics” [42], the form of the representation clearly follows
the modern form:
\[
f(t) = \int_{-\infty}^{\infty} e^{itx} dF(x).
\]
which is considered by Cramér to be the general form. In this work, Cramér cites
Wiener [219, 222] and Bochner [18] as inspirations whose work foreshadowed his
developments.

3This work is also occasionally called Bochner’s Theorem due to his work in [18].
Cramér also supervised the PhD thesis of Herman Wold [225]. These two researchers are responsible for significant work in exhibiting the spectral representation theorem (Cramér, as described above), development of, and coining the terms for, the “moving-average process” and “autoregressive process” (Wold), and doing significant work on the decomposition theorem for stationary processes (joint). Peter Whittle trained with Wold at Uppsala as well, and should be mentioned for both his work on prediction [216, 218] and for formulating the autoregressive moving-average (ARMA) process [214].

A history would be incomplete without mentioning Kari Karhunen and Michel Loéve, whose work was jointly honored with the naming of the *Karhunen-Loéve Theorem*. Karhunen completed his thesis in Finland in 1947, with the results published that same year in the Annals of the Finnish Academy of Sciences4 [85], and Loéve his in 1941 at Université de Paris under Paul Lévy. Emmanuel Parzen points to [104, 105] as the critical works that, combined with Karhunen’s paper, provide the Karhunen-Loéve representation for random functions of second order on a finite interval [131, Section 4]. This representation allows for a function to be expanded in a series of orthonormal functions, typically with uncorrelated coefficients of expansion. More details can be found in Davenport and Root [44, pgs. 96–99] and Helstrom [78, pgs. 124–133]

Theoretically, much of the groundwork of the progress made in the 1940s (during World War II and immediately following) had been laid between 1930 and 1940 by Norbert Wiener and Raymond Paley [219, 127]. Wiener’s *Generalized Harmonic Analysis* played a fundamental role in placing spectral estimation on a firm theoretical basis, rigorously defining for the first time the concepts of *power spectral density* and *autocorrelation*. This was also the first place where what would become the *spectral representation theorem* was demonstrated (Cramér’s work being published in 1940

---

4Annales Academiae Scientiarum Fennicae.
and 1942 [40, 42], as detailed above), in Wiener’s notation as

\[ x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega n} B(\omega) E(\omega) d\omega. \]  (2.3)

The paper also included the theorem that the autocovariance and spectrum are a Fourier transform pair, a theorem that is now known as the Einstein-Wiener-Khinchin theorem. Corresponding to this work, in 1941 Kolmogorov published his complementary (yet independent) work [93]5.

Raymond Paley’s untimely death in 1933 cut short a promising research career, but Wiener continued his work, and during the war worked extensively on projects for the US Army. These projects led in 1949, after the war, to the publishing of the infamous book, “the yellow peril” [220], summarizing a large amount of previously classified research on prediction theory that occurred between 1938 and 1946. We must note here the importance of Norman Levinson’s work in translating Wiener’s results into engineering language [101, 100]. More details on this are given in the next section.

The two worlds of theoretical and empirical came together in June 1949 with the Wood’s Hole “Symposium on Autocorrelation Analysis Applied to Physical Problems”. The highlight of the symposium was a paper presented by John Tukey [204] that presented a method for accurately computing power spectra from empirical (short-run) data series. Previous methods of computing power spectra from empirical autocorrelation functions were too erratic to be useful [161]. This, combined with non-military access to the first computers, led to an explosion of research and progress in the field.

Two years later, in 1949, “The Measurement of Power Spectra” by R.B. Blackman and John Tukey [16] was published, introducing a generation of engineers to the ideas

---

5By all accounts, Kolmogorov and Wiener worked completely independently of one another, and only after their work was complete did both realize that the other had foreseen the same results. Kolmogorov’s work is somewhat more theoretical, which may be due to Wiener’s extensive involvement with military projects and the general atmosphere at MIT at the time.
of spectral estimation, smoothing, and tapers. In their book, two approaches were laid out: *direct* and *indirect* spectral estimation. In the direct method [212], the estimate is made from the data directly, possibly using a taper to control bias. The indirect method instead estimates first the autocorrelations of the series, then transforms these to the spectrum using the Fourier transform, an application of the Einstein-Wiener-Khinchin theorem. We here briefly mention the contributions of Richard Hamming, not least through his joint work with Tukey on smoothing of the periodogram [205].

Following the conclusion of World War II, M.S. Bartlett worked on similar problems at Manchester, and then Cambridge, publishing a book on stochastic processes [8]. He further published several influential papers [5, 6] on periodogram estimation, proposing an averaging method for multiple periodogram estimates that is a special case of Welch’s method, requiring no overlap. Some additional work was completed by Kendall [89, 90], with the two men additionally collaborating on several additional topics within the field.

The next major advance that would impact the world of spectral estimation was the (re)discovery of the Fast Fourier Transform algorithm in 1965 by Cooley and Tukey [38, 39]. This allowed the rapidly advancing computational tools of the day to be leveraged and dramatically changed the signal processing community. With this major advance, incredible computational gains were made, spreading from limited use in signal processing and geophysical exploration to practically every applied science field. Much work was done on developing different tapers for bias control, with an excellent overview given by Harris [73]. The further work of Welch [213] foreshadowed the eventual development of multitaper by introducing the idea of using multiple estimates of the spectrum together to control the variance. This can be further traced back to the work of Schuster [165, 166, 167, 168], especially that of [167].

From 1965, when the FFT was reinvented, through to 1982, significant work was done in expanding the field from its base to more complicated and adventurous areas.
Emmanuel Parzen contributed many pioneering papers, including [130, 133, 132, 134, 137, 136]. The 1961 paper [133] was particularly important, as in it Parzen: a) demonstrated the fundamental equation for spectral estimation, b) showed clearly why the periodogram is an inconsistent estimate of the spectral density function, and c) advocated for the use of the biased form of the sample autocovariance function to obtain both positive definiteness and reduced mean-square error. We especially note here point (b) which had also been mentioned by Rayleigh in 1912 [156] (although without modern statistical words, and in a physics context). It is unfortunate that this appears to have escaped the attention of many people in the years since – it is clear both from the formulation and discussion in this paper that the serious workers in the field were both aware of the deficiencies of the periodogram, and actively working to mitigate them. It is unclear how something that was so clearly known to be flawed in 1961 by Parzen, Bartlett, Tukey, Wiener, etc. should still be accepted in peer-reviewed publications today, 50 years later.

The next major change to occur in the field was completed, as many of the previous changes were, at Bell Labs. The work on the WT-4 waveguide project led David J. Thomson to work on both robust estimates of the power spectrum [193] and the multitaper formulation [194]. This work built on the discovery, also made at Bell Labs, by David Slepian, H.O. Pollak and H.J. Landau [173, 95, 96, 171, 172] (among others), of the optimally frequency-concentrated time-limited functions / sequences, named prolate spheroidal functions / sequences. These prolates, now known as Slepian functions and Slepian sequences (or Slepians for short if the context is clear), arose from the consideration of the eigenfunctions of the discrete Fourier transform, this work proceeding in parallel with the developments of Cooley, Tukey, Blackman and Welch, mentioned above.

With the multitaper, the scientific world finally had access to a consistent estimator of the power spectrum, with the first reasonable justification for tapering that had
been published\textsuperscript{6}. It is interesting to note here that it appears that Norbert Wiener was also aware of the use of tapers in spectral estimation, and may have in fact used them himself as part of the war work he completed. This is taken from a letter from Wiener to Tukey, quoted in David Brillinger’s article commemorating John Tukey’s death [20, pg. 1615].

With the multitaper in hand, significant work began in fitting the new estimator into the statistical framework that had been built up around spectral estimation. The series of papers by Mullis and Scharf, Bronez, and Stoica and Sundin [120, 23, 178] firmly placed the multitaper on strong theoretical footing, showing that the theoretical quadratic formulation implies the need for multiple tapers, and that such solutions are approximately maximum-likelihood.

Much of the work since 1990 has been concerned with nonstationary processes, with the energy spent being focused on wavelets and other (similar) formulations that attempt to simultaneously estimate the time and frequency structure. This work is quite orthogonal to the approach pioneered by Thomson, and has had little impact to the developments of this thesis. Possible exceptions that contribute developments on the multitaper approach include the conference paper by Thomson [199] which demonstrates possible uses for the eigencoeficients of the multitaper formulation, leveraging their Gaussian nature in statistical testing, and the theses of K. Q. Lepage [99] and G. Prieto [146] which both attempted to address the local properties of the multitaper spectral estimate.

We conclude here with a note about a particularly vexing property of spectral estimates. In all cases mentioned above, the analyst requires contiguous\textsuperscript{7} data. This introduces the curse of “missing data” – unfortunately, real scientific data is often

\textsuperscript{6}Previously, there had been some significant furor over the use of tapers, with the “pro” side using them “because John Tukey said so” (and as a way to work around the clearly understood problems with the consistency of the periodogram), and the “con” side refusing to use them for reasons best summarized by Burg [24, 25].

\textsuperscript{7}Evenly sampled, complete data – discretely sampled, but continuous within that context.
missing measurements, whether due to instrumental failure, environmental conditions or experiment error. All techniques mentioned above require no missing samples in their data span in order to provide estimates of the spectrum.

2.2 Interpolating Stationary Stochastic Processes

Interpolation is an ancient problem, dating back at least to Gauss and in some form, to the Greek empire. Erik Meijering has an excellent overview of parts of the field that begins with ancient astronomy and passes through sampling theory developments on its way to modern image processing [119]. For our purposes, we are specifically interested in algorithms that attempt to reconstruct a gappy (or irregularly sampled) series in some statistical sense, as we make the assumption from the beginning that we are dealing with a weakly stationary stochastic process, possibly extended to the case of a continuous background with embedded line components. The real data series we consider can be easily constrained to fit into this world; they are observations of environmental systems that mostly likely do have a continuous background of some sort, with the caveat that embedded into this background are numerous (possibly millions) line components, each of which should be considered to be only approximately sinusoidal. For our purposes, we will not extend the theory so far as to fully model these line components, although that extension has been considered and initial results indicate that the technique is possible.

The first major modern work on interpolation of stationary stochastic processes can be considered to be Wiener’s book [220], mentioned in the last section. In this book the problems of extrapolation and interpolation are taken up from a theoretical viewpoint, and the majority of the derivations completed. The theory developed in this book was subsequently constrained into algorithmic form by Levinson [101, 100]
and published in the Journal of Mathematics and Physics. These papers were subsequently included as Appendices B and C of future printings of [220]. In particular, the exposition in the following chapter owes a tremendous amount to Appendix C of [220, 100].

Wiener’s work between 1947 and 1964, the year of his death, did not delve further into the problem of univariate prediction of stationary stochastic processes, in no small part because [220] had firmly solved the linear case of univariate prediction. He did examine multivariate processes and nonlinear processes, and we will revisit this work later in this section.

In economics, Friedman [61] was concerned with the interpolation of time series by related series, which more directly influences the work of Chapter 4. The Friedman paper is not directly applicable to our work here, being more interested with geometric arguments for interpolation – in effect, consideration of difference equations and the use of related series to interpolate some small number of points.

We have examined the literature that extends and builds upon the Friedman paper [61]. Unfortunately, the primary application in econometrics is disaggregation, a technique whereby data observed on a coarse scale is decomposed in such a way as to increase the observing frequency. For example, data recorded quarterly can be disaggregated to form data points observed on a monthly scale. Significant work in this field is summarized in [139]. Of the extensive work reviewed in this paper, the only portion applicable to our particular problem (i.e., irregularly spaced missing points) is that reviewed in Section 7 of [139], primarily [54, 81, 134, 17, 56]. Some of the authors referenced there summarized and extended their work in the proceedings of the 1983 “Symposium on Time Series Analysis of Irregularly Sampled Data” [138], with several significant results presented. Outside of the econometrics literature, significant work was completed between 1960 and 1983, the year of the symposium on irregularly sampled data. Significant contributors include R.H. Jones [81, 82, 83, 84],
Emmanuel Parzen [132, 134], Peter Bloomfield [17], and H.E. Doran [54]. Of these, the most influential could be considered [134], in which Brillinger proposed a general framework for missing data time series. In this framework the observed (gappy) data is considered to be the product of two additional time series, the first an indicator function representing the available data (and known), and the second the underlying true process, which is unknown. This *amplitude modulation* format proved to be quite influential in the work that followed.

In the proceedings of the 1983 symposium mentioned earlier, three papers stand out as being particularly influential in the development of spectral estimators for irregularly sampled data. David Brillinger’s paper [19] gives an excellent summary of the ideas of inference for such problems, and furthermore gives an excellent demonstration of the underlying problem, reproduced here. Consider three processes $X$, $Y$ and $\mu$, where $X$ is the underlying process being considered, $\mu$ is a random measure determining which observations are available, and $Y$ is the observed process, i.e., the convolution of $X$ and $\mu$. In this case, when assuming that $X$ and $\mu$ are independent, that $X$ has mean 0 and that $\mu$ has mean $c_\mu$, we can write

$$f_{YY}(\lambda) = c^2_\mu f_{XX}(\lambda) + \int f_{\mu\mu}(\lambda - \alpha)f_{XX}(\alpha)d\alpha,$$

(2.4)

where $f_{XX}$, $f_{YY}$, $f_{\mu\mu}$ are the power spectral densities of the respective processes. From this expression we gain considerable intuitive understanding: firstly, the spectrum of $Y$ will exhibit many of the characteristics of the spectrum of $X$; secondly, if $c_\mu$ is not large, then considerable structure from $f_{\mu\mu}$ will exhibit in $f_{YY}$; and thirdly, this is an integral equation of the second kind. Thus, we can think of solving for $f_{XX}$ as a sort of inverse problem, and many of the tools from that field are now applicable.

There are several other developments in this paper, but the most important thing we must note is the philosophical suggestion given at the end [19, pg. 52–53]:

In the case of an ordinary time series, one might be well advised to proceed
as follows when missing values are scattered. We would prefer to work with a white noise series, hence we might: i) interpolate original values crudely, ii) prewhiten, iii) delete interpolated values and reinterpolate prewhitened series, iv) recolor as needed. (Moving to a white noise process has the advantage that it is easier to interpolate and it is easier to estimate its spectrum).

While this approach is not identical to the eventual algorithm given in Chapter 3, the philosophy is the same! We leverage the power of the Wiener forward/backward interpolator as the optimal (in the least-squares sense) interpolator for stationary stochastic processes, and much of the framework of our algorithm is concerned with the transformation to and from this state.

The second significant paper in [138] is entitled “Direct Quadratic Spectrum Estimation (DQSE) with Irregularly Spaced Data” [111], written by Donald Marquardt. From a citation search this paper has been largely ignored in the years since, a slight the author blamed [71] on a lack of publicity due to the paper being published as a book chapter rather than in a journal. Primarily, this paper suggests a spectral estimate of the form

\[ Z(\omega) = \frac{1}{T} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{i,j} Y(t_i)Y(t_j) \]  

(2.5)

for observations \((t_i, Y_i), i = 1, 2, \ldots, n\) with \(0 \leq t_i \leq T\), with the bulk of the paper spent developing the framework necessary to put this quadratic form together. Notably, the weight function is defined as \(W_{i,j} = D(\tau)F(t_i, t_j)\cos2\pi\omega(\tau)\), allowing for the use of a lag window and a “data spacing factor” that is equal to the area corresponding to the product of the \(i^{th}\) and \(j^{th}\) data steps. As an example, if we had four points \(t_i = 1, 2, 4, 5\), then

\[ F(t_2, t_3) = \left[ \frac{6}{2} - \frac{3}{2} \right] \left[ \frac{9}{2} - \frac{6}{2} \right] = (1.5)(1.5) = 2.25. \]
In a contiguous case, this factor would be $1 \cdot 1 = 1$, and it is intended to correspond to non-overlapping rectangular areas, each corresponding to a covariance ordinate.

The primary reason this work by Marquandt was largely passed over is due to self-admitted flaws in the approach. The spectral window for irregularly spaced data in this DQSE approach has two differences when compared to the default (single-taper direct estimate) spectral window. The first is that the central lobe of the window is less sharp, leading to higher variance in the estimate – which is to say, unacceptable levels of variance, as the single-taper direct estimate is barely acceptable for many physical applications. Secondly, there is the possibility for significant aliasing due to the spread of the sidelobes, as for irregularly spaced data the window may have side lobes spread throughout the frequency range. Unfortunately, these two conditions are unacceptable for serious analysis, explaining why the paper was largely ignored in further work.

The third significant paper from the symposium was presented by Elias Masry, entitled “Spectral and Probability Density Estimation From Irregularly Observed Data” [115]. The problem of having only the observations $\{X(t_k)\}_{k=1}^n$ is considered along with the case of having the observations $\{X(t_k), t_k\}_{k=1}^n$, i.e., the case of knowing the sample times. As our problem is not one of true irregular sampling, we are in the latter category, as we know exactly when the observations are missing. The work in this paper is primarily concerned with certain representations of the missing points, namely point processes $\{t_k\}$ that are “alias-free”, defined also in this paper as requiring that any two unique spectral densities also invoke unique covariance measures $\mu_Z$ for the orthogonal increment process $Z(dt) = X(t)N(dt)$. We cannot ensure this for an arbitrarily observed (i.e., gappy) time series, so the results of this paper may not always be applicable. The key result to take from this paper is that it is possible to form a new estimate of the power spectral density as a modified periodogram with a constrained window choice such that the two converge in the
same manner, but with the variance of the modified version increased from that of the default periodogram. As we already struggle with the variance properties of the periodogram, and these results only hold asymptotically, there appears to be little here of use to us.

We now digress somewhat from the work on irregularly sampled data, and return to the work on prediction and interpolation begun with Wiener [220]. It appears the two communities worked in parallel through much of the 1960–1980 timeline, the irregularly sampled data community being concerned with asymptotics and generalizations of sampling schemes, and the interpolation / prediction community being concerned with more practical data-driven techniques and extensions of the linear prediction theory of Wiener.

Emmanuel Parzen wrote a technical report [135] for the US Navy in 1967 which details much of the development to that point. The origins of much of the document are clear, tracing back to Wiener and Levinson in 1947 [220, 100, 101]. The prediction covered in this report is primarily that of the minimum mean-square error linear predictor, although the author briefly digresses into discussion of state-space approaches (Kalman filtering) and adaptive exponential smoothing. There are no new results regarding the Wiener-type linear prediction given in this report, but it does nicely summarize the development in notation and philosophy that occurred between 1942 and 1967.

For the multivariate version of the prediction problem, Wiener states that the earliest work of note was Zasuhin in 1941 [234]. This was followed closely by Wiener’s own work [220, 221] and by the excellent (if terse and complex) book of Doob [53]. Peter Whittle (mentioned earlier as a student of Wold’s) also presented work in 1953 [215] which tied together much of what had come before, deriving some relations for the “purely nondeterministic case”. Interestingly, in Whittle’s paper he briefly examines a two-series problem using separate hemisphere sunspot observations (120
half-yearly observations) using a very simple autocovariance model. This is expanded upon in a later paper [217] in Astrophysical Journal, and is a fascinating example of this type of analysis from before the age of easily accessible computing.

The papers by Zasuhin and Whittle were followed by a trio of comprehensive papers by Wiener and Masani [223, 224] and Masani alone [114]. The first paper is especially important for its extension of [220] to the vector-valued random processes setting and the careful examination of spectral analysis, also in this setting. The second of these papers continued by obtaining a linear predictor for the multivariate discrete-parameter stationary stochastic process. It is notable for establishing a “generating function” approach to the prediction problem, and (in [224, Section 6]) for deriving expressions for the linear prediction and prediction error matrix in terms of the spectral density.

Bhansali was responsible for extending the work of Wiener and Wiener-Masani in a series of papers from 1974 to 1983, [12, 13, 14, 15]. Notable achievements include the derivation of expressions for the asymptotic mean-square error for one- and many-step prediction [12, 14] the asymptotic properties of truncated versions of the linear predictor and the autoregressive spectral density function [13], and the re-expression of the prediction problem in a linear filtering setting [15]. The 1978 paper [14] is particularly recommended as a clear and comprehensive paper.

As our first example of an engineering paper, in 1975 John Makhoul wrote a tutorial review of linear prediction [107]. In this paper the author focuses on modeling the signal spectrum by a pole-zero spectrum\(^8\), with the majority of the modeling done for all-pole (autoregressive) models. Significant explanation of these terms and the context of the signal processing problem is also given, and the framing of the problem in both time- and frequency-domain is particularly valuable.

In the 1970s and early 1980s a combination of approaches became known as the

\(^8\)Familiar to those in the time-series world as the ARMA formulation of the rational spectrum.
Lomb-Scargle periodogram technique. In 1976 Nicholas Lomb pointed out that the (then in current vogue) least-squares approach to fitting sinusoids (periodocities) bore strong resemblance to the periodogram. This was extended by Jeffery Scargle of NASA [162, 163] to a definition of a periodogram with unequally spaced sample points. Unfortunately, this estimator shares all of the same problems as its source estimator, as detailed in the previous section, and is insufficient for serious analysis. Despite this it continues to be used in several applied fields to this day, in particular astronomy.

A nice reframing of the linear prediction problem into Toeplitz matrix formulation by Steven Kay [86] yields some nice intuition regarding minimum interpolation error power. The author shows that for the special case of autoregressive processes that the minimum interpolation error power occurs for prediction at the center of a span (i.e., data points on $-M, -M + 1, \ldots, -1, 1, \ldots, M - 1, M$ and interpolating $i = 0$). The results in this paper are not necessarily new, but the use of the Toeplitz matrix formulation and the interpretation gained were worth mentioning.

Battaglia and Bhansali [9] have a summary and extension paper in 1987 that nicely encapsulates the work on interpolation error variance. They further detail some findings for defining an index of linear determinism $A$. This index measures the proportion of total variance explained by the interpolation when interpolating a point $X_t$ from $\{X_s, s \neq t\}$. The paper is the first mentioned in this review that carefully explores the use of the inverse covariance function, see [69, pp. 82–85]. These inverse correlations were also studied by Cleveland [36], Chatfield [34], and Rao and Gabr [151], among many others.

The 1988 paper by Mohsen Pourahmadi [145] was influential in the development of Chapter 3. In it the author details a modification of the Kolmogorov-Wiener algorithm for interpolation that solves the problem in a finite time framework. He additionally extends the algorithm into an expectation-maximization (EM) algorithm
setting and suggests several matrix algebraic possibilities for speedy and efficient computation.

A paper by Høst-Madsen and Caspersen [80] develop a new method for estimating the spectrum of Poisson-sampled stochastic processes. They show that their new method is asymptotically unbiased, but as their method attempts to estimate the autocorrelation using an adaptive Bartlett-style technique, then applies an autocorrelation window and a Fourier transform to find the spectrum, it is obviously inconsistent and not sufficient for our purposes. The assumptions made by the authors are also highly restrictive and eliminate many of the interesting problems we would like to work with.

The same approach is taken by R.J. Martin 4 years later [113], with more framework and effort, but essentially the same results. Attempting to estimate the autocorrelation function using adaptive, specialized techniques that ignore the quadratic requirements of [120, 23, 178] seems inevitably doomed to fail for real-world mixed spectra.

Without citing any of the considerable additional work, we briefly note that a great deal of effort has gone into attempting to estimate autoregressive parameters in the presence of missing samples or the framework of irregular sampling. Unfortunately, there appear to be very few of the researchers in this field who fit their work into the greater framework of consistent estimators of the spectrum, excepting the few mentioned in the following paragraphs.

Rivoira and Fleury continue the work of Masry mentioned above [160], working within the framework of random irregularly sampled series. The framework the authors introduce restricts the system to the case of the random sampling scheme being Poisson, and the signal being Gaussian. They proceed to derive a nonparametric estimator for the power spectrum. The authors make the statement that this estimator
is consistent, as they are able to show that the variance goes to 0 as $O(N^{-2/5})$. Unfortunately, the restriction to Gaussian data and the slow convergence rate for both bias and variance make this estimator useless for our needs.

The 2005 paper of Wang, Stoica, Li and Marzetta is philosophically similar to our approach [211]. In this paper the authors consider nonparametric spectral estimation from data series with missing (arbitrarily sampled) samples. They develop two non-parametric amplitude and phase estimation algorithms which make use of the EM algorithm. The papers of Bronez [22] and Fodor and Stark [59] are mentioned and passed over due to the authors’ desire for higher resolution than the DFT can offer. This is not primarily a concern for our approach. The main connection between our method and that of these authors is the philosophical EM-approach to nonparametric estimation of the power spectrum, although we do not use the complex formulation of this paper.

Petre Stoica’s group continued to study the problem of estimating the power spectrum of nonuniformly sampled data in [177]. In this paper the authors present an adaptive least-squares periogogram (i.e., Lomb-Scargle periodogram, LSP) that iteratively weights the least-squares solution. The results are significantly improved from that of the LSP, although the spectral window is not particularly strong, and even though the estimator is iteratively adapted to provide stronger distinction between signal and noise, the algorithm does not seem to have strong broadband bias protection, as it is functionally still a non-windowed non-quadratic estimate.

There is one additional technique that has been tried that is similar philosophically to the irregular sampling work of Brillinger [19] and those that followed. In these works, the multitaper estimate is taken and the prolate spheroidal tapers modified so as to provide an orthogonal basis that has support equal to the sampling process. In particular, we note [22], [59] and [175, 176] as examples of this approach. Unfortunately, the change in taper structure appears to destroy much of the benefit typically
gained, and the broadband bias protection is significantly weakened. Finally, we note that some work has been done by K.Q. Lepage [99] on extension of the multitaper method to irregularly sampled data.

To recap, there are essentially three separate distinct approaches to estimation of power spectra in the presence of gaps:

1. Iterative interpolation:
   - Interpolate the data using some technique based on estimation of the autocovariance sequence,
   - Estimate the spectrum
   - Use the spectrum to estimate the autocovariance, iterate,

2. Direct estimation of the spectrum using specially designed windows that “work around” the gaps, or

3. The deconvolution approach that attempts to remove the effect of the gaps directly, leaving a series whose spectrum can then be estimated.

Our focus in this thesis will be on the first option, as the second appears to have intractable problems (see [176] for performance discussion) and the third similarly appears hopelessly naive for real-world data series for which assumption of smoothly varying background processes is too wrong to be useful.

2.2.1 Conclusion and Transition

We are primarily concerned in the following chapter with single-series data sets, evenly sampled but with missing values. The particular scientific problems being considered require the full length of the series be available, so sub-selection is not a viable option. As detailed in this chapter, there are a variety of possible spectral estimators
that “work around” the gaps. The majority of these estimators either do not provide sufficient broadband bias protection for our purposes, or make unwarranted assumptions about the underlying time-series model. We have thus chosen to attempt to reconstruct (interpolate) data to fill the gaps in our series in such a way as to provide an accurate estimate of the power spectrum of the underlying process.
Chapter 3

Univariate Interpolation of Time Series

As mentioned in the literature review in the previous chapter, Wiener’s 1949 book *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* [220] was augmented by Levinson’s more algorithmic approach in the second printing. These algorithms have been used to great effect for implementation of interpolators which are the best possible under certain constraints. Despite this, the exposition given in Wiener is analytic in nature, and it is easier to begin from a more computationally minded description. Our exposition in owes much to that given by Grenander and Rosenblatt [69] and especially that of Brockwell and Davis [21, §2.7, 5.1–5.3].

### 3.1 Hilbert Space Background

For background, begin with a probability space $(\Omega, \mathcal{F}, P)$ and consider the collection $C$ of random variables $X$ defined on $\Omega$ that have finite second moment. This collection is a vector space. Define an inner product on $X, Y \in C$ as

$$
\langle X, Y \rangle = E\{XY\},
$$

(3.1)
with \( E \) denoting statistical expectation. This definition does not quite satisfy the requirements of an inner product, as \( \langle X, X \rangle = 0 \) gives only \( P(X = 0) = 1 \), not \( X(\omega) = 0 \ \forall \omega \), so divide \( C \) into equivalence classes such that the random variables \( X \) and \( Y \) are equivalent if \( P(X = Y) = 1 \). Then, the collection of these equivalence classes with inner product as given is denoted as \( L^2(\Omega, \mathcal{F}, P) = L^2 \). The proof of completeness under the norm given is provided in [21], and with this completeness, \( L^2 \) is a Hilbert space.

Now define \( \mathcal{M} \) as a closed subspace of \( L^2 \) (and hence itself a Hilbert space) and take \( Y \in L^2 \). In this case the best mean square predictor of \( Y \) in \( \mathcal{M} \) is the random variable \( \hat{Y} \) such that

\[
\left\| Y - \hat{Y} \right\|^2 = \inf_{Z \in \mathcal{M}} \| Y - Z \|^2 = \inf_{Z \in \mathcal{M}} E |Y - Z|^2. \tag{3.2}
\]

It is debatable whether “best” should mean “minimizes the mean squared error”, but for our purposes this is a nice formulation that provides a closed form solution that is computationally tractable. The Projection Theorem provides us with a unique solution for this minimization.

**Theorem 1** (Thm. 2.3.1, [21]). If \( \mathcal{M} \) is a closed subspace of the Hilbert space \( L^2(\Omega, \mathcal{F}, P) \) and \( X \in L^2 \), then

(i) there is a unique element \( \hat{X} \in \mathcal{M} \) such that

\[
\left\| X - \hat{X} \right\| = \inf_{y \in \mathcal{M}} \| X - y \|, \quad \text{and}
\]

(ii) \( \hat{X} \in \mathcal{M} \) and \( \left\| X - \hat{X} \right\| = \inf_{y \in \mathcal{M}} \| X - y \| \) if and only if \( \hat{X} \in \mathcal{M} \) and \( (X - \hat{X}) \in \mathcal{M}^\perp \).

Note that \( \mathcal{M}^\perp = \{ y \in L^2(\Omega, \mathcal{F}, P) | E \{ xy \} = 0 \ \forall x \in \mathcal{M} \} \). Define by \( \mathcal{M}(Z_1, Z_2, \ldots, Z_n) \) the closed subspace of \( L^2 \) consisting of all random variables \( X \in L^2 \) of the form \( \phi(Z_1, \ldots, Z_n) \) for some Borel function \( \phi : \mathbb{R}^n \to \mathbb{R} \) and define by \( E_{\mathcal{M}(Z_1, \ldots, Z_n)} \) the
conditional expectation

\[ \mathbf{E} \{X | Z_1, \ldots, Z_n\} = \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} . \]  \hspace{2cm} (3.4)

Then, by the Projection Theorem, this conditional expectation \( \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} \) is the best mean square predictor of \( X \) in \( \mathcal{M}(Z_1, \ldots, Z_n) \). However, the determination of a projection onto \( \mathcal{M}(Z_1, \ldots, Z_n) \) is quite difficult as \( \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} \) is the unique random variable in \( \mathcal{M} \) such that

\[ \mathbf{E} \{W \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} \} = \mathbf{E} \{WX\} \text{ for all } W \in \mathcal{M}. \]  \hspace{2cm} (3.5)

This formulation is quite complex, as the structure of \( \mathcal{M}(Z_1, \ldots, Z_n) \) may not be readily accessible. If instead \( Z_1, \ldots, Z_n \in L^2 \), then we can instead consider the span (\( \mathcal{sp} \)) of these random variables augmented with the constant 1:

\[ \mathcal{H}_n = \mathcal{sp} \{1, Z_1, \ldots, Z_n\} \subseteq \mathcal{M}(Z_1, \ldots, Z_n). \]  \hspace{2cm} (3.6)

This is much easier, as we can then write the projection as

\[ P_{\mathcal{sp}\{1,Z_1,\ldots,Z_n\}}(X) = \sum_{i=0}^n \alpha_i Z_i, \text{ with } Z_0 = 1 \]  \hspace{2cm} (3.7)

and restriction

\[ \sum_{i=0}^n \alpha_i \mathbf{E} \{Z_i Z_j\} = \mathbf{E} \{XZ_j\}, \text{ } j = 0, 1, \ldots, n. \]  \hspace{2cm} (3.8)

for some \( \alpha_i, i = 0, \ldots, n \) scalars.

Now, the Projection Theorem guarantees that a solution exists, and when substituted into Eqn. 3.7 gives the best linear predictor of \( X \) in terms of 1, \( Z_1, \ldots, Z_n \). As this projection is onto a subspace of \( \mathcal{M}(Z_1, \ldots, Z_n) \), it cannot have smaller mean squared error than \( \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} \), but it has three main advantages:

(a) Computation is easier than \( \mathbf{E}_{\mathcal{M}(Z_1, \ldots, Z_n)} \{X\} \),

(b) It depends only on the first and second-order moments of the joint distribution of \( (X, Z_1, \ldots, Z_n) \).
(c) If \((X, Z_1, \ldots, Z_n)\) has a multivariate normal distribution, then the two are equal:

\[ P_{\mathcal{F}(1,Z_1,\ldots,Z_n)} \{ X \} = \mathbb{E}_{\mathcal{M}(Z_1,\ldots,Z_n)} \{ X \}. \] (3.9)

### 3.1.1 Prediction Equations for Stationary Processes

Begin with the one-step forward prediction problem and let \(\hat{X}_{n+1}, n \geq 0\) denote the one-step predictors, defined theoretically by the recursive relationship

\[ \hat{X}_{n+1} = \begin{cases} 0 & \text{if } n = 0 \\ P_{\mathcal{H}_n}(X_{n+1}) & \text{if } n \geq 1. \end{cases} \] (3.10)

Now, since by construction \(\hat{X}_{n+1} \in \mathcal{H}_n, n \geq 1\), write

\[ \hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}, \quad n \geq 1, \] (3.11)

where the \(\phi_{n1}, \ldots, \phi_{nn}\) satisfy the prediction equations (3.7) and (3.8). These equations can be rewritten by applying the linearity of the expectation operator as

\[ \sum_{j=1}^{n} \phi_{nj} \gamma(i-j) = \gamma(j), \quad j = 1, \ldots, n, \] (3.12)

or equivalently in matrix formulation as

\[ \Gamma_n \phi_n = \gamma_n \] (3.13)

with \(\Gamma_n = [\gamma(i-j)]_{i,j=1,\ldots,n}\), \(\gamma_n = (\gamma(1), \ldots, \gamma(n))^T\) and \(\phi_n = (\phi_{n1}, \ldots, \phi_{nn})^T\). This matrix equation has one solution if and only if \(\Gamma_n\) is non-singular, for which the following condition provides sufficiency.

**Theorem 2** (Prop. 5.1.1, [21]). *If \(\gamma(0) > 0\) and \(\gamma(h) \to 0\) as \(h \to \infty\) then the covariance matrix \(\Gamma_n\) of \((X_1, \ldots, X_n)^T\) is non-singular for every \(n\).*

Under this condition the covariance matrix \(\Gamma_n\) is non-singular, and thus is invertible, giving the unique solution

\[ \hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}, \quad n = 1, 2, \ldots \] (3.14)

for \(\phi_n = (\phi_{n1}, \ldots, \phi_{nn})' = \Gamma_n^{-1} \gamma_n\). In this notation, the \(\phi_{nj}\) are related to the partial
autocorrelation coefficients, with the \( n \) indicating the total data length (and fixed), and the \( i \) the indexing term. See [21, pg. 101].

There are at least two algorithms available for recursive computation of the \( \phi \) coefficients. However, with careful implementation of matrix computation code it is possible to forego much of the computational burden involved in computing the matrix inverse \( \Gamma^{-1}_n \) and thus the implementation of this algorithm in our interpolator will be in the format given above.

### 3.1.2 Extending the Algorithm to Interpolation

The extension of this one-step prediction algorithm to many-step or to interpolation is quite straightforward. Consider \( \{X_{-n}, \ldots, X_{-1}, X_1, \ldots, X_n\} \) a stationary stochastic process with one missing observation \( X_0 \). Now, assuming that \( \gamma(h) \) (the Autocovariance Function, or ACVF) is known, we can establish similar equations to Eqn. 3.14. Write

\[
\hat{X}_0 = \sum_{j=-n}^{n} \alpha_j X_j
\]

Then, as before, recall the secondary condition on this estimate, namely that

\[
E\{\hat{X}_0, X_j\} = E\{X_0, X_j\}, \quad j = -n, \ldots, -1, 1, \ldots, n.
\]

By the linearity of expectation, we rewrite this as

\[
\sum_{j=-n}^{n} \alpha_j E\{X_j, X_j\} = E\{X_0, X_j\}, \quad j = -n, \ldots, -1, 1, \ldots, n
\]

\[
\sum_{j=-n}^{n} \alpha_j \gamma(i - j) = \gamma(j).
\]

Thus, as before, we have a number of linear equations (in this case, \( 2n \) of them) in the ACVF lags, and we can rewrite this in matrix form as

\[
\Gamma \alpha = \gamma
\]
where $\mathbf{\Gamma} = [\gamma(i-j)]_{i,j=-n,\ldots,-1,1,\ldots,n}$ and $\gamma = (\gamma(n), \gamma(n-1), \ldots, \gamma(1), \gamma(1), \ldots, \gamma(n))'$. It may seem as if the symmetry of the equations would lead to duplicated rows, but this is not the case, and the full $2n \times 2n$ $\mathbf{\Gamma}$ matrix and $2n$-element $\gamma$ vector are necessary for a full solution. There is, however, symmetry in the solution. There is additionally a reduction that can be done that uses the expression

$$\gamma_i = \sum_{j=1}^{n} \alpha_j [\gamma(i-j) + \gamma(i+j)]$$

although this is not strictly necessary unless working with very large $n$.

The extension from a one-point interpolator to a many-point interpolator is similarly straightforward. The hardest part of the implementation is ensuring that the “tags” (i.e., the vector of differences between the point to be interpolated and the non-missing points) is properly established and converted to all-positive lags so as to match with the symmetric ACVF. Code has been implemented that does all of this and is available in the package tsinterp for R; see Appendix B.

### 3.1.3 Estimating Error in the Gaussian Case

When dealing with a zero-mean wide-sense stationary Gaussian process with ACVF $\gamma(\cdot)$ for which Theorem 2 holds, the best linear solution for an $h$-step prediction is

$$P_nX_{n+h} = \gamma_n \Gamma_n^{-1}X^T_n$$

for $\gamma_n = (\gamma(n+h-1), \gamma(n+h-2), \ldots, \gamma(h))$, $\Gamma_n$ the $n \times n$ covariance matrix for $X_n$, and $X_n = (X_1, X_2, \ldots, X_n)$. This can be rewritten as

$$P_nX_{n+h} = \mathbf{E}\{X_{n+h}|X_1, \ldots, X_n\}.$$  \hspace{1cm} (3.21)

Then the prediction error

$$\Delta_n(h) = X_{n+h} - P_nX_{n+h} = X_{n+h} - \mathbf{E}\{X_{n+h}|X_1, \ldots, X_n\}$$

where $\mathbf{\Gamma} = [\gamma(i-j)]_{i,j=-n,\ldots,-1,1,\ldots,n}$ and $\gamma = (\gamma(n), \gamma(n-1), \ldots, \gamma(1), \gamma(1), \ldots, \gamma(n))'$. It may seem as if the symmetry of the equations would lead to duplicated rows, but this is not the case, and the full $2n \times 2n$ $\mathbf{\Gamma}$ matrix and $2n$-element $\gamma$ vector are necessary for a full solution. There is, however, symmetry in the solution. There is additionally a reduction that can be done that uses the expression

$$\gamma_i = \sum_{j=1}^{n} \alpha_j [\gamma(i-j) + \gamma(i+j)]$$

although this is not strictly necessary unless working with very large $n$.

The extension from a one-point interpolator to a many-point interpolator is similarly straightforward. The hardest part of the implementation is ensuring that the “tags” (i.e., the vector of differences between the point to be interpolated and the non-missing points) is properly established and converted to all-positive lags so as to match with the symmetric ACVF. Code has been implemented that does all of this and is available in the package tsinterp for R; see Appendix B.

### 3.1.3 Estimating Error in the Gaussian Case

When dealing with a zero-mean wide-sense stationary Gaussian process with ACVF $\gamma(\cdot)$ for which Theorem 2 holds, the best linear solution for an $h$-step prediction is

$$P_nX_{n+h} = \gamma_n \Gamma_n^{-1}X^T_n$$

for $\gamma_n = (\gamma(n+h-1), \gamma(n+h-2), \ldots, \gamma(h))$, $\Gamma_n$ the $n \times n$ covariance matrix for $X_n$, and $X_n = (X_1, X_2, \ldots, X_n)$. This can be rewritten as

$$P_nX_{n+h} = \mathbf{E}\{X_{n+h}|X_1, \ldots, X_n\}.$$  \hspace{1cm} (3.21)

Then the prediction error

$$\Delta_n(h) = X_{n+h} - P_nX_{n+h} = X_{n+h} - \mathbf{E}\{X_{n+h}|X_1, \ldots, X_n\}$$

where $\mathbf{\Gamma} = [\gamma(i-j)]_{i,j=-n,\ldots,-1,1,\ldots,n}$ and $\gamma = (\gamma(n), \gamma(n-1), \ldots, \gamma(1), \gamma(1), \ldots, \gamma(n))'$. It may seem as if the symmetry of the equations would lead to duplicated rows, but this is not the case, and the full $2n \times 2n$ $\mathbf{\Gamma}$ matrix and $2n$-element $\gamma$ vector are necessary for a full solution. There is, however, symmetry in the solution. There is additionally a reduction that can be done that uses the expression

$$\gamma_i = \sum_{j=1}^{n} \alpha_j [\gamma(i-j) + \gamma(i+j)]$$

although this is not strictly necessary unless working with very large $n$.
is normally distributed with mean zero and variance

\[ \sigma^2_n(h) = E\{(\Delta_n(h))^2\}. \]  

(3.23)

This is computed using the autocovariance matrix \( \kappa(i,j) = E\{X_iX_j\} \) \cite[pp.173–175] as:

\[ \sigma^2_n(h) = ||X_{n+h}||^2 - ||P_nX_{n+h}||^2 \]

\[ = \kappa(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta^2_{n+h-1, j} v_{n+h-j-1}, \]  

(3.24)

with the \( \theta_{n,m} \) the coefficients in the recursive innovations expansion, and the \( v_n \) the mean-squared errors from previous steps. Unfortunately, this method is recursive: we would rather a closed-form solution for the errors as the rest of the algorithm we use is non-recursive. For 1-point prediction / interpolation, the derivation below for Kolmogorov’s formula suffices.

Thus, in the case of an underlying zero-mean wide-sense stationary Gaussian process there exist confidence intervals such that \( X_{n+h} \) lies between the bounds

\[ P_nX_{n+h} \pm \Phi_{1-\alpha/2}\sigma_n(h) \]  

(3.25)

with probability \( 1 - \alpha \). Unfortunately, restricting ourselves to the case where we can realistically expect the underlying process to be both truly stationary and also Gaussian seems to be too conservative for most practical examples.

### 3.1.4 Kolmogorov’s Formula

More generally, consider a real-valued zero-mean stationary process \( \{X_t\} \) with spectral distribution function \( F_X \) with associated density \( f = F'_X \). The spectral density function is assumed to exist and be both continuous and strictly positive. Furthermore, for \( X_t \) real we have that \( f(-\lambda) = f(\lambda), 0 \leq \lambda \leq \pi \). The following result was first derived by Szegö \cite{Szego} (in 1939) for the absolutely continuous case, and was extended to the general case by Kolmogorov \cite{Kolmogorov} (in 1941). The full sequence of papers
which led to this derivation are [92, 93, 94], with the last a translation of the original Russian paper.

**Theorem 3** (Thm. 5.8.1, [21]). *The one-step mean-square prediction error of the stationary process \( \{X_t\} \) is*

\[
\sigma^2_E(1) = 2\pi \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln f(\lambda) d\lambda \right\} = \exp \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln (2\pi f(\lambda)) d\lambda \right\}. \quad (3.26)
\]

This can be rewritten on the domain \([-1/2, 1/2]\) as

\[
\sigma^2_E(1) = \exp \left\{ \int_{-1/2}^{1/2} \ln f(\lambda) d\lambda \right\}. \quad (3.27)
\]

Further derivational notes from [94] give that

\[
\sigma^2_E(m) = \exp \left\{ \frac{1}{\pi} \int_0^\pi \ln f(\lambda) d\lambda \right\} \cdot (1 + r_1^2 + \cdots + r_m^2) \quad (3.28)
\]

where the \( r_s \) terms are determined from

\[
\exp \left\{ a_1\zeta + a_2\zeta^2 \right\} = 1 + r_1\zeta + r_2\zeta^2 + \cdots. \quad (3.29)
\]

and the \( a_k \) are the Fourier expansion coefficients

\[
a_k = \frac{1}{\pi} \int_0^\pi \cos(k\lambda) \ln f(\lambda) d\lambda. \quad (3.30)
\]

As is shown by Kolmogorov [94] and further detailed by Yaglom [230], the *interpolation* variance has a simpler form:

\[
\sigma^2_I(1) = \frac{1}{\int_{-1/2}^{1/2} \frac{1}{S(f)} df} = \left( \int_{-1/2}^{1/2} S^{-1}(f) df \right)^{-1}, \quad (3.31)
\]

which is written on the original \((-\pi, \pi]\) domain as

\[
\sigma^2_I(1) = \frac{(2\pi)^2}{\int_{-\pi}^{\pi} f_{xx}(\lambda) d\lambda} = \frac{\pi}{\int_0^\pi w(\lambda) d\lambda}. \quad (3.32)
\]

In this last, \( w(\lambda) = 2\pi f_{xx}(\lambda) \).

Finally, we note that this formulation is not directly applicable to most problems because assuming that the spectrum is continuous is far from true. In the simulations that follow, we will apply these formulas only to the continuous background portion of
the spectrum, essentially ignoring the error in the estimation of the line components as being intractable. We further will ignore the fact that the 1-point interpolation variance applies only to: a) asymptotic choice of \( n \) (which is only approximately attained), and b) one-point gaps. Certainly as gap length goes out the variance of the errors will increase in similar fashion to the prediction variance (Eqn. 3.27), scaling with the scaled amplitudes of the prediction coefficients. This is a topic for further work.

3.1.5 Notes

The prediction algorithm detailed in this section provides the best linear predictor of the value at a given missing time. However, the restrictions on this algorithm are more severe than may be apparent from the development. We require that the non-missing elements of the process be in a Hilbert space \( \mathcal{H} \). By its construction, this implies that the process must be wide-sense (covariance) stationary. This further implies that the process has constant mean and variance and has covariance that is a function only of the time lag between the two points. Truly wide-sense stationary time series data are rare, so the algorithm detailed in the following section is constructed specifically to address this problem.

3.2 Interpolation Algorithm

As mentioned at the conclusion of the previous section, the technique developed in [211] is philosophically similar to our approach. The algorithm of that paper and the following algorithm can generally be thought of as Expectation-Maximization (EM) algorithms [46]. Loosely, given a data series with missing values, our \( E \) (expectation) step consists of estimation of the multitaper power spectrum of a copy of this series with the missing values filled in some way. This multitaper power spectrum is then
inverted through the Einstein-Wiener-Khintchine theorem to obtain a consistent estimate of the ACVF of the process, although care must be taken [118, 72]. The M (maximization) step consists of the estimation of the missing values of the series using this ACVF. The bulk of the computational burden in this algorithm resides in the M step, and the convergence properties of this are discussed in the following.

3.2.1 E Step

The E step is simple given a realization of the time series \( \{x_t\}_{t=0}^{N-1} \) where observations are assumed to be contiguous. Estimation of the power spectrum is done through the multitaper technique [194], formally

\[
\hat{S}(f) = \frac{1}{K} \sum_{k=0}^{K-1} \left| \sum_{t=0}^{N-1} x_t v_t^{(k)} e^{-i2\pi ft} \right|^2
\]

(3.33)

where the \( \{v_t^{(k)}\}_{k=0}^{K-1}, t = 0, \ldots, N-1 \) are discrete prolate spheroidal sequences (or Slepians for short), with bandwidth chosen to be concentrated on an “inner bandwidth” \((-W, W)\). Typically for these problems \( W \) will be chosen to be \( \sim 4/N \). This spectrum estimation is straightforward and well-understood, and the reader unfamiliar with its properties is directed to [194] for the genesis of the method, [143, 148] for practical implementation details, and [196, 195, 200] for more advanced techniques. Note that implementations of the multitaper method are available in MATLAB[103], R [148], C[98] and F90[147].

We note here that in practice our M step actually contains numerous copies of this E step. In fact, this entire algorithm could be rephrased as multiple individual EM-like steps that are chained together in a global EM-like algorithm.
3.2.2 M Step

The M (maximization) step is significantly more complicated than the E step. Obtaining the ACVF from the E step is necessary for only a portion of the M step, as we cannot assume “fairy tales” (i.e., wide-sense stationarity, Gaussian data or noise, etc.) about our series of interest.

Consider a classic model for a time series:

\[ X_t = M_t + T_t + \xi_t \]  

(3.34)

where \( M_t \) is a fully deterministic polynomial function of time, representative of the mean of the series; \( T_t \) is a fully deterministic finite linear combination of sinusoids, i.e., a trend function; and \( \xi_t \) is a generalized function of a zero-mean wide-sense stationary stochastic process. The last term, \( \xi_t \), is taken for our considerations to be an auto-regressive (AR) process of undefined order. This model is remarkably flexible, allowing for the inclusion of multiple sinusoids (line components) in the \( T_t \) trend term, a non-zero time-varying mean in the \( M_t \) mean term, and a moderately complicated time series model in the \( \xi_t \) process term. This allows for extension beyond the simple case of a stationary stochastic process to real-world data sets containing trends, line components, seasonal mean variation (i.e., a yearly periodic mean), and other interesting and useful features.

In practice, the estimation of these separate terms is philosophically similar to Cleveland’s STL procedure [35], which decomposes a time series into trend, seasonality and remainder using an iterative backfitting algorithm and smoothers [74]. Unlike the STL procedure, we do not estimate the seasonality term as discretely separate elements – we are more interested in a smooth resultant function that is a linear combination of a number of different cosinusoids.

Consider as a working example the series \( \{X_t\}_{t=1}^{1000} \) where

\[ X_t = \left( \frac{t}{500} \right)^2 + \sin \left( 2\pi \frac{5 \cdot t}{1000} \right) + \sum_{k=1}^{3} \sin \left( 2\pi \frac{k \cdot t}{10} \right) + \xi_t \]  

(3.35)
for $\zeta_t \sim \mathcal{N}(0, 1)$. A realization of this series is displayed in Figure 3.1. For the moment, this series is simulated without any missing values – we are interested in how we should go about estimating the different components of the series, namely: the degrees and coefficients of the $M_t$ term; the amplitudes and phases of the $T_t$ cosinusoids; and an estimate of the realization of $\zeta_t$.

### 3.2.3 $M_t$: The Mean Term

To begin, the first step will always be to estimate the (zeroth order) trend, that is, the mean. We will assume that $M_t$ is actually itself composed of two parts, the first a constant non-varying mean element, and the second a varying polynomial trend element. This algorithm could be rephrased to move the polynomial trend components to the $T_t$ trend term, but there are certain advantages to dividing the components this way that will hopefully become clear as we proceed.
There are any number of ways available to use to estimate the mean: we choose
the one that is most conducive to spectrum analysis as that is our eventual goal.
The mean is estimated through use of the zeroth-frequency transfer functions of
the Slepian sequences [194, 198]. Formally, given a series \( \{X_t\}_{t=0}^{N-1} \) and given user-
selected parameters \( NW \) and \( K \leq 2NW \), we generate a set of Slepian sequences
\( \{ v_t(k) \}_{k=0}^{K-1}, t = 0, \cdots, N - 1 \). Now, consider
\[
U_k(f) = \sum_{t=0}^{N-1} v_t(k) e^{-i2\pi ft} \tag{3.36}
\]
and upon setting \( f = 0 \) obtain \( U_k(0) = \sum_{t=0}^{N-1} v_t(k) \). For the odd-order Slepians this
sum is 0, so we write
\[
S_w(0) = \sum_{k=0,2,4,\cdots}^{K-1} U_k^2(0) \tag{3.37}
\]
Next, compute the eigencoefficients (eigencoefficients are the Fourier transform of the
tapered data series, and are individually direct estimates of the spectrum):
\[
y_k(f) = \sum_{t=0}^{N-1} x_t v_t(k) e^{-i2\pi ft}, \tag{3.38}
\]
and evaluate at \( f = 0 \) to obtain \( y_k(0) \), \( k = 0, \ldots, K - 1 \). Then compute the mean
estimate:
\[
\mu = \frac{\sum_{k=0,2,4,\cdots}^{K-1} U_k(0) \cdot y_k(0)}{S_w(0)} = \frac{\sum_{k=0}^{K-1} U_k(0) \cdot y_k(0)}{\sum_{k=0,2,4,\cdots}^{K-1} U_k^2(0)} \tag{3.39}
\]
This can be considered to be a special case of the harmonic F-test statistic restricted
to frequency \( f = 0 \) (see [194] for more details).

Extending this technique to allow for linear or higher polynomial order trends is
also straightforward. Following the derivation given in [198], begin by defining as
\( R_j(n) \) a polynomial of degree \( j \) sampled on a contiguos discrete mesh. Expanding
this polynomial on a Slepian projection basis as above, we obtain

$$U_{k,j} = \sum_{n=0}^{N-1} v_n^{(k)} \cdot R_j(n),$$

(3.40)

with orthogonality constraint

$$\sum_{k=0}^{K-1} U_{k,i} \cdot U_{k,j} = \delta_{i,j}. \quad (3.41)$$

To determine these $R_j(n)$ polynomials, begin with the $K = \lfloor 2NW \rfloor$ lowest-order Slepian sequences $v_n^{(k)}$ in matrix form:

$$\mathbb{V} = \begin{bmatrix}
    v_0^{(0)} & v_0^{(1)} & \cdots & v_0^{(K-1)} \\
    v_1^{(0)} & v_1^{(1)} & \cdots & v_1^{(K-1)} \\
    \vdots & \vdots & \ddots & \vdots \\
    v_{N-1}^{(0)} & v_{N-1}^{(1)} & \cdots & v_{N-1}^{(K-1)}
\end{bmatrix}, \quad (3.42)$$

which has orthogonality property

$$\mathbf{I}_K = \mathbb{V}^T\mathbb{V}. \quad (3.43)$$

Similarly, define

$$\mathbb{R} = \begin{bmatrix}
    R_0(0) & R_1(0) & \cdots & R_{P-1}(0) \\
    R_0(1) & R_1(1) & \cdots & R_{P-1}(1) \\
    \vdots & \vdots & \ddots & \vdots \\
    R_0(N-1) & R_1(N-1) & \cdots & R_{P-1}(N-1)
\end{bmatrix}, \quad (3.44)$$

the $N \times P$ matrix composed of polynomials of orders 0 through $P - 1$ evaluated at $t = 0, 1, \ldots, N - 1$. Note that these polynomials $R_j(n)$ are not simple polynomials like $R_1(n) = n$ or $R_2(n) = n^2$. The $R_j(n)$'s must satisfy $U^T U = \mathbf{I}_p$, for $U = \mathbb{V}^T \mathbb{R}$, so careful construction via the Gram-Schmidt procedure is necessary.

Further define a data block of $N$ samples as $\mathbf{X}$ (size $N \times k$) and the corresponding eigencoeficients as

$$\mathbf{Y} = \mathbb{V}^T \mathbf{X}. \quad (3.45)$$
Separate these eigencoefficients into polynomial
\[ a = U^T Y, \quad (3.46) \]
and residual terms
\[ r = Y - Ua. \quad (3.47) \]
The degree \( P - 1 \) reconstruction is then
\[ \hat{X} = Vr + Ra. \quad (3.48) \]

Figure 3.2: Polynomial (order 2) fit (in red) to the example (in black) presented in Eqn. 3.35. The blue dashed line represents the true polynomial portion of this series.

As an example, consider \( P = 3 \) so that we are fitting a quadratic function to our series. As the data we are examining is synthetic, we can obviously identify the correct choice of \( P \) \textit{a priori}. In a real-world data analysis, this will not be possible. A possible extension of the algorithm we are describing would do a search on the space of possible polynomials to determine the best option, or possibly use
additional information about the underlying process to determine a reasonable choice. This is a weakness of this algorithm, although as the $M_t$ term is included mainly to approximately zero-mean the series, the difference between a quadratic and a cubic is not all that significant.

Return to the example begun in Eqn. 3.35 and construct $V$, $R$, and $Y$. The resultant polynomial portion (given as $Ra$ above) shown in Figure 3.2 is presented as the red line, with the blue dashed line indicating the true polynomial portion of the example. Note how the quadratic fit is quite close to the true “climbing” effect of the true polynomial underlying trend. This ideal example does not illustrate all of the difficulty one often encounters with noisy data, in that it is almost impossible to differentiate a slowly varying sinusoid from a polynomial! The slowly varying sinusoid included in the example is far enough from 0 frequency so as to not interfere with this fit – in general, if our series were to contain a sinusoid with period approximately the same length as the data length, the fit will not be as reliable. However, the algorithm we will develop in the next section will remove the trend components either in this stage or the next, so the result will still have the desired properties.

### 3.2.4 $T_t$: The Trend Term

The next stage of the fitting process is to take the (approximately) zero-mean residual series after removing $M_t$ and to estimate sinusoidal trends in it. It will be clear to the experienced reader that our usage of trend and mean is somewhat fuzzy – this is due to the allocation of “mean and things relating to the change in mean level” to $M_t$ and the allocation of “trends that vary periodically” to $T_t$.

There are many techniques available in the literature for detection and estimation of sinusoids in the presence of noise. The technique we will use for detection of line components is a spectrum-based algorithm known as the harmonic F-test statistic,
first published by Thomson [194]. This statistic is a variance ratio test with a null hypothesis of no signal present, and relies on an assumption of local “whiteness”. The derivation is very similar to that given in the previous section for estimation of the mean.

Begin with contiguous observations of a time series \( \{X_t\}_{t=0}^{N-1} \), assumed to have a stationary background with one or more line components additively embedded in it. Estimate the eigencoefficients for this series with under parameters \( NW \) and \( K \):

\[
y_k(f) = \sum_{t=0}^{N-1} X_t v_t^{(k)} e^{-i2\pi ft}.
\]  

(3.49)

Defining, as before, the spectral windows \( U_k(f) \), write

\[
U_k(f) = \sum_{t=0}^{N-1} v_t^{(k)} e^{-i2\pi ft},
\]  

(3.50)

and then

\[
\hat{\mu}(f) = \frac{\sum_{k=0}^{K-1} U_k(0)y_k(f)}{\sum_{k=0}^{K-1} U_k^2(0)},
\]  

(3.51)

where the only difference from (3.36) and (3.39) is that \( y_k \) is defined at a general \( f \) rather than \( f = 0 \). The variance of this estimated mean depends on the local continuous part of the spectrum [194],

\[
\text{var} \{\hat{\mu}(f)\} = \frac{S(f)}{\sum_{k=0}^{K-1} U_k^2(0)},
\]  

(3.52)

and subtracting the estimated means from the eigencoefficients gives an estimate of the continuous spectrum. Comparing the two values gives a variance-ratio test [58] with an \( F(2, 2K - 2) \) distribution for the significance of an estimated line component:

\[
F(f) = \frac{(K - 1) |\hat{\mu}(f)|^2 \sum_{k=0}^{K-1} U_k(0)^2}{\sum_{k=0}^{K-1} |y_k(f) - \hat{\mu}(f)U_k(0)|^2}.
\]  

(3.53)
The reader may note that earlier we discussed the implausibility of truly Gaussian
data, yet here we glibly assume normality for the construction of a chi-squared ra-
tio. This formulation is acceptable due to the Fourier transformation applied to the
original data, see [158, 159] for background and [108, 109] for full details.

In practice the $F$-test is extremely precise, often identifying central frequencies for
line components at precisions exceeding the chosen zero-padded frequency resolution\(^1\). For this reason, when using this test to find significant sinusoids in a series the zero-
padding level is adaptively modified to allow for increased precision.

\begin{center}
\begin{tabular}{|c|}
\hline
$T_t$ Algorithm \\
\hline
1. Compute a pilot estimate of the spectrum and harmonic $F$-test. \\
2. Identify possible significant line components. \\
3. For each line component: \\
\hspace{1cm} (a) Determine a choice of zero-padding using low-order prime powers such \\
\hspace{1cm} that the estimated center of the line component falls directly in the \\
\hspace{1cm} middle of a bin. \\
\hspace{1cm} (b) Re-compute the spectrum estimate and $F$-test using this refined zero-
\hspace{1cm} padding amount. \\
\hspace{1cm} (c) If the center frequency has moved more than some $\epsilon$, repeat until con-
\hspace{1cm} vergence. \\
4. Estimate each line component in the time or frequency domain, obtaining \\
\hspace{1cm} estimates of the amplitude and phase from the (complex) eigencoefficients. \\
\hline
\end{tabular}
\end{center}

This algorithm is necessary for removal of a line component in the frequency
domain. For example, consider a toy example with a single sinusoid at frequency
$f_0 = 0.0998$ Hz, sampled 1000 times at 1 Hz in the presence of noise. The Nyquist
frequency for this example is 0.5 Hz, and a typical number of FFT points would be
2048. This gives a Fourier “bin width” of $0.488 \text{mHz}$. The closest bins to 0.0998 Hz
\(^1\)In fact, the harmonic $F$-test has been shown to come close to attaining the Cramér-Rao Bound
[201].

41
are at 99.6094m and 100.098mHz. If we were to choose the center frequency for this periodicity, it would be at the maximum of the F-test, which will occur at the closest bin, i.e., 99.6094mHz. This is 0.191mHz off from the actual frequency, and across 1000 samples is enough to induce a phase drift of ~ 0.2 cycles. This problem scales up with N, although an increase in N also decreases the bin error.

If instead we were to zero-pad adaptively (i.e., in an iterative fashion) until the maximized F-test estimate of the center frequency converged to within some ε, say ε = 1e − 10, we can obtain significantly improved estimates of central frequency. As our estimates of amplitude and phase (and eventually, the entire Tt estimate) are dependent upon the choices of these parameters, it seems to be worth the time to estimate these frequencies accurately.

Now, consider (again, without any missing data) estimation of the frequencies for the example given in Eqn. 3.35. The pilot harmonic F-test is shown in Figure 3.3. Note the 4 peaks above 99.9% significance level. With the (N · Δt)−1 = 1mHz resolution available to us we can easily differentiate the 0.005, 0.1, 0.2 and 0.3Hz signals in the data series. Iterating the algorithm above, we obtain the results of Table 3.1.

<table>
<thead>
<tr>
<th>Sinusoid</th>
<th>Frequency True</th>
<th>Est</th>
<th>Amplitude True</th>
<th>Est</th>
<th>Phase (Rad) True</th>
<th>Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.00mHz</td>
<td>5.00mHz</td>
<td>1.0</td>
<td>0.929</td>
<td>0.00</td>
<td>-0.041</td>
</tr>
<tr>
<td>2</td>
<td>100.00mHz</td>
<td>100.01mHz</td>
<td>1.0</td>
<td>1.009</td>
<td>0.00</td>
<td>-0.043</td>
</tr>
<tr>
<td>3</td>
<td>200.00mHz</td>
<td>199.95mHz</td>
<td>1.0</td>
<td>1.033</td>
<td>0.00</td>
<td>+0.258</td>
</tr>
<tr>
<td>4</td>
<td>300.00mHz</td>
<td>299.99mHz</td>
<td>1.0</td>
<td>1.046</td>
<td>0.00</td>
<td>+0.068</td>
</tr>
</tbody>
</table>

Table 3.1: Estimates of four sinusoids from a realization of Eqn. 3.35.

Note that there are two equally valid approaches to estimation of the amplitude and phase coefficients given a frequency estimate. The first is the least-squares solution, which suffers from global bias from the noise that is also in the series, and also has end-point problems [3]. The second involves estimating the power spectral
density and then inverting the density to obtain the original series. If one masks the spectral density so as to isolate a single sinusoid (and its complex conjugate pair), it is possible to obtain a spectral estimate of the parameters of a sinusoid. This is what is done in Table 3.1.

### 3.2.5 Sub-Iteration

Once the sinusoidal components of the series have been estimated, we iterate the two steps (estimate $M_t$, estimate $T_t$) on the residuals after accounting for the other component. This is a form of *backfitting algorithm* such as can be used for Generalized Additive Models [74], and will be discussed further in later chapters. Iteration proceeds to convergence under a choice of metric. Our convergence metric will be defined as the maximum pointwise difference between successive iterations of the overall
backfit. That is, if at step \( j \) we have estimates \( T^{(j)}_t \) and \( M^{(j)}_t \), then iterate until

\[
\max \left\{ \max_t \left| M^{(j)}_t - M^{(j-1)}_t \right|, \max_t \left| T^{(j)}_t - T^{(j-1)}_t \right| \right\} < \epsilon
\]

(3.54)

for some suitable \( \epsilon \). Given that our data will typically be observational and have some experimental (or recorded) precision of \( \pm \delta \), iterating to convergence level \( \epsilon = \delta/10 \) is a reasonable rule-of-thumb.

Implementing this iterative process for our sample from Eqn. 3.35, we obtain residuals which are reasonably white. For the general case, the residuals can have arbitrarily complicated structure, although by assumption they will be of the background process, which is presumably continuous. Figure 3.4 shows the time- and frequency-domain plots for the same realization exhibited in the previous plots.

### 3.2.6 Best Least-Squares Prediction (Interpolation)

As detailed in Sections 3.1 through 3.1.5, once we have obtained an approximately white series, we apply the best mean-square error linear interpolator to the residual series. Specifically, apply Eqn. 3.18 to the residual data from the previously iterated \( M_t \) and \( T_t \) steps, with missing data carried through in each step. Each estimate provides values for missing data points, and these points are preserved, but not passed on to successive iterative steps. The output from this stage will be referred to below as \( W_t \).

### 3.2.7 Global M Step Iteration

There are three sub-steps to each \( M \) step, of which two (the \( M_t \) and \( T_t \) steps) are internally iterated to convergence. The outputs of these individual component steps are passed to the \( E \) step where they are combined (additively) to produce an estimate \( \hat{x}_t = M_t + T_t + W_t \). Each missing point \( x_t \) is then replaced by its estimate \( \hat{x}_t \), and the new overall series passed back into the \( M \) steps.
Figure 3.4: Time- and Frequency-domain representations for one realization of Eqn. 3.35. Estimation of $M_t$ and $T_t$ was iterated until the maximum difference per step was less than $1e-4$. 
$M_t$ and $T_t$ are estimated on the interpolated (from the previous-stage $M$ estimate) series wherever possible. The Wiener interpolation step uses the difference of the interpolated series and the new $M_t + T_t$ partial estimate to produce the ACVF and then computes the interpolation weights based only on the “real” non-missing values.

### 3.2.8 Global Algorithm

On the next page we present a more algorithmic representation of the steps that are taken in the global interpolation algorithm. This algorithm matches the code that is included in the tsinterp package for R, see Appendix B. In all simulation examples that follow, $\epsilon_{\text{fixed}} = 1 \times 10^{-3}$. This should not be set too small: most observational data is only recorded to 1 or 2 decimal accuracy, so requiring that our interpolator converge to anything more stringent than one additional decimal is excessive. It is also possible to estimate the quantization of data which would allow for automatically adapting this parameter to different data sets without user input.

### 3.3 Examples Using Simulation

In this section we demonstrate two simple examples each with two forms of randomly placed gaps. We must make a note about our methodology. When looking at interpolation points, we have categorized all of the “missing” points based on their distance from the nearest “good” point. This distance is referred to from here on as $g_d$ (gap distance). When measuring the errors for points with common $g_d$, we will refer to the distribution we expect those points to have been drawn from. To test the points we use the Anderson-Darling test [4] for goodness-of-fit, with $p$-value calculation based on [112] and code from the ADGofTest package [11]. The formula for the test statistic $A$ which assesses if sorted errors $y_1 < \ldots < y_n$ come from a distribution with CDF
Global Algorithm

Initialize Step

1. Linearly interpolate the gappy original data series, label this $Z_t^{(0)}$.

2. Compute a pilot estimate of the spectrum and invert to obtain the pilot ACVF.

3. Estimate $\hat{M}_t^{(0)}$ using $Z_t^{(0)}$ as input.

4. Estimate $\hat{T}_t^{(0)}$ using $Z_t^{(0)} - \hat{M}_t$ as input.
   - Fix the frequencies obtained in this zeroth step for the duration of the algorithm.
   - Iterate the $M_t$, $T_t$ sequence until convergence, obtaining final estimates of $\hat{M}_t^{(0)}$ and $\hat{T}_t^{(0)}$.

5. Estimate $\hat{W}_t^{(0)}$ using $Z_t^{(0)} - \hat{M}_t^{(0)} - \hat{T}_t^{(0)}$ as input.

6. Take $\hat{M}_t^{(0)}$, $\hat{T}_t^{(0)}$ and $\hat{W}_t^{(0)}$ and add them together to give $\hat{Z}_t^{(1)}$. (E step)

Convergence Step $k$

1. Estimate the spectrum (and corresponding ACVF) of $\hat{Z}_t^{(k)}$. (E step)

2. Estimate $\hat{M}_t^{(k)}$ using $\hat{Z}_t^{(k)}$ as input. (M step)

3. Estimate $\hat{T}_t^{(k)}$ using $\hat{Z}_t^{(k)} - \hat{M}_t^{(k)}$ as input.
   - Iterate the $M_t$, $T_t$ steps until convergence, obtaining final estimates of $\hat{M}_t^{(k)}$ and $\hat{T}_t^{(k)}$.

4. Estimate $\hat{W}_t^{(k)}$ using $\hat{Z}_t^{(k)} - \hat{M}_t^{(k)} - \hat{T}_t^{(k)}$ as input.

5. Take $\hat{M}_t^{(k)}$, $\hat{T}_t^{(k)}$ and $\hat{W}_t^{(k)}$ and add them together to give $\hat{Z}_t^{(k+1)}$. (E step)

6. Evaluate the maximum absolute difference $\delta$ between $\hat{Z}_t^{(k+1)}$ and $\hat{Z}_t^{(k)}$. If $\delta > \epsilon_{\text{fixed}}$ then iterate.
\[ F \]
\[ A^2 = -n - \left( \sum_{k=1}^{n} \frac{2k-1}{n} [\ln(F(y_k)) + \ln(1 - F(y_{n+1-k}))] \right) \]  
(3.55)

which can then be tested against the critical values of the \( F \) distribution. For our purposes \( F \) will be normal, e.g. \( \mathcal{N}(0, 1) \) with the variance depending on the particular example.

### 3.3.1 Extending Eqn. 3.35

Consider again Eqn. 3.35, as explored in the previous \( M_t \) and \( T_1 \) derivations. We consider five gap situations: 5%, 10%, 15%, 20% and 25% of the data missing. Each simulated series has length \( N = 1000 \), and we restrict the missing points to the center 800 points (80%) of the time span, and randomize the gaps for each realization. That is, for each realization up to three things change: (a) the normal random variable portion of \( x_t \); (b) the location of the missing points; and (c) the number of missing points. For each gap-percentage we simulate 100 realizations, and iterate the gapfilling routine of the previous section.

Distributional results for the errors of the interpolated points are displayed in Table 3.2; note that due to the IID \( \mathcal{N}(0, 1) \) nature of the background, we do not expect \( W_t \) to predict anything at all, and thus the errors for each missing point should be approximately \( \mathcal{N}(0, 1) \). We apply Eqn. (3.27) (the extension of Kolmogorov’s formula to interpolation) to obtain an estimate of the lag-1 interpolation variance, obtaining a mean estimate (across the 100 realizations) of \( \sigma^2 = 0.858 \). Testing the error categories in Table 3.2 against this distribution gives comically small \( p \)-values for the cases which have sufficiently large collections of points, i.e. it is not a good fit. Note that the Anderson-Darling test is very sensitive to departures from normality, so that may contribute somewhat to the results presented.

Figure 3.5 shows a histogram of all of the \( g_d = 1 \) interpolation points for the 15%
Table 3.2: Using the model of Eqn. 3.35 and the randomly scattered gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, standardized 4th moment, and Anderson-Darling test for goodness-of-fit to a $\mathcal{N}(0, \sigma^2)$ (with $\sigma^2 = 0.858$) of the collection of points that meet the distance criterion is evaluated and recorded. The AD test report is the $p$-value for the test statistic. If less than 10 points are available across the 100 realizations, the statistics are not computed.

<table>
<thead>
<tr>
<th>$g_d$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4980</td>
<td>9856</td>
<td>14447</td>
<td>18794</td>
<td>22489</td>
</tr>
<tr>
<td></td>
<td>0.0129</td>
<td>-0.0169</td>
<td>0.0051</td>
<td>0.0040</td>
<td>0.0026</td>
</tr>
<tr>
<td></td>
<td>1.0391</td>
<td>1.0196</td>
<td>1.0583</td>
<td>1.0629</td>
<td>1.0576</td>
</tr>
<tr>
<td></td>
<td>-0.0261</td>
<td>0.0212</td>
<td>-0.0224</td>
<td>0.0223</td>
<td>0.0174</td>
</tr>
<tr>
<td></td>
<td>2.9839</td>
<td>3.0148</td>
<td>3.0184</td>
<td>2.9974</td>
<td>3.0235</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>143</td>
<td>529</td>
<td>1132</td>
<td>2264</td>
</tr>
<tr>
<td></td>
<td>0.0084</td>
<td>0.0051</td>
<td>-0.0161</td>
<td>-0.0269</td>
<td>-0.0185</td>
</tr>
<tr>
<td></td>
<td>0.9458</td>
<td>0.9433</td>
<td>1.1599</td>
<td>1.0773</td>
<td>1.1055</td>
</tr>
<tr>
<td></td>
<td>0.4185</td>
<td>-0.0328</td>
<td>-0.0587</td>
<td>-0.0035</td>
<td>-0.0029</td>
</tr>
<tr>
<td></td>
<td>2.3904</td>
<td>2.5477</td>
<td>3.1127</td>
<td>2.9836</td>
<td>2.9666</td>
</tr>
<tr>
<td></td>
<td>0.8422</td>
<td>0.1418</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>69</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-0.0621</td>
<td>0.0124</td>
<td>0.0345</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>0.8974</td>
<td>1.0333</td>
<td>1.1136</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-0.3847</td>
<td>0.0640</td>
<td>0.1331</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>2.5046</td>
<td>3.3891</td>
<td>3.2161</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>0.9063</td>
<td>0.5841</td>
<td>0.0248</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.0723</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.6062</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.4924</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.4254</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.7593</td>
<td>-</td>
</tr>
<tr>
<td>$g_d$</td>
<td>5%</td>
<td>10%</td>
<td>15%</td>
<td>20%</td>
<td>25%</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>948</td>
<td>1780</td>
<td>2516</td>
<td>3180</td>
<td>3660</td>
</tr>
<tr>
<td></td>
<td>-0.0232</td>
<td>-0.0201</td>
<td>-0.0249</td>
<td>-0.0006</td>
<td>0.0096</td>
</tr>
<tr>
<td></td>
<td>1.0900</td>
<td>1.1073</td>
<td>1.1023</td>
<td>1.1378</td>
<td>1.0829</td>
</tr>
<tr>
<td></td>
<td>-0.0067</td>
<td>0.0294</td>
<td>0.0781</td>
<td>0.0688</td>
<td>-0.0918</td>
</tr>
<tr>
<td></td>
<td>3.0220</td>
<td>2.8767</td>
<td>3.0379</td>
<td>3.0445</td>
<td>2.9446</td>
</tr>
<tr>
<td></td>
<td>0.3957</td>
<td>0.0828</td>
<td>0.0147</td>
<td>0.0092</td>
<td>0.0243</td>
</tr>
<tr>
<td>2</td>
<td>948</td>
<td>1780</td>
<td>2516</td>
<td>3180</td>
<td>3660</td>
</tr>
<tr>
<td></td>
<td>0.0379</td>
<td>-0.0046</td>
<td>-0.0024</td>
<td>0.0144</td>
<td>-0.0216</td>
</tr>
<tr>
<td></td>
<td>1.0562</td>
<td>1.1195</td>
<td>1.1012</td>
<td>1.1158</td>
<td>1.0758</td>
</tr>
<tr>
<td></td>
<td>-0.1355</td>
<td>0.0346</td>
<td>-0.0266</td>
<td>0.0019</td>
<td>-0.0218</td>
</tr>
<tr>
<td></td>
<td>3.0388</td>
<td>3.2878</td>
<td>3.0235</td>
<td>2.9144</td>
<td>2.9240</td>
</tr>
<tr>
<td></td>
<td>0.0658</td>
<td>0.2248</td>
<td>0.1250</td>
<td>0.0127</td>
<td>0.0408</td>
</tr>
<tr>
<td>3</td>
<td>948</td>
<td>1780</td>
<td>2516</td>
<td>3180</td>
<td>3660</td>
</tr>
<tr>
<td></td>
<td>0.0006</td>
<td>-0.0181</td>
<td>0.0052</td>
<td>-0.0131</td>
<td>-0.0295</td>
</tr>
<tr>
<td></td>
<td>1.0503</td>
<td>1.1509</td>
<td>1.0731</td>
<td>1.0488</td>
<td>1.0673</td>
</tr>
<tr>
<td></td>
<td>0.0241</td>
<td>-0.0099</td>
<td>0.0186</td>
<td>-0.0757</td>
<td>0.0198</td>
</tr>
<tr>
<td></td>
<td>2.6626</td>
<td>3.0047</td>
<td>2.9409</td>
<td>3.0359</td>
<td>3.0198</td>
</tr>
<tr>
<td></td>
<td>0.3703</td>
<td>0.0153</td>
<td>0.2741</td>
<td>0.3554</td>
<td>0.0268</td>
</tr>
<tr>
<td>4</td>
<td>948</td>
<td>1780</td>
<td>2516</td>
<td>3180</td>
<td>3660</td>
</tr>
<tr>
<td></td>
<td>0.0386</td>
<td>0.0168</td>
<td>0.0012</td>
<td>-0.0307</td>
<td>-0.0192</td>
</tr>
<tr>
<td></td>
<td>1.0917</td>
<td>1.1007</td>
<td>1.0870</td>
<td>1.0931</td>
<td>1.1114</td>
</tr>
<tr>
<td></td>
<td>0.1089</td>
<td>0.0353</td>
<td>0.0077</td>
<td>0.0515</td>
<td>0.0227</td>
</tr>
<tr>
<td></td>
<td>3.2076</td>
<td>2.9276</td>
<td>3.0217</td>
<td>2.9307</td>
<td>2.9067</td>
</tr>
<tr>
<td></td>
<td>0.3143</td>
<td>0.1408</td>
<td>0.2319</td>
<td>0.0092</td>
<td>0.0058</td>
</tr>
<tr>
<td>5</td>
<td>948</td>
<td>1780</td>
<td>2516</td>
<td>3180</td>
<td>3660</td>
</tr>
<tr>
<td></td>
<td>-0.0108</td>
<td>-0.0238</td>
<td>0.0294</td>
<td>0.0113</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td>1.0205</td>
<td>1.0970</td>
<td>1.0810</td>
<td>1.0792</td>
<td>1.1363</td>
</tr>
<tr>
<td></td>
<td>-0.0195</td>
<td>0.0645</td>
<td>0.0056</td>
<td>-0.0041</td>
<td>0.0338</td>
</tr>
<tr>
<td></td>
<td>2.8250</td>
<td>3.0521</td>
<td>3.0114</td>
<td>2.9537</td>
<td>3.1169</td>
</tr>
<tr>
<td></td>
<td>0.7699</td>
<td>0.1232</td>
<td>0.0739</td>
<td>0.1274</td>
<td>0.0130</td>
</tr>
</tbody>
</table>

Table 3.3: Using the model of Eqn. 3.35 and the “block” gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, standardized 4th moment, and Anderson-Darling test for goodness-of-fit to a $\mathcal{N}(0, 1)$ of the collection of points that meet the distance criterion is evaluated and recorded. The AD test report is the $p$-value for the test. If less than 10 points are available across the 100 realizations, the statistics are not computed.
<table>
<thead>
<tr>
<th>(g_d)</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>48</td>
<td>192</td>
<td>381</td>
<td>639</td>
<td>966</td>
</tr>
<tr>
<td></td>
<td>-0.2581</td>
<td>0.0320</td>
<td>-0.0798</td>
<td>-0.0834</td>
<td>0.0052</td>
</tr>
<tr>
<td></td>
<td>0.8292</td>
<td>1.0373</td>
<td>1.2375</td>
<td>1.1276</td>
<td>1.1365</td>
</tr>
<tr>
<td></td>
<td>0.4249</td>
<td>0.0423</td>
<td>0.0967</td>
<td>-0.0373</td>
<td>0.0250</td>
</tr>
<tr>
<td></td>
<td>2.9222</td>
<td>2.8724</td>
<td>2.6374</td>
<td>3.0497</td>
<td>2.9341</td>
</tr>
<tr>
<td></td>
<td>0.0387</td>
<td>0.9267</td>
<td>0.0153</td>
<td>0.0403</td>
<td>0.2514</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>148</td>
<td>291</td>
<td>520</td>
<td>808</td>
</tr>
<tr>
<td></td>
<td>-0.0332</td>
<td>-0.0979</td>
<td>0.0162</td>
<td>-0.0445</td>
<td>-0.0040</td>
</tr>
<tr>
<td></td>
<td>1.5258</td>
<td>1.0036</td>
<td>1.1133</td>
<td>1.0776</td>
<td>1.0703</td>
</tr>
<tr>
<td></td>
<td>-0.0952</td>
<td>0.0877</td>
<td>0.3286</td>
<td>-0.0918</td>
<td>-0.0623</td>
</tr>
<tr>
<td></td>
<td>2.5884</td>
<td>3.3563</td>
<td>3.5790</td>
<td>2.8757</td>
<td>2.9832</td>
</tr>
<tr>
<td></td>
<td>0.4141</td>
<td>0.3466</td>
<td>0.8975</td>
<td>0.4339</td>
<td>0.7273</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>112</td>
<td>223</td>
<td>393</td>
<td>642</td>
</tr>
<tr>
<td></td>
<td>0.1959</td>
<td>0.1075</td>
<td>-0.0601</td>
<td>0.0386</td>
<td>0.0236</td>
</tr>
<tr>
<td></td>
<td>1.4106</td>
<td>0.9292</td>
<td>1.1899</td>
<td>1.0026</td>
<td>1.0852</td>
</tr>
<tr>
<td></td>
<td>-0.6422</td>
<td>-0.0182</td>
<td>-0.2548</td>
<td>-0.0222</td>
<td>-0.0416</td>
</tr>
<tr>
<td></td>
<td>3.3740</td>
<td>2.5490</td>
<td>3.4006</td>
<td>3.0566</td>
<td>3.0821</td>
</tr>
<tr>
<td></td>
<td>0.1315</td>
<td>0.2721</td>
<td>0.5404</td>
<td>0.6440</td>
<td>0.6694</td>
</tr>
<tr>
<td>9</td>
<td>26</td>
<td>75</td>
<td>139</td>
<td>259</td>
<td>454</td>
</tr>
<tr>
<td></td>
<td>-0.1943</td>
<td>0.0680</td>
<td>0.0832</td>
<td>-0.0861</td>
<td>0.0016</td>
</tr>
<tr>
<td></td>
<td>0.6536</td>
<td>1.0493</td>
<td>1.0523</td>
<td>0.9442</td>
<td>1.0976</td>
</tr>
<tr>
<td></td>
<td>-0.1977</td>
<td>0.3487</td>
<td>0.2763</td>
<td>-0.2458</td>
<td>-0.0620</td>
</tr>
<tr>
<td></td>
<td>3.1627</td>
<td>3.1393</td>
<td>3.5544</td>
<td>2.9378</td>
<td>3.1272</td>
</tr>
<tr>
<td></td>
<td>0.3499</td>
<td>0.8575</td>
<td>0.7270</td>
<td>0.3750</td>
<td>0.8752</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>35</td>
<td>88</td>
<td>133</td>
<td>294</td>
</tr>
<tr>
<td></td>
<td>0.0824</td>
<td>-0.1503</td>
<td>0.0472</td>
<td>-0.0254</td>
<td>0.0145</td>
</tr>
<tr>
<td></td>
<td>1.4257</td>
<td>1.2173</td>
<td>1.1914</td>
<td>1.0650</td>
<td>1.0663</td>
</tr>
<tr>
<td></td>
<td>-0.2467</td>
<td>-0.2183</td>
<td>0.1928</td>
<td>0.2427</td>
<td>0.0670</td>
</tr>
<tr>
<td></td>
<td>2.5486</td>
<td>2.0993</td>
<td>2.7955</td>
<td>3.3020</td>
<td>2.6446</td>
</tr>
<tr>
<td></td>
<td>0.4885</td>
<td>0.3729</td>
<td>0.7757</td>
<td>0.9307</td>
<td>0.6409</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>73</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.0811</td>
<td>-0.1207</td>
<td>0.0111</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1.8248</td>
<td>1.1105</td>
<td>1.1306</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.3447</td>
<td>-0.2305</td>
<td>0.2882</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>3.0536</td>
<td>2.6949</td>
<td>3.3769</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.2242</td>
<td>0.5187</td>
<td>0.9440</td>
</tr>
</tbody>
</table>

Table 3.4: Continuation of Table 3.3. Results for \(g_d \in \{12, \ldots, 16\}\) are left off this table for sake of space.
Figure 3.5: Histogram of errors in interpolations of $g_d = 1$ missing points for the 15% case of simulation 1. The red curve is a $\mathcal{N}(0, \sigma^2)$ density for $\sigma^2 = 0.858$ and the $\sigma$ axis corresponds to this distribution.

missing case. The red curve overdrawn is a $\mathcal{N}(0, \sigma^2)$ distribution, where $\sigma^2 = 0.858$ was found using Eqn. (3.31) and a multitaper estimate of the power spectrum. As can clearly be seen from both this figure and the first four moments in Table 3.2, the distribution of the errors is similar to this normal distribution. This is reassuring, as a white noise background should not be predictable, and the errors matching the distribution of this background indicate that the interpolator is performing at least as well as a (Wiener) prediction-only interpolator might using $W_t$ only.

\[^{2}\]That is, find $\sigma_I^2(1) = \left[2 \cdot \int_{0}^{1/2} S^{-1}(f) df \right]^{-1}$. 

52
In the first simulation, the gaps were randomly placed one point at a time throughout the [101, 900] domain. In the second simulation, we instead randomly place 10 point blocks throughout the same domain, restricting ourselves to cases where the gaps do not overlap by more than 2 points. This can provide slightly less missing points than the first simulation, but ensures that the points that are missing will be composed of longer gap lengths. The results for this simulation are exhibited in Tables 3.3 and 3.4, and although more gap distances are available, present the same story as the first simulation and Table 3.2. For these tables we have used $\sigma^2 = 1.0$ to match theory, with similar results to the previous run where we allowed $\sigma^2 = 0.858$. One thing of interest is that as $g_d$ increases (see especially Table 3.4) the variance of the errors stabilizes at $\sim 1.1$, slightly higher than the theoretical background white noise variance.

### 3.4 An ARMA(4,4) Example

Consider now an auto-regressive moving-average (ARMA) process with 4 lags for both the AR and MA components. We choose the coefficients used by Randolph et al. [150]:

$$X_t + 0.1X_{t-1} + 1.66X_{t-2} + 0.093X_{t-3} + 0.8649X_{t-4} = Z_t + 0.0266Z_{t-1} + 0.8175Z_{t-2} + 0.0595Z_{t-3} + 0.0764Z_{t-4}.$$  \hspace{1cm} (3.56)

Note here that, as far as the author is aware, there are no industry-standard examples that should be compared for interpolation. A review of the literature (given in the previous chapter) showed no consistent use of standard data sets for demonstration of efficacy or power for newly proposed interpolation methods.

The power spectrum of a single realization of the ARMA process chosen is shown in Figure 3.6 with the theoretical spectrum plotted over top in blue. As in the previous example, 100 realizations of both randomly scattered and 10-point “block”
scattered gaps were created and the series then interpolated. The results for the random scattering are shown in Table 3.5 and the “block” scattering in Tables 3.6 and 3.7.

The moments of this ARMA(4,4) process (taken as a non-serially correlated set of sample realizations) are approximately distributed as $\mathcal{N}(0, 4)^3$. In Table 3.5, the mean, skewness, and standardized 4th moment also reflect a normal distribution, but with variance that is shows more variation between cases than previously. Note that the one-point interpolation variance is $\sigma_f^2(1) = 0.473$. This is highly encouraging, as it says the interpolator is performing at least as well as the best least-squares linear predictor could be expected to. The higher variances for the $g_d = 3$ and $g_d = 4$ cases

---

3As can be derived using [117], the variance of these samples is 3.955.
are less inspiring, but the total number of points considered is quite small. Figure 3.7 shows an example from the 20% missing case for $g_d = 2$ which shows the underlying process distribution relative to the errors. The red curve is a reasonable fit to the histogram of the errors, although clearly with smaller variance than the data.

The message to take from Tables 3.6 and 3.7 is the slowly increasing variance of the errors as both $g_d$ and percentage of missing data increase. Linear prediction

![Figure 3.7: Histogram of errors in interpolations of $g_d = 2$ missing points for the 20% case of simulation 2 using randomly scattered gaps. The red curve is a $\mathcal{N}(0, \sigma^2)$ density with $\sigma^2 = 0.473$, computed again using Kolmogorov's formula. For comparison with the underlying ARMA(4,4) process, taken as random samples from a normal distribution, the blue curve shows a $\mathcal{N}(0, 4)$. The $\sigma$ axis corresponds to the former, Kolmogorov-derived, distribution.](image-url)
Table 3.5: Using the ARMA(4,4) model of Eqn. 3.56 and the randomly scattered gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, standardized 4th moment, and Anderson-Darling test for goodness-of-fit to a $\mathcal{N}(0,\sigma^2)$ of the collection of points that meet the distance criterion is evaluated and recorded for $\sigma^2 = 0.473$. The AD test report is the $p$-value for the test. If less than 10 points are available across the 100 realizations, the statistics are not computed.

<table>
<thead>
<tr>
<th>gd</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4982</td>
<td>9862</td>
<td>14472</td>
<td>18804</td>
<td>22549</td>
</tr>
<tr>
<td></td>
<td>0.0049</td>
<td>-0.0045</td>
<td>0.0042</td>
<td>-0.0070</td>
<td>-0.0102</td>
</tr>
<tr>
<td></td>
<td>0.6421</td>
<td>0.6722</td>
<td>0.7536</td>
<td>0.8333</td>
<td>0.9087</td>
</tr>
<tr>
<td></td>
<td>-0.0676</td>
<td>0.0037</td>
<td>0.0006</td>
<td>-0.0426</td>
<td>0.0084</td>
</tr>
<tr>
<td></td>
<td>3.0673</td>
<td>3.1210</td>
<td>3.2534</td>
<td>3.1797</td>
<td>3.2638</td>
</tr>
<tr>
<td></td>
<td>1.2e-7</td>
<td>6.1e-8</td>
<td>4.1e-8</td>
<td>3.2e-7</td>
<td>2.7e-2</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>137</td>
<td>511</td>
<td>1131</td>
<td>2234</td>
</tr>
<tr>
<td></td>
<td>0.0492</td>
<td>0.0889</td>
<td>0.0194</td>
<td>0.0133</td>
<td>-0.0379</td>
</tr>
<tr>
<td></td>
<td>0.3600</td>
<td>0.9448</td>
<td>0.7443</td>
<td>0.7616</td>
<td>0.8715</td>
</tr>
<tr>
<td></td>
<td>0.2190</td>
<td>0.1376</td>
<td>-0.0603</td>
<td>0.0085</td>
<td>-0.0298</td>
</tr>
<tr>
<td></td>
<td>2.5191</td>
<td>3.6328</td>
<td>2.6797</td>
<td>2.9703</td>
<td>3.4447</td>
</tr>
<tr>
<td></td>
<td>5.4e-1</td>
<td>4.4e-6</td>
<td>1.2e-6</td>
<td>5.3e-7</td>
<td>2.7e-7</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>17</td>
<td>60</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>0.0401</td>
<td>-0.1768</td>
<td>0.0494</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>1.1658</td>
<td>1.8953</td>
<td>1.5897</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-0.0228</td>
<td>-0.0267</td>
<td>0.1507</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>2.2964</td>
<td>2.6963</td>
<td>2.9671</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>3.2e-4</td>
<td>1.0e-5</td>
<td>3.2e-6</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.2030</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.3889</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.3013</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.9418</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.6e-5</td>
</tr>
<tr>
<td>( g_d )</td>
<td>5%</td>
<td>10%</td>
<td>15%</td>
<td>20%</td>
<td>25%</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>942</td>
<td>1774</td>
<td>2490</td>
<td>3162</td>
<td>3702</td>
</tr>
<tr>
<td></td>
<td>-0.0250</td>
<td>-0.0028</td>
<td>-0.0042</td>
<td>-0.0010</td>
<td>0.0065</td>
</tr>
<tr>
<td></td>
<td>0.9749</td>
<td>1.0106</td>
<td>1.0531</td>
<td>1.0470</td>
<td>1.0716</td>
</tr>
<tr>
<td></td>
<td>-0.0138</td>
<td>-0.0781</td>
<td>-0.0033</td>
<td>-0.0033</td>
<td>0.0029</td>
</tr>
<tr>
<td></td>
<td>3.1526</td>
<td>2.9428</td>
<td>3.1724</td>
<td>3.0354</td>
<td>3.1078</td>
</tr>
<tr>
<td></td>
<td>6.4e-7</td>
<td>3.4e-7</td>
<td>2.4e-7</td>
<td>1.9e-7</td>
<td>1.6e-7</td>
</tr>
<tr>
<td>2</td>
<td>942</td>
<td>1774</td>
<td>2490</td>
<td>3162</td>
<td>3702</td>
</tr>
<tr>
<td></td>
<td>0.0255</td>
<td>0.0209</td>
<td>-0.0120</td>
<td>-0.0153</td>
<td>0.0113</td>
</tr>
<tr>
<td></td>
<td>0.9460</td>
<td>0.9477</td>
<td>1.0411</td>
<td>1.0321</td>
<td>1.0161</td>
</tr>
<tr>
<td></td>
<td>0.0066</td>
<td>-0.0112</td>
<td>0.0127</td>
<td>-0.0814</td>
<td>-0.0177</td>
</tr>
<tr>
<td></td>
<td>2.8541</td>
<td>2.9510</td>
<td>3.1067</td>
<td>2.9290</td>
<td>2.9803</td>
</tr>
<tr>
<td></td>
<td>6.4e-7</td>
<td>3.4e-7</td>
<td>2.4e-7</td>
<td>1.9e-7</td>
<td>1.6e-7</td>
</tr>
<tr>
<td>3</td>
<td>942</td>
<td>1774</td>
<td>2490</td>
<td>3162</td>
<td>3702</td>
</tr>
<tr>
<td></td>
<td>-0.0154</td>
<td>0.0007</td>
<td>-0.0043</td>
<td>0.0077</td>
<td>-0.0096</td>
</tr>
<tr>
<td></td>
<td>1.3846</td>
<td>1.4987</td>
<td>1.5503</td>
<td>1.6200</td>
<td>1.6013</td>
</tr>
<tr>
<td></td>
<td>0.0133</td>
<td>0.0264</td>
<td>-0.0160</td>
<td>-0.0175</td>
<td>-0.0520</td>
</tr>
<tr>
<td></td>
<td>3.1340</td>
<td>3.1389</td>
<td>3.1180</td>
<td>2.9574</td>
<td>3.0395</td>
</tr>
<tr>
<td></td>
<td>6.4e-7</td>
<td>3.4e-7</td>
<td>2.4e-7</td>
<td>1.9e-7</td>
<td>1.6e-7</td>
</tr>
<tr>
<td>4</td>
<td>942</td>
<td>1774</td>
<td>2490</td>
<td>3162</td>
<td>3702</td>
</tr>
<tr>
<td></td>
<td>-0.0711</td>
<td>-0.0448</td>
<td>-0.0260</td>
<td>-0.0137</td>
<td>-0.0032</td>
</tr>
<tr>
<td></td>
<td>1.5353</td>
<td>1.3960</td>
<td>1.5666</td>
<td>1.5831</td>
<td>1.5634</td>
</tr>
<tr>
<td></td>
<td>-0.0702</td>
<td>-0.0795</td>
<td>-0.0908</td>
<td>-0.0550</td>
<td>-0.0269</td>
</tr>
<tr>
<td></td>
<td>2.8997</td>
<td>3.4014</td>
<td>2.8780</td>
<td>3.0247</td>
<td>2.8985</td>
</tr>
<tr>
<td></td>
<td>6.4e-7</td>
<td>3.4e-7</td>
<td>2.4e-7</td>
<td>1.9e-7</td>
<td>1.6e-7</td>
</tr>
<tr>
<td>5</td>
<td>942</td>
<td>1774</td>
<td>2490</td>
<td>3162</td>
<td>3702</td>
</tr>
<tr>
<td></td>
<td>0.0558</td>
<td>0.0208</td>
<td>0.0015</td>
<td>0.0039</td>
<td>0.0048</td>
</tr>
<tr>
<td></td>
<td>1.5149</td>
<td>1.5872</td>
<td>1.6998</td>
<td>1.7869</td>
<td>1.7546</td>
</tr>
<tr>
<td></td>
<td>-0.1354</td>
<td>-0.0844</td>
<td>0.0697</td>
<td>0.0247</td>
<td>0.0574</td>
</tr>
<tr>
<td></td>
<td>2.8127</td>
<td>3.2390</td>
<td>2.9026</td>
<td>2.8923</td>
<td>2.9625</td>
</tr>
<tr>
<td></td>
<td>6.4e-7</td>
<td>3.4e-7</td>
<td>2.4e-7</td>
<td>1.9e-7</td>
<td>1.6e-7</td>
</tr>
</tbody>
</table>

Table 3.6: Using the ARMA\((4,4)\) model of Eqn. 3.56 and the “block” gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, standardized 4th moment, and Anderson-Darling test for goodness-of-fit to a \( \mathcal{N}(0, \sigma^2) \) of the collection of points that meet the distance criterion is evaluated and recorded, for \( \sigma^2 = 0.473 \) the asymptotic one-point interpolation error variance. The AD test report is the \( p \)-value for the test. If less than 10 points are available across the 100 realizations, the statistics are not computed.
<table>
<thead>
<tr>
<th>$gd$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>57</td>
<td>191</td>
<td>407</td>
<td>642</td>
<td>923</td>
</tr>
<tr>
<td></td>
<td>0.1734</td>
<td>0.0254</td>
<td>0.0288</td>
<td>0.0258</td>
<td>-0.0508</td>
</tr>
<tr>
<td></td>
<td>2.1570</td>
<td>2.0236</td>
<td>1.7558</td>
<td>1.7114</td>
<td>1.9434</td>
</tr>
<tr>
<td></td>
<td>0.0199</td>
<td>0.2410</td>
<td>-0.1071</td>
<td>-0.1801</td>
<td>0.0375</td>
</tr>
<tr>
<td></td>
<td>2.627</td>
<td>3.4585</td>
<td>3.4258</td>
<td>3.6518</td>
<td>3.1672</td>
</tr>
<tr>
<td></td>
<td>1.1e-5</td>
<td>3.1e-6</td>
<td>1.5e-6</td>
<td>9.3e-7</td>
<td>6.5e-7</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>156</td>
<td>337</td>
<td>524</td>
<td>766</td>
</tr>
<tr>
<td></td>
<td>0.3555</td>
<td>-0.0193</td>
<td>0.0210</td>
<td>0.0258</td>
<td>0.0559</td>
</tr>
<tr>
<td></td>
<td>1.5850</td>
<td>1.7191</td>
<td>2.1183</td>
<td>1.9356</td>
<td>1.9673</td>
</tr>
<tr>
<td></td>
<td>-0.4897</td>
<td>0.0472</td>
<td>-0.0686</td>
<td>-0.1490</td>
<td>-0.1877</td>
</tr>
<tr>
<td></td>
<td>2.5360</td>
<td>2.9136</td>
<td>3.0273</td>
<td>2.8858</td>
<td>2.8449</td>
</tr>
<tr>
<td></td>
<td>1.3e-5</td>
<td>3.8e-6</td>
<td>1.8e-6</td>
<td>1.1e-6</td>
<td>7.8e-7</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>126</td>
<td>274</td>
<td>403</td>
<td>595</td>
</tr>
<tr>
<td></td>
<td>-0.2938</td>
<td>0.1739</td>
<td>0.0123</td>
<td>-0.0369</td>
<td>0.0405</td>
</tr>
<tr>
<td></td>
<td>1.9606</td>
<td>1.7428</td>
<td>1.8794</td>
<td>1.7793</td>
<td>1.9417</td>
</tr>
<tr>
<td></td>
<td>0.0530</td>
<td>-0.1857</td>
<td>0.1282</td>
<td>0.1834</td>
<td>-0.0687</td>
</tr>
<tr>
<td></td>
<td>2.5592</td>
<td>2.6968</td>
<td>2.7294</td>
<td>2.8847</td>
<td>2.7910</td>
</tr>
<tr>
<td></td>
<td>1.7e-5</td>
<td>4.8e-6</td>
<td>2.2e-6</td>
<td>1.5e-6</td>
<td>1.0e-6</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>86</td>
<td>220</td>
<td>279</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>-0.3952</td>
<td>-0.1030</td>
<td>-0.0730</td>
<td>-0.0319</td>
<td>-0.0720</td>
</tr>
<tr>
<td></td>
<td>2.7807</td>
<td>1.5222</td>
<td>2.1840</td>
<td>1.9454</td>
<td>2.1054</td>
</tr>
<tr>
<td></td>
<td>-0.2109</td>
<td>0.0627</td>
<td>-0.0232</td>
<td>0.2082</td>
<td>-0.0627</td>
</tr>
<tr>
<td></td>
<td>1.8551</td>
<td>3.1964</td>
<td>2.9278</td>
<td>3.4441</td>
<td>2.9903</td>
</tr>
<tr>
<td></td>
<td>2.6e-5</td>
<td>7.0e-6</td>
<td>2.7e-6</td>
<td>2.2e-6</td>
<td>1.4e-6</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>43</td>
<td>112</td>
<td>155</td>
<td>278</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.1584</td>
<td>-0.0022</td>
<td>-0.0486</td>
<td>0.0137</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>1.9387</td>
<td>2.0939</td>
<td>2.1977</td>
<td>2.1836</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.0618</td>
<td>-0.0098</td>
<td>-0.3326</td>
<td>0.0596</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>1.4e-5</td>
<td>5.4e-6</td>
<td>3.9e-6</td>
<td>2.2e-6</td>
</tr>
</tbody>
</table>

Table 3.7: Continuation of Table 3.6.
high of $\sim 3.1$ for the longest $g_d$ and most number of missing points (bottom right of Table 3.7). Finally, Figure 3.8 shows a case (20% missing points, $g_d = 5$) where the variance of the errors exceeds the Kolmogorov’s formula, but remains smaller than the underlying process variance. Note that Kolmogorov’s formula is only for a single lag, so exceeding it in this case is not surprising.
3.4.1 Conclusions from the Simulations

The results of the simulations were encouraging. For the first two cases tested using the hybrid time series (continuous background and embedded line components) the variance of the errors was comparable to both that expected by the Kolmogorov one-point interpolation error variance and also the variance of the underlying noise process. In the case of the ARMA(4,4) process, the variance of the errors exceeded the Kolmogorov one-point interpolation error variance in all cases. This may be partially due to the application of the global algorithm to a series which contained no mean trend and no periodic components. Examination of the residuals from the estimation procedure showed that the algorithm detected no spurious periodic components, but that a non-zero mean trend was removed. This may have been sufficient to bias the estimation of the underlying ACVF which in turn would have produced biased prediction coefficients. More work can certainly be done to improve the performance of this algorithm for this type of test case, although this is not our primary focus.

3.5 Observational Example: Noon Solar Flux

In this section we consider two examples of daily noon 10.7cm solar flux. This series was chosen for its possible usefulness in our analysis of Chapter 6. The daily noon solar flux is an extremely valuable scientific data series. It can be used as a proxy for a number of other solar measurements, see [191], and comprises the longest running instrumental solar data set in existence\textsuperscript{4}. There are two primary observational sets of solar flux data: the Dominion Radio Astronomical Observatory (DRAO) series, in near-continuous observation since 1949, and the Radio Solar Telescope Network.

\textsuperscript{4}A longer-running solar series is the sunspot series. This series runs for hundreds of years, but is quite coarse and based almost entirely on human observers rather than instruments.
(RSTN) operated by the United States Air Force in five separate locations\textsuperscript{5} around the globe semi-continuously since 1966.

![Graph showing daily noon solar flux](image)

Figure 3.9: Daily noon solar flux (in dark grey) as observed by the RSTN site at Sagamore Hill, Mass., USA. 1,316 values of this series are missing. The black curve is a smooth of the daily flux to help show the cyclostationary variance pattern.

The data used in this analysis were obtained from two sources. The Penticton data were obtained from [121] and modified slightly in consultation with Dr. Ken Tapping of the National Research Council of Canada, permanent staff at DRAO. The Sagamore Hill Observatory data was obtained from the NGDC [124] in consultation with Capt. Annette Parsons, USAF and Rob Steenburgh of NOAA. Both data sets are available in the form used in this thesis upon request to the author.

\textsuperscript{5}Learmouth, WA, Australia; Sagamore Hill, Massachusetts, USA; Holloman AFB, New Mexico, USA; Palehua, Hawai‘i, USA; and San Vito Observatory, San Vito dei Normanni, Italy.
The Penticton data is remarkably complete, spanning a period from 1949, when the project was initiated, through to present day. The Sagamore Hill data is much less complete, with a number of extensive gaps due to instrument failure and repair. Both instruments observe the noon solar flux at their respective locations, and both report their results in standardized URSI-D solar flux units (SFU). For a condensed history of solar flux observations, see [37].

![Example of daily noon solar flux as observed by the RSTN site at Sagamore Hill, Mass., USA. Light grey curves are the values provided by the RSTN, and the black curves are the interpolated values. The rug plot at the bottom provides a references for which values were missing.](image)

The Penticton series will be used more thoroughly in the next chapter. Here we will focus on the Sagamore Hill series. The series used runs from May 1, 1966 to Jan. 31, 2008, a total of 15,251 daily observations. Of these, 1,316 are missing (∼ 8.6%). Figure 3.9 shows the gappy time series. It is clear from this figure that this series
is not stationary – not only are there strong periodic structures, but the variance of the series is time-dependent, being significantly larger at the peaks (solar maximum) than the levels at the troughs (solar minimum). For this reason, when interpolating this series we restrict attention to a rough span of 1000 days at a time (corresponding to \( \approx 3 \) years which is slightly more than half the estimate of the half-solar period), which helps keep the “local” variance approximately constant. As a demonstration, consider Figure 3.11. There may be a more optimal choice given the cyclostationarity inherent in this series, but explorations yielded no insight.

Further examine Figure 3.10 which shows the interpolation for the \( \sim 1980–1982 \) span in which there were severe instrument failures at Sagamore Hill and the data
accordingly spotty. In a highly non-scientific back-of-the-envelope sense, the interpolator did a reasonable job: a) there are no strange excursions from the time series that do not seem to fit; b) periodic (structural) patterns through the series appear to be left intact by the interpolator; and c) there is some evidence that structure is present in the interpolated values that would not be selectable using naive techniques. Unfortunately, there is very little we can do to evaluate this interpolation, as we do not know the original series.

To work around the problem of not knowing the underlying missing values, we take the contiguous Penticton series and create gaps in the exact pattern forced upon us by the Sagamore Hill data set. We then interpolate the newly gappy Penticton flux series in exactly the same fashion as before. The resultant fills are compared to the original observational data and the errors exhibited in Table 3.8. This table shows what seem at first glance to be extremely high error variances, however applying Kolmogorov’s formula gives $\sigma_1^2 = 1.905 \times 10^{11}$, and this is excessively higher than the actual variance observed for the synthetic gaps. Applying the Anderson-Darling test against these errors gives the result that none of them match the distribution – hardly surprising given the wild disparity in variances.

### 3.6 Summary and Conclusions

To summarize, we wish to make one salient fact clear to the reader: interpolation is an art, not a science. Every time series is different, and it appears highly improbable, if not impossible, to develop a universal algorithm that will interpolate a series with missing values regardless of context. Take as an example the case of the Penticton/Sagamore Hill series that were demonstrated in the previous section. Very little was done to leverage the prior knowledge we have about the two series. Daily solar measurements are known to have extremely strong periodic structure on solar mode...
<table>
<thead>
<tr>
<th>( g_d )</th>
<th>Moments</th>
<th>( g_d )</th>
<th>Moments</th>
<th>( g_d )</th>
<th>Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>787</td>
<td>6</td>
<td>31</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>0.7645</td>
<td>-3.7133</td>
<td>-3.2675</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47.9109</td>
<td>659.4464</td>
<td>536.7974</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1425</td>
<td>0.8051</td>
<td>-0.5998</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19.0421</td>
<td>3.7870</td>
<td>2.9456</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.6e-07</td>
<td>1.9e-05</td>
<td>3.9e-04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>172</td>
<td>7</td>
<td>24</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>0.4515</td>
<td>-3.0539</td>
<td>-7.2318</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>158.7033</td>
<td>823.1099</td>
<td>344.9243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1106</td>
<td>0.6968</td>
<td>-0.3250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.5283</td>
<td>3.7997</td>
<td>2.9200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5e-06</td>
<td>4.8e-05</td>
<td>3.9e-04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>74</td>
<td>8</td>
<td>22</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>-0.7556</td>
<td>-4.0089</td>
<td>-4.8010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>235.0826</td>
<td>978.9214</td>
<td>389.5103</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8513</td>
<td>0.6612</td>
<td>0.2442</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.9738</td>
<td>3.5874</td>
<td>3.3300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.1e-06</td>
<td>9.1e-05</td>
<td>5.8e-04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>9</td>
<td>21</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>412.5614</td>
<td>833.5324</td>
<td>297.0196</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9360</td>
<td>0.0997</td>
<td>-0.4593</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3236</td>
<td>2.3373</td>
<td>3.2847</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4e-05</td>
<td>1.3e-04</td>
<td>2.9e-03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>10</td>
<td>18</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>-3.5124</td>
<td>-1.0076</td>
<td>-9.6786</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>575.7352</td>
<td>637.1321</td>
<td>265.0385</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1934</td>
<td>-0.1590</td>
<td>-0.3442</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.9062</td>
<td>2.9689</td>
<td>2.6430</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.7e-05</td>
<td>3.9e-04</td>
<td>6.9e-03</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.8: Total number of gap points corresponding to each \( g_d \) and the first four central moments of errors in interpolation for gaps artificially placed in the Penticton daily noon solar flux series. The final entry in each category is the Anderson-Darling test against a normal distribution with variance computed by Kolmogorov’s formula.
frequencies such as 11.1 years. It is possible to adaptively work with such time series and integrate this prior knowledge in an *ad hoc* way so as to improve significantly upon the results demonstrated in this chapter. Furthermore, as we will see in the next chapter, having two observational series such as these which are independent observations of a common data stream can allow us to use the common structure and noise that is due to the point source, and eliminate some of the bias due to instrument-specific noise.
Chapter 4

Bivariate Interpolation of Time Series

Often with observational series we have access to secondary observations from separate instruments, possibly located at different geographic locations. If the observations are of the same point source (e.g. the Sun) then regardless of differences in instrumentation and local background noise, the fundamental process underlying the observations should be shared between the sites. As a first approximation, consider observations $Z_k(t), k = 1, \cdots, p$ for $p \geq 2$ with local zero-mean additive noise processes $\xi_k(t)$. Assume an underlying hybrid (i.e. not stationary or non-deterministic due to line components and time-varying mean) process for site $k$ with instrumental and site-specific effects $\xi_k(t)$. To keep things simple we assume that the processes are sampled at the same rate, at the same time, with independent instrument/local site effects noise processes. There are myriad other complications that could be considered such as: aliases (sample at the same rate but at different times), changes in calibration and drift for the instruments, etc., but these are beyond the scope of this thesis.

Now, split the underlying hybrid process into three parts in the style of Chapter 3,
so that the $k$th series is

$$Z_k(t) = M_k(t) + T_k(t) + W_k(t) + \xi_k(t).$$  \hspace{1cm} (4.1)$$

Then, lump together the trend and periodic terms $M_k(t)$ and $T_k(t)$ into some “mean” term $N_k(t)$ and the noise terms $W_k(t)$ and $\xi_k(t)$ into a “random” term $X_k(t)$ so that

$$Z_k(t) = N_k(t) + X_k(t).$$  \hspace{1cm} (4.2)$$

We then have, explicitly,

$$Z_1(t) = N_1(t) + X_2(t)$$

$$Z_2(t) = N_2(t) + X_1(t)$$

$$\ldots$$

$$Z_p(t) = N_p(p) + X_p(t)$$

Clearly we have independence of $N_k(t)$ from $\xi_k(t + \tau)$ for all $i$ and $\tau$ as the first is deterministic and the second is composed of two (assumed) zero-mean random processes. In similar fashion to the derivation given in Chapter 3, begin by assuming that $X_m(t_0)$ is missing, for $m \in \{1, 2, \ldots, p\}$. Then, using the the Hilbert space formed as the span of all of the $X_k(t)$'s (an obvious extension of the one detailed in the univariate case), write

$$\hat{X}_m(t_0) = \sum_{i=1}^{p} \sum_{q=1}^{n_i} \beta_{i,q} X_i(t_0 + \tau_i(q))$$  \hspace{1cm} (4.4)$$

so that the linear prediction of the missing point is a linear combination of all other points from all $p$ series. The $\tau_i(q)$ is a “tag” array that gives the relative positions of the $X_k(t)$ terms relative to $t_0$.

Write the mean-square error:

$$e = \mathbb{E} \left\{ \left| \hat{X}_m(t_0) - X_m(t_0) \right|^2 \right\} = \mathbb{E} \left\{ \left( \hat{X}_m(t_0) - X_m(t_0) \right) \left( \hat{X}_m^*(t_0) - X_m^*(t_0) \right) \right\}$$  \hspace{1cm} (4.5)$$

where superscript $*$ refers to complex conjugation, and then differentiate it with
respect to the \( \beta_{i,q}^* \) expansion coefficients from Eqn. 4.4:

\[
\frac{\partial e}{\partial \beta_{j,r}^*} = 0 \text{ which implies } \frac{\partial e}{\partial \beta_{j,r}^*} = 0 \Rightarrow E \left\{ X_j^*(t_0 + \tau_j(r))X_i(t_0 + \tau_i(q)) \right\} = \sum_{i=1}^{p} \sum_{q=1}^{n_i} \beta_{iq} E \left\{ X_j^*(t_0 + \tau_j(r))X_i(t_0 + \tau_i(q)) \right\}.
\]

Further write the \( \tau \)-lagged cross-covariance between two series \( Z_t^{(i)} \) and \( Z_t^{(j)} \) as

\[
R_{ij}(\tau) = E \left\{ X_i(t + \tau)X_j^*(\tau) \right\}, \quad i, j = 1, 2, \ldots, p,
\]

and then evaluate Eqn. 4.6, conjugating both sides:

\[
R_{jm}(-\tau_j(r)) = \sum_{i=1}^{p} \sum_{q=1}^{n_i} \beta_{iq} R_{ji}(\tau_j(r) - \tau_i(q)).
\]

This then provides the fundamental (algebraic) equation which will be setup in matrix form similar to the implementation of Chapter 3. Now, consider the derivation for the univariate \( W_t \) interpolator, Equations 3.15, 3.17 and 3.18. There we showed how the linear prediction problem could be summarized as a linear equation in selected lags of the autocovariance function. As we will be using two series, we will be considering each series’ autocovariance function (ACVF) as well as the cross-covariance function (CCVF) between the two. As before, we will estimate the process ACVFs using the multitaper two-step estimation technique, first finding a pilot estimate of the spectrum and then inverting it to obtain the ACVF or CCVF. The multitaper cross-spectrum is used for the same reason as the univariate multitaper spectrum estimate\(^1\).

Note that the cross-spectrum \( \hat{S}_{12} \) can be found in a similar way by combining the eigencoefficients \( y_k(f) \) as

\[
\hat{S}_{12} = \frac{1}{K} \sum_{k=0}^{K-1} y_k^{(1)}(f)y_k^{(2)*}(f)
\]

where \( \left\{ y_k^{(1)} \right\}_{k=0}^{K-1} \) and \( \left\{ y_k^{(2)} \right\}_{k=0}^{K-1} \) are the eigencoefficients for the multitaper spectrum estimate of series 1 and series 2, respectively. This gives the complex spectrum

\(^1\)As mentioned by [228, pg. 322], while the cross-periodogram may be asymptotically unbiased, the estimate is inconsistent and behaves erratically.
\( \hat{S}_{12}(f) \) which can further be Fourier transformed to obtain the real cross-covariance function \( R_{12} \). Note that unlike the univariate case, the cross-covariance function is not necessarily symmetric with respect to variable order – i.e. \( R_{12}(\tau) = R_{21}^*(-\tau) \) [228, pg. 311]. In addition, the cross-correlation functions need not be even in the case of real processes.

Consider the bivariate case, \( m = 2 \), and assume we have estimates of \( R_{11}, R_{12}, R_{21}, \) and \( R_{22} \), the auto- and cross-covariances between the two series. As we are interested in interpolation, determine the number of non-missing points for each series, say \( n_1 \) and \( n_2 \) respectively. We then form a \((n_1 + n_2) \times (n_1 \times n_2)\) matrix \( A \) block-wise:

\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix},
\]  
(4.10)

where each \( A_{ij} \) represents the corresponding lagged auto- or cross-covariance elements from \( R_{11}, \) etc.²

The mean-square error can also be determined:

\[
e = \mathbb{E}\left\{ \left| \hat{X}_m(t_0) - X_m(t_0) \right|^2 \right\}
= \mathbb{E}\left\{ \left( X_m(t_0) - \hat{X}_m(t_0) \right) \left( X_m^*(t_0) - \hat{X}_m^*(t_0) \right) \right\}
\]  
(4.11)

(and recalling that \( X_m(t_0) - \hat{X}_m(t_0) \perp \hat{X}_t(t_0) \))

\[
= \mathbb{E}\left\{ \left( X_m^*(t_0) - \hat{X}_m^*(t_0) \right) X_m(t_0) \right\}
= \mathbb{E}\left\{ X_m^2(t_0) \right\} - \mathbb{E}\left\{ \hat{X}_m^*(t_0)X_m(t_0) \right\}
= \sigma_m^2 - \sum_{i=1}^{p} \sum_{q=1}^{n_i} \beta_{iq}^* \mathbb{E}\left\{ X_m(t_0)X_i^*(t_0 + \tau_i(q)) \right\}
= \sigma_m^2 - \sum_{i=1}^{p} \sum_{q=1}^{n_i} \beta_{iq}^* R_{mi}(-\tau_i(q)).
\]

We do not apply this relation in the following examples, but it can be extended (as we have estimates of \( \sigma_m^2 \) and \( R_{mi}(\tau) \)) to compare the performance of the overall

²The implementation of this is available in the \texttt{tsinterp} package, see Appendix B.
global interpolation algorithm with the interior $W_t$ bivariate best least-squares linear predictor/interpolator. This is a topic for further work.

4.1 Bivariate Interpolation Algorithm

<table>
<thead>
<tr>
<th>Algorithm: Series $i$ ($i \in {1, 2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialize Step</td>
</tr>
<tr>
<td>1. Linearly interpolate the gappy original data series, label this $Z_{i,t}^{(0)}$.</td>
</tr>
<tr>
<td>2. Compute a pilot estimate of the spectrum, $\hat{S}<em>{ii}^{(0)}(f)$ and invert to obtain the pilot ACVF $R</em>{ii}^{(0)}$.</td>
</tr>
<tr>
<td>3. Estimate $\hat{M}<em>{i,t}^{(0)}$ using $Z</em>{i,t}^{(0)}$ as input.</td>
</tr>
<tr>
<td>4. Estimate $\hat{T}<em>{i,t}^{(0)}$ using $Z</em>{i,t}^{(0)} - \hat{M}_{i,t}^{(0)}$ as input.</td>
</tr>
<tr>
<td>• Fix the frequencies obtained in this zeroth step for series 1 for the duration of the algorithm.</td>
</tr>
<tr>
<td>• Iterate the $M_{i,t}$, $T_{i,t}$ sequence until convergence, obtaining final estimates of $\hat{M}<em>{i,t}^{(0)}$ and $\hat{T}</em>{i,t}^{(0)}$.</td>
</tr>
<tr>
<td>5. Estimate $W_{i,t}^{(0)}$ using the bivariate estimator below using initial $\hat{S}_{ii}^{(0)}(f)$ eigencoefficients, $i = 1, 2$.</td>
</tr>
<tr>
<td>6. Take $\hat{M}<em>{i,t}^{(0)}$, $\hat{T}</em>{i,t}^{(0)}$ and $\hat{W}<em>{i,t}^{(0)}$ and add them together to give $\hat{Z}</em>{i,t}^{(1)}$.</td>
</tr>
</tbody>
</table>

The interpolation algorithm is essentially a bivariate version of the algorithm given in the previous chapter. Each series is independently interpolated to obtain spectrum estimates $\hat{S}_{11}$, $\hat{S}_{12}$, $\hat{S}_{21}$ and $\hat{S}_{22}$. These are inverted to obtain $\hat{R}_{11}$ and so on. The $M_t$ and $T_t$ references from this point on should be taken to refer to the same $M_t$ and $T_t$ estimators used in the last chapter, specifically Chapters 3.2.3 and 3.2.4.
Algorithm: Bivariate $W_{i,t}^{(j)}$ Estimator for Series $i \in \{1, 2\}$

$j$ Step, $j > 0$

1. Given eigencoefficients $y_{i,k}^{(j)}(f)$ for the two series, form the two cross-spectral estimates $\hat{S}_{12}^{(j)}(f)$ and $\hat{S}_{21}^{(j)}(f)$.

2. Invert the complex spectrum to form $\hat{R}_{12}^{(j)}(\tau)$ and $\hat{R}_{21}^{(j)}(\tau)$.

3. Create the $A$ block-matrix and corresponding $b$ block-vector for each missing point in each series and solve using $X_{i,t}$; this gives $\hat{W}_{i,t}^{(j)}$ for $i = 1, 2$.

Finally, iterate this process to convergence:

Algorithm: Series $i$ ($i \in \{1, 2\}$)

Convergence Step $j$

1. For $i = 1, 2$:
   - Compute an estimate of the spectrum, $\hat{S}_{ii}^{(j)}(f)$ using the $(j - 1)^{\text{st}}$ interpolation of $X_{i,t}$, i.e. $\hat{Z}_{it}^{(j-1)}$. Invert this to obtain the $j^{\text{th}}$ ACVF $R_{ii}^{(j)}$.
   - Estimate $\hat{M}_{i,t}^{(j)}$ using $Z_{i,t}^{(j-1)}$ as input.
   - Estimate $\hat{T}_{i,t}^{(j)}$ using $Z_{i,t}^{(j-1)} - \hat{M}_{i,t}^{(j-1)}$ as input.

2. Compute an estimate of one of the cross-spectra, $\hat{S}_{12}^{(j)}(f)$ and $\hat{S}_{21}^{(j)}(f)$. Invert this to obtain the $j^{\text{th}}$ CCVFs $R_{12}^{(j)}(\tau)$ and $R_{21}^{(j)}(\tau)$, $R_{21}(\tau) = R_{12}^{*}(-\tau)$.

3. Estimate $W_{i,t}^{(j)}$ using the bivariate estimator jointly for $i = 1, 2$.

4. For $i = 1, 2$:
   - Take $\hat{M}_{i,t}^{(j)}$, $\hat{T}_{i,t}^{(j)}$ and $\hat{W}_{i,t}^{(j)}$ and add them together to give $\hat{Z}_{i,t}^{(j+1)}$.

Coarsely, the main advantage of this approach over the univariate interpolator detailed in the previous chapter is that it performs significantly better when applied to two highly correlated series, one of which is gappy and the other is not. In effect, the non-gappy series prevents the linear predictor from decaying to zero in the middle.
of non-trivial gaps, a characteristic that is driven in the univariate case by the decay of the ACVF function. We demonstrate the performance of the interpolator in the next section.

4.2 Simulated Data

In this section we use the same structural form for our examples as in the previous chapter, extending them to the bivariate case with additive noise.

4.2.1 Bivariate Realization of Eqn. 3.35

In the previous chapter we examined the univariate interpolation of a sample series, Eqn. 3.35. We reproduce this formulation here and extend it to a bivariate realization.

\[ X_t = \left( \frac{t}{500} \right)^2 + \sin \left( 2\pi \frac{5 \cdot t}{1000} \right) + \sum_{k=1}^{3} \sin \left( 2\pi \frac{k \cdot t}{10} \right) + \zeta_t \]

\[ Y_t = X_t + \eta_t \]

for \( \zeta_t \sim \mathcal{N}(0,1) \) and \( \eta_t \sim \mathcal{N}(0,0.25) \). These series are highly correlated, which reflects their status as an example of the assumptions contained in Eqn. 4.3: observations of the same point source with additive local instrument noise. In this case we do not add instrument noise to \( X_t \), though this is entirely possible as an alternative example.

As in Chapter 3, we generate 100 realizations of \( X_t \) and \( Y_t \) for \( t \in 1, \cdots, 1000 \), and run two gap models, the first consisting of randomly scattered missing values, and the second of 10-point block-wise gaps. The algorithm of Section 4.1 is used to interpolate \( X_t \)'s missing values. Also as previously, we consider 5, 10, 15, 20 and 25% missing for each realization (reusing each realization 5 times), and then repeat the same (again reusing the realizations) for 5, 10, 15, 20 and 25% missing in 10-point blocks. An example of realizations of \( X_t \) and \( Y_t \) together with (randomly) selected blocks of interpolated values are presented in Figures 4.1, 4.2, 4.3 and 4.4.
Figure 4.1: Realization of $X_t$ for Eqn. 4.12 with 25% missing values, randomly scattered.

Figure 4.2: Realization of $Y_t$ for Eqn. 4.12: no missing values. Series is Figure 4.1 with additive $\mathcal{N}(0, \frac{1}{4})$ noise and all observations present.
Figure 4.3: Example of bivariate interpolation for Eqn. 4.12, 25% missing data, randomly scattered. Red line is interpolation, and light grey / black lines are the original series.

Figure 4.4: Second example of bivariate interpolation for Eqn. 4.12, 25% missing data, randomly scattered. Red line is interpolation, and light grey / black lines are the original series.
Now, before we present the aggregate error performance for these series, first consider what we should expect the performance to be. In the univariate case, we saw that the errors for the $g_d = 1$ interpolated points were equivalent to the distribution of the underlying noise process, due to the white character of it. In this case, we have two series which are identical up to an additive low-variance noise series. We also know from the previous chapter that the construction of the series allows $M_t$ and $T_t$ to accurately estimate the trend and sinusoidal terms, so we expect the bivariate interpolator to roughly act on the two series

$$
\hat{X}_{W,t} = X_t - \hat{M}_{1,t} - \hat{T}_{1,t} \approx \zeta_t
$$

$$
\hat{Y}_{W,t} = Y_t - \hat{M}_{2,t} - \hat{T}_{2,t} \approx \zeta_t + \eta_t.
$$

These two series have ACVF and CCVF approximately equal to

$$
R_{11}(\tau) \approx \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \quad R_{12}(\tau) \approx \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}
$$

$$
R_{22}(\tau) \approx \begin{cases} \frac{5}{4} & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \quad R_{21}(\tau) \approx \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}
$$

(4.14)

since $E \{\zeta_t \eta_u\} = 0$ for all $t, u$ and both $\zeta_t$ and $\eta_t$ are white. But then this implies that, for example, a $g_d = 1$ interpolation will have weight predominately placed upon the $Y_t$ value at that point, which as we saw above, is approximately $\zeta_t + \eta_t$. And, in practice, evaluation of the weights obtained from the bivariate interpolator shows that the largest weight by far is placed on $\hat{Y}_{W,t}$ (above, Eqn. 4.13).

The moments for these runs are displayed in Table 4.1. We see from the variance of the errors that they are only slightly higher than the underlying extra variance in the $Y_t$ series; $\sim 0.30$ versus $0.25$. There is some slight indication that there are higher error variances for series with more gaps, which is also expected. It is difficult to generalize the results for increased gap length (moving vertically down Table 4.1) as there are very few available points due to the randomly scattered gap structure. We
The higher error variances in these tables can be attributed to more error in the...
estimation of the auto- and cross-covariances. Recall that in the algorithm the series is first linearly interpolated, and then the spectrum estimated and inverted to obtain the covariances. With longer gaps, the linear interpolator has increasing levels of error, and the estimation of the covariances is correspondingly compromised.

Figure 4.5 demonstrates the variance errors graphically. As noted above, the variance of the errors for the block gap structure is uniformly higher than for the randomly scattered gap structure, running approximately at $\sigma^2 = 0.50$. While this is poorer performance than the randomly scattered case, it is still an improvement over the univariate case, although it is difficult to generalize the results given the artificial nature of the construction.

Figure 4.5: Variance of errors in bivariate interpolation using Eqn. 4.12 for five different percent-missing cases, as indicated, across gap distances of 1 to 15 points (from the closest non-missing).
<table>
<thead>
<tr>
<th>$g_d$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>966</td>
<td>1788</td>
<td>2484</td>
<td>3186</td>
<td>3694</td>
</tr>
<tr>
<td></td>
<td>0.0045</td>
<td>-0.0102</td>
<td>0.0190</td>
<td>0.0106</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>0.4184</td>
<td>0.4409</td>
<td>0.4694</td>
<td>0.4505</td>
<td>0.4724</td>
</tr>
<tr>
<td></td>
<td>-0.0839</td>
<td>0.0069</td>
<td>0.0879</td>
<td>-0.0036</td>
<td>-0.0893</td>
</tr>
<tr>
<td></td>
<td>2.8895</td>
<td>3.1739</td>
<td>2.8733</td>
<td>3.0114</td>
<td>3.1033</td>
</tr>
<tr>
<td>2</td>
<td>966</td>
<td>1788</td>
<td>2484</td>
<td>3186</td>
<td>3694</td>
</tr>
<tr>
<td></td>
<td>0.0244</td>
<td>0.0112</td>
<td>-0.0138</td>
<td>0.0088</td>
<td>-0.0005</td>
</tr>
<tr>
<td></td>
<td>0.4109</td>
<td>0.4102</td>
<td>0.4440</td>
<td>0.4582</td>
<td>0.4557</td>
</tr>
<tr>
<td></td>
<td>0.1660</td>
<td>-0.0556</td>
<td>-0.0233</td>
<td>0.0211</td>
<td>-0.0242</td>
</tr>
<tr>
<td></td>
<td>2.8684</td>
<td>3.0190</td>
<td>3.0448</td>
<td>3.1389</td>
<td>3.0511</td>
</tr>
<tr>
<td>3</td>
<td>966</td>
<td>1788</td>
<td>2484</td>
<td>3186</td>
<td>3694</td>
</tr>
<tr>
<td></td>
<td>0.0181</td>
<td>0.0120</td>
<td>-0.0150</td>
<td>-0.0281</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>0.4240</td>
<td>0.4554</td>
<td>0.4500</td>
<td>0.4669</td>
<td>0.4358</td>
</tr>
<tr>
<td></td>
<td>0.1081</td>
<td>-0.0114</td>
<td>0.0293</td>
<td>0.0205</td>
<td>0.0562</td>
</tr>
<tr>
<td></td>
<td>2.9443</td>
<td>3.1306</td>
<td>2.8197</td>
<td>3.1093</td>
<td>3.2643</td>
</tr>
<tr>
<td>4</td>
<td>966</td>
<td>1788</td>
<td>2484</td>
<td>3186</td>
<td>3694</td>
</tr>
<tr>
<td></td>
<td>0.0291</td>
<td>-0.0091</td>
<td>-0.0073</td>
<td>0.0028</td>
<td>-0.0007</td>
</tr>
<tr>
<td></td>
<td>0.4311</td>
<td>0.4220</td>
<td>0.4359</td>
<td>0.4687</td>
<td>0.4568</td>
</tr>
<tr>
<td></td>
<td>0.0548</td>
<td>-0.0546</td>
<td>0.0146</td>
<td>-0.0304</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>3.0382</td>
<td>3.1163</td>
<td>2.9403</td>
<td>3.0831</td>
<td>3.0021</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>101</td>
<td>250</td>
<td>406</td>
<td>595</td>
</tr>
<tr>
<td></td>
<td>0.0487</td>
<td>0.0438</td>
<td>0.0341</td>
<td>0.0563</td>
<td>0.0406</td>
</tr>
<tr>
<td></td>
<td>0.4776</td>
<td>0.4221</td>
<td>0.4391</td>
<td>0.4465</td>
<td>0.4278</td>
</tr>
<tr>
<td></td>
<td>0.2151</td>
<td>-0.0856</td>
<td>0.0158</td>
<td>-0.2545</td>
<td>0.0996</td>
</tr>
<tr>
<td></td>
<td>2.3034</td>
<td>2.5709</td>
<td>3.1072</td>
<td>3.0630</td>
<td>3.3399</td>
</tr>
<tr>
<td>9</td>
<td>–</td>
<td>77</td>
<td>174</td>
<td>306</td>
<td>425</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.0716</td>
<td>0.0683</td>
<td>-0.0050</td>
<td>-0.0046</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.4800</td>
<td>0.4371</td>
<td>0.4200</td>
<td>0.4426</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.5025</td>
<td>-0.0221</td>
<td>-0.1073</td>
<td>0.0561</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>3.8563</td>
<td>3.0494</td>
<td>2.9040</td>
<td>3.3777</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>39</td>
<td>98</td>
<td>165</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.3133</td>
<td>-0.0080</td>
<td>0.0232</td>
<td>-0.0223</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.4896</td>
<td>0.4062</td>
<td>0.4597</td>
<td>0.4241</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.2064</td>
<td>-0.2032</td>
<td>0.0698</td>
<td>0.0061</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>2.8266</td>
<td>3.0929</td>
<td>2.8322</td>
<td>2.5223</td>
</tr>
</tbody>
</table>

Table 4.2: Using the model of Eqn. 4.12 and the 10-point block gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, and standardized 4th moment of the collection of points that meet the distance criterion is evaluated and recorded. If less than 10 points are available across the 100 realizations, the statistics are not computed. Table shows results for only a subset of available $g_d$ distances.
4.2.2 Bivariate Realization of Eqn. 3.56

As in the previous chapter, we again consider an ARMA(4,4) model:

\[ X_t + 0.1X_{t-1} + 1.66X_{t-2} + 0.093X_{t-3} + 0.8649X_{t-4} = \]
\[ Z_t + 0.0266Z_{t-1} + 0.8175Z_{t-2} + 0.0595Z_{t-3} + 0.0764Z_{t-4}. \]  

(4.15)

To more accurately follow the form of Eqn. 4.3, we generate 100 realizations of \( X_t \), then form two new series from this underlying process as

\[ Y_t = X_t + \zeta_t \]
\[ Z_t = X_t + \eta_t, \]

(4.16)

for \( \zeta_t, \eta_t \sim N(0, 1) \) and \( E\{\zeta_t\eta_u\} = 0 \) for all \( t, u \). The realizations of \( Y_t \) then have gaps randomly placed in them in the same fashion as previously: 5, 10, 15, 20 and 25% missing and both randomly scattered and 10-point block gap structure. The series are then interpolated using the bivariate algorithm.

As in the previous example case, before we present the aggregate error performance for these series, first consider what we should expect the performance to be. In the univariate case, we saw that the errors for the \( g_d = 1 \) interpolated were distributed with significantly lower variance than the underlying noise process, a result we attributed to the autoregressive structure of the series allowing the interpolator to perform well. In this case, we have two series which begin identical, but then have \( \mathcal{N}(0, 1) \) random “instrument noise” added to them. As there are no line components or time-varying mean in these series, both \( M_t \) and \( T_t \) will fit largely to noise, and the bivariate interpolator should fit approximately on \( Y_t \) and \( Z_t \).

Write:

\[ \hat{Y}_{W,t} \approx Y_t \]
\[ \hat{Z}_{W,t} \approx Z_t. \]

(4.17)

With the additive “instrument” noise, the two series \( Y_t \) and \( Z_t \) have \( \sigma^2 = 4.955^3 \),

\(^3\)Determined using the technique of [117].
Table 4.3: Using the model of Eqn. 4.16 and randomly scattered gaps, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, and standardized 4th moment of the collection of points that meet the distance criterion is evaluated and recorded. If less than 10 points are available across the 100 realizations, the statistics are not computed.

<table>
<thead>
<tr>
<th>$g_d$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4978</td>
<td>9865</td>
<td>14469</td>
<td>18796</td>
<td>22487</td>
</tr>
<tr>
<td></td>
<td>0.0065</td>
<td>0.0044</td>
<td>0.0131</td>
<td>0.0031</td>
<td>0.0077</td>
</tr>
<tr>
<td></td>
<td>2.2648</td>
<td>2.6806</td>
<td>3.2402</td>
<td>4.1314</td>
<td>5.0569</td>
</tr>
<tr>
<td></td>
<td>0.0090</td>
<td>0.0093</td>
<td>-0.0127</td>
<td>0.0007</td>
<td>-0.0301</td>
</tr>
<tr>
<td></td>
<td>2.9941</td>
<td>3.0416</td>
<td>3.1601</td>
<td>3.3204</td>
<td>3.9108</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>133</td>
<td>518</td>
<td>1136</td>
<td>2273</td>
</tr>
<tr>
<td></td>
<td>-0.4177</td>
<td>0.1308</td>
<td>0.0625</td>
<td>-0.0655</td>
<td>0.0545</td>
</tr>
<tr>
<td></td>
<td>0.8608</td>
<td>2.7800</td>
<td>3.0328</td>
<td>4.1982</td>
<td>5.5027</td>
</tr>
<tr>
<td></td>
<td>-0.1722</td>
<td>-0.2223</td>
<td>-0.2068</td>
<td>-0.0085</td>
<td>-0.0177</td>
</tr>
<tr>
<td></td>
<td>3.6473</td>
<td>3.5656</td>
<td>2.9384</td>
<td>3.4191</td>
<td>4.0147</td>
</tr>
<tr>
<td>3</td>
<td>–</td>
<td>–</td>
<td>13</td>
<td>66</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>0.2220</td>
<td>-0.2986</td>
<td>-0.2868</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>2.8304</td>
<td>4.2571</td>
<td>4.6703</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>-0.2728</td>
<td>0.3871</td>
<td>-0.1866</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>3.7090</td>
<td>3.5443</td>
<td>3.3567</td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>-0.1912</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>4.0271</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.5527</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>3.8139</td>
</tr>
</tbody>
</table>

since the ARMA(4,4) portion of each of them is independent of the noise terms. As there are no periodic components in these series, and we expect both $M_t$ and $T_t$ to be effectively fit to noise, we expect the variance of the errors to converge upwards towards the variance of the underlying process (taken as samples only), that is, $\sigma^2 = 4.955$.

The results of Table 4.3 (simulation using randomly scattered gaps) effectively show the results we expect. At low percentages of missing data, the errors have
<table>
<thead>
<tr>
<th>$g_d$</th>
<th>5%</th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>960</td>
<td>1804</td>
<td>2506</td>
<td>3188</td>
<td>3690</td>
</tr>
<tr>
<td></td>
<td>0.0028</td>
<td>0.0287</td>
<td>-0.0163</td>
<td>-0.0968</td>
<td>-0.0229</td>
</tr>
<tr>
<td></td>
<td>3.0452</td>
<td>3.3333</td>
<td>3.8661</td>
<td>4.2394</td>
<td>4.6222</td>
</tr>
<tr>
<td></td>
<td>0.0856</td>
<td>0.0054</td>
<td>-0.0061</td>
<td>0.0386</td>
<td>0.0832</td>
</tr>
<tr>
<td></td>
<td>3.0633</td>
<td>3.1069</td>
<td>3.0606</td>
<td>3.2639</td>
<td>3.2603</td>
</tr>
<tr>
<td>2</td>
<td>960</td>
<td>1804</td>
<td>2506</td>
<td>3188</td>
<td>3690</td>
</tr>
<tr>
<td></td>
<td>0.1437</td>
<td>-0.0288</td>
<td>0.0261</td>
<td>0.0011</td>
<td>0.0134</td>
</tr>
<tr>
<td></td>
<td>3.0677</td>
<td>3.4526</td>
<td>3.8439</td>
<td>4.2585</td>
<td>4.5497</td>
</tr>
<tr>
<td></td>
<td>0.1185</td>
<td>0.0282</td>
<td>0.0411</td>
<td>-0.0010</td>
<td>-0.0744</td>
</tr>
<tr>
<td></td>
<td>3.2215</td>
<td>3.0777</td>
<td>3.3918</td>
<td>3.3365</td>
<td>3.2277</td>
</tr>
<tr>
<td>3</td>
<td>960</td>
<td>1804</td>
<td>2506</td>
<td>3188</td>
<td>3690</td>
</tr>
<tr>
<td></td>
<td>-0.0180</td>
<td>0.0327</td>
<td>-0.0202</td>
<td>0.0407</td>
<td>-0.0330</td>
</tr>
<tr>
<td></td>
<td>2.9109</td>
<td>3.2824</td>
<td>3.5266</td>
<td>4.0508</td>
<td>4.4584</td>
</tr>
<tr>
<td></td>
<td>-0.1341</td>
<td>0.0035</td>
<td>-0.0053</td>
<td>0.0667</td>
<td>-0.0047</td>
</tr>
<tr>
<td></td>
<td>3.1638</td>
<td>3.1397</td>
<td>2.9852</td>
<td>3.1754</td>
<td>3.4739</td>
</tr>
<tr>
<td>4</td>
<td>960</td>
<td>1804</td>
<td>2506</td>
<td>3188</td>
<td>3690</td>
</tr>
<tr>
<td></td>
<td>0.0627</td>
<td>0.0489</td>
<td>0.0049</td>
<td>0.0310</td>
<td>-0.0153</td>
</tr>
<tr>
<td></td>
<td>3.0139</td>
<td>3.4746</td>
<td>3.6830</td>
<td>4.3479</td>
<td>4.5306</td>
</tr>
<tr>
<td></td>
<td>-0.0235</td>
<td>-0.0484</td>
<td>-0.0059</td>
<td>-0.0799</td>
<td>0.0578</td>
</tr>
<tr>
<td></td>
<td>3.1288</td>
<td>3.1117</td>
<td>3.1107</td>
<td>3.2516</td>
<td>3.3327</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>113</td>
<td>245</td>
<td>394</td>
<td>624</td>
</tr>
<tr>
<td></td>
<td>0.3009</td>
<td>0.1968</td>
<td>-0.1464</td>
<td>-0.0027</td>
<td>0.0384</td>
</tr>
<tr>
<td></td>
<td>4.0484</td>
<td>3.3690</td>
<td>3.9504</td>
<td>3.8546</td>
<td>4.7262</td>
</tr>
<tr>
<td></td>
<td>-0.0087</td>
<td>-0.1029</td>
<td>-0.1941</td>
<td>0.0110</td>
<td>0.1517</td>
</tr>
<tr>
<td></td>
<td>2.7598</td>
<td>3.1878</td>
<td>3.4504</td>
<td>3.1295</td>
<td>4.6059</td>
</tr>
<tr>
<td>9</td>
<td>–</td>
<td>60</td>
<td>182</td>
<td>259</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.2441</td>
<td>0.0199</td>
<td>0.2362</td>
<td>0.0416</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.4021</td>
<td>-0.1562</td>
<td>0.1174</td>
<td>-0.1936</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>2.7822</td>
<td>2.6310</td>
<td>3.4064</td>
<td>5.0123</td>
</tr>
<tr>
<td>10</td>
<td>–</td>
<td>29</td>
<td>112</td>
<td>152</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>-0.1281</td>
<td>-0.0790</td>
<td>-0.1485</td>
<td>-0.1412</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>2.5488</td>
<td>3.2909</td>
<td>4.7538</td>
<td>4.2080</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>0.0962</td>
<td>-0.4290</td>
<td>0.3546</td>
<td>0.0888</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>3.2881</td>
<td>2.8906</td>
<td>4.2586</td>
<td>2.7407</td>
</tr>
</tbody>
</table>

Table 4.4: Using the model of Eqn. 4.16 and the 10-point block gap structure, each missing point across all 100 realizations is evaluated for distance from the nearest non-missing point (first column). For each set of realizations, the mean, variance, skew, and standardized 4th moment of the collection of points that meet the distance criterion is evaluated and recorded. If less than 10 points are available across the 100 realizations, the statistics are not computed. Table shows results for only a subset of available $g_d$ distances.
variance $< 3$, and as the percentages climb, the error variances climb with them toward $\sigma^2 \approx 5$. Table 4.4 shows the results for the 10–point block gap structure, and reflects similar behaviour. The variance of the 25% missing errors is interesting, as it shows very little variation along the $g_d$s. This is graphically presented in Figure 4.6. The performance of the bivariate interpolator on this set is particularly intriguing due to the fact that it seems to scale very well with gap length and with percentage missing. In the case of 5% missing, the variance of the errors is $\approx 3$ through increasing $g_d$; contrast this with the underlying process variance of $\approx 5$.

We make one final note about the bivariate interpolator applied to the ARMA(4,4) example. This is effectively a worst-case example of a pair of series for which this
interpolator would be tried for the following reasons: a) the series have no periodic or trend components; b) the instrument noise is large relative to the process magnitude; c) the high degree of correlation in the underlying series will cause the cross-covariance weights to be quite high, carrying the full noise sequence into the final estimate.

4.3 Real Data: Revisiting the Solar Flux

![Bivariate interpolation of daily noon solar flux from Sagamore Hill observatory, Massachusetts. The second series used was the flux from Penticton, DRAO.](image)

Figure 4.7: Bivariate interpolation of daily noon solar flux from Sagamore Hill observatory, Massachusetts. The second series used was the flux from Penticton, DRAO.

The daily noon solar flux series, jointly observed at Sagamore Hill, Massachusetts and Penticton, British Columbia, was the original impetus behind the development of the hybrid bivariate interpolator. Data sets like this do occur quite often in the physical sciences (i.e. environmental observations like humidity or temperature) and
we would like to be able to reconstruct missing values from one series with some reliability. As we briefly discussed in Section 3.5, these series are observations of the noon solar flux, measured in the aptly named “Solar Flux Units”, using distinct instruments at geographically isolated sites. Despite this, the observations are of the Sun, and we do not expect them to differ drastically between sites – in fact, we are counting on the fact that these two series are so similar. In addition, this series has a number of very strong periodic components in it, which allows the $T_t$ estimator to shine.

As in the previous section, we restrict attention to blocks of at most 1000 points in order to work around the strong changes in variance that occur with the solar cycle. Figure 4.7 shows an example of the fill using the same time span used previous in Figure 3.10.

As in the univariate case, there is little we can do to determine the accuracy of this fill as we do not know the missing values. However, recall\footnote{The reader is forgiven if this has been forgotten in the lengthy discourse that followed.} that the original purpose of developing these interpolators was to allow us to examine the spectrum of the process. In Figure 4.8 and 4.9 we compare the spectra of the Penticton DRAO daily noon solar flux and the bivariate interpolator copy of the Sagamore Hill flux. These figures show that the majority of the structure of the power spectra are the same through these frequencies – the main difference is that the Sagamore Hill series has a higher level of variance as $f \to f_N$, the Nyquist frequency. Some level of error variance is unavoidable, and this variance comes through as spread-spectrum noise.

\section*{4.4 Conclusions}

We presented an alternative interpolation technique designed for the case where two series share an underlying process structure (up to and including the same line and
trend components). This technique proved to be quite strong for the first case of line components against a white background, but performed more poorly for the case of high noise against an autoregressive-type background. In the case of the daily noon solar flux, the interpolator performed very well, providing reasonable and consistent estimates of the missing measurements of the Sagamore flux, comparing very well with the complete Penticton series against which it was interpolated. Contrasting the spectra of the two series shows that a formerly gap-ridden daily flux series has been converted to a contiguous estimate of daily noon solar flux with similar structure and behaviour to other observational series of the same source.
Figure 4.9: Multitaper power spectra comparison between Penticton and Sagamore Hill - II. Continuation (in frequency) of Figure 4.8.
Chapter 5

Smoothers and Generalized Additive Models for Epidemiologic Time-Series Analysis

The work presented in this chapter was done in conjunction with Dr. Glen Takahara of Queen’s University. The work was inspired and indirectly funded by Health Canada through Contract # 4500250926. The primary contact at Health Canada was Hwashin H. Shin, PhD, Project Scientist in the Population Studies Division. Health Canada is gratefully thanked for allowing the presentation of models run on privately-held data.

5.1 Epidemiologic Models and Air Pollution

It seems clear from a basic understanding of human health that breathing in unclean air can have adverse health effects. We begin with two extreme examples. The condition known colloquially as “black lung” which so often affects coal miners is one of the most visible of these possible effects [169], and there is a clear connection between the extended periods spent below-ground breathing suspended particulate
coal dust and the miner’s later illness, and often, death.

In December 1958, the combination of an extreme (for England) cold front and an eventual temperature inversion\(^1\) caused 5 days of extreme loss of visibility due to tremendous amounts of coal smoke and suspended particulate matter. Estimates range from 3,000 to 14,000\(\mu g/m^3\) for the PM\(_{10}\) concentration\(^2\) in the air in London during December 5–10; typical levels today are around 30 \(\mu g/m^3\) [169]. The estimated number of deaths due to this prolonged period of extreme smog was 12,000. Approximately two-third of the deaths were of people older than 65, and infant mortality also doubled during the period.

Given these two well-documented results, an obvious question is whether or not we can statistically tie atmospheric presence of low levels of PM\(_{10}\) to increased risk of mortality or morbidity\(^3\). While it is well-established that extreme amounts of exposure (such as a life-long coal miner might experience) will almost certainly result in premature death due to diseases of the cardiorespiratory system, it is much more difficult to measure the effects of short-term exposure on the subtle scales we might find in a typical North American city. As noted by Dockery and Pope [48], what remained (at that time) to be determined was the level of pollution that would significantly affect human health.

Beginning in the early 1990s and proceeding through to the present, tremendous research effort has been expended in exploration of the impact of relevant (i.e., at levels comparable to everyday city exposures) concentrations on health. For a review up to 2003, see [65], and for some of the work in the 10 years since, see [64, 1, 144, 47, 77]. It should also be noted that environmental stressors have strong health effects (e.g.,

---

\(^1\)A temperature inversion is a meteorological condition where the air close to the ground is cooler than the air above it.

\(^2\)In keeping with common usage, we will denote particulate matter of size \(\leq 10\) micrometers in diameter as PM\(_{10}\); similarly, particles of size \(\leq 2.5\) micrometers will be denoted as PM\(_{2.5}\).

\(^3\)We take morbidity to mean an increased rate of disease, and typically tie it to the number of hospital admissions as a convenient source of data.
temperature), and significant work has been completed on temperature-mortality re-
sponse relationships, especially in the context of distributed lag or distributed lag
non-linear models (dlnlms), see [62, 66]. In addition, we note here that we must be
careful in linking air pollution directly to the respiratory and/or cardiovascular sys-
tems, as these arguments are not fully developed and may not actually be correct.
Having said this, it seems clear that there is some form of population health risk due
to air pollution, and the estimation of this relationship is our primary focus in what
follows.

When attempting to statistically estimate the population health risk due to sus-
pended particulate matter or air pollution, the first difficulty is deciding what data
is appropriate to use. This is better posed as an epidemiology question, as we seek a
scientific basis for including variates in our models. It seems clear that using mortal-
ity counts is the closest we can get to a direct effect, as the counts measure the actual
deaths that occur. It is equally clear that most deaths are unrelated to air pollution –
most people who die do so due to entirely separate causes. It thus behooves us to
restrict our attention to deaths which can at least be attributed to cardiorespiratory
or cardiopulmonary causes or disease; that is, to deaths which are tied to the lungs or
the circulatory system. From a medical perspective, this ties the deaths to the origi-
nal “attack vector” – the lungs, and the air the victims breathe in the time preceeding
their deaths. Mortality data in Canada and the United States is categorized accord-
ing to the International Classification of Diseases (ICD) system [125], using either
version 9 or 10 depending on the year. Under this system it is possible to distinguish
between accidental, all-cause non-accidental (AC), non-accidental cardiopulmonary
(CP) and non-accidental non-cardiopulmonary (NCP) deaths.

In any statistical model, we seek to include explanatory variables to account for
confounding\textsuperscript{4} between our response and primary predictor. If we were to use a model which attempted to fit mortality as some link function applied solely to air pollution, we would obtain a risk estimate which would only be valid if all correlation between the two variables corresponded to our sought-after risk. However, this is clearly not the case for mortality and air pollution – PM\textsubscript{10} varies seasonally, as does mortality, although they are out-of-phase with one another\textsuperscript{5}. Simply because the two variables are not in phase does not mean that the relationship should be negative. Similarly, we can easily imagine a situation where an environmental factor is the trigger for death, but where this factor varies in a correlated fashion to the pollutant.

After season and trend, the most important predictors of mortality are weather terms. Plausibly, the weather variates that are most likely to be associated with mortality (especially cardiopulmonary mortality) are temperature and humidity. Both are known to exacerbate respiratory illnesses like asthma, and especially in the case of temperature, can cause death directly in extreme cases. However, if we include a term (or terms) in the model to account for the seasonality and trend of the mortality, the weather variables will be present to carry information about the effects of short-term variations in weather on mortality [170]. Further, the dependence of mortality on weather is typically nonlinear: very low temperatures (near or below the freezing point of water) contribute to increased mortality\textsuperscript{6}, then as temperature increases mortality decreases smoothly up to a point. However, at very high temperatures (e.g., 30° C or higher), the combination of increased heat and corresponding

\textsuperscript{4}A confounder is an extraneous variable in a statistical model that correlates with both the dependent and independent variable. This contrasts with interaction, which refers to the simultaneous influence of two (or more) variables on another variable failing to be additive.

\textsuperscript{5}Note that PM\textsubscript{2.5} is not in phase with NO\textsubscript{2} and Ozone, at least for Canadian data. Instead it appears to be several months behind mortality, but not the approximate six months of (say) NO\textsubscript{x} pollutants.

\textsuperscript{6}This relationship contributes to increased mortality in the winter together with influenza epidemics.
humidity contributes to sharp rises in mortality. This possibly gives the overall relationship a $U$ or $V$ shape. For an example, see the top panels of Fig. 6.11. To model this nonlinear relationship, typically a smooth function of temperature with a small number of degrees-of-freedom is typically used. In this context, degrees of freedom refers to the number of basis vectors used in the creation of the smoother’s model matrix, e.g. if working with a $df = 4$ cubic regression spline, there are a total of 4 effective free parameters in the expansion, consisting of the coefficients of the $B$-spline basis vectors. This will be made more clear later in this chapter.

As well as weather variates, other contributing factors that should be examined in a mortality–pollutant model include day-of-week (as is well-known, deaths are typically highest on Mondays), and possibly additional pollutants. Finally, a term is conventionally included to compensate for slowly-varying risk factors.

This last element deserves more thought. Mortality has a strong seasonal variation, largely driven by the seasons of our Northern Hemisphere climate. However, this strong seasonal variation is also present in most air pollutants, influenza epidemics, and other strong risk factors. This is especially true for unmeasured confounding terms – there are any number of causal factors that contribute to non-accidental cardiopulmonary mortality (e.g. flu epidemics, extremely high temperatures, out-of-season unusual temperatures, etc.), and most of them are not measured routinely or are not easily incorporated when they are measured. Accordingly, these patterns are too common to allow them to contribute to the correlation between air pollution and daily mortality. Thus, the question becomes: at what frequency (or time-scale) do we believe that this contribution decreases to acceptable levels? And having determined such a frequency, how do we ensure that these patterns are removed from consideration?
5.2 The Use of Time-Based Smoothers

As briefly mentioned in the previous section, it is clear that it is impossible to include all possible confounding variates when modeling health risk due to air pollution. Examples of variates that fail to be included can be split into two classes: variates that are not included because researchers are not aware that they have causal impact, and variates that are not included because it is not possible to accurately measure them. In epidemiologic models, examples of the latter type of confounder include influenza epidemics (which typically peak in mid-winter for the northern hemisphere) and long-term demographic shifts. As we will demonstrate in the following brief literature review, the impact of these confounding variates has been considered. The solution proposed was the use of smooth functions of time that vary on long timescales, that is, timescales longer than a few weeks.

We begin in 1990 with the first significant study by Schwartz and Allan [170]. The authors consider the problem

[...] that even within a single season, collinearity with omitted weather variables that vary slowly over the season (long wave-length) may artificially enhance the correlation. A number of approaches have been used to eliminate such longer-term effects from mortality data so as to better assess other suspected causes of excess daily mortality. These include subtraction of sinusoidal (seasonal) trends and subtraction of 15-day moving averages of daily mortality⁷. None of these have dealt with the high degree of short-term autocorrelation in the data and the subsequent lack of independence of the regression residuals.

The context was estimation of population health risk due to air pollution, specifically a time-series analysis of data from London during the era of British smoke, i.e., heavy

⁷Note that this 15-day moving average is a non-causal filter.
concentrations of suspended soft coal particulate matter that led to severe health effects including death. In 1990 when Schwartz and Allan’s paper was written, the best approach any researchers had tried was to subtract seasonal trends and moving averages from the mortality series. What was intended was to obtain a high-pass filtered version of the series, but this was not as effective as hoped. It is important to not criticize the researchers overly much, as computing resources were rarer and less accessible than they are now, and researchers in epidemiology may not have even been aware of much of the work being done in time series and signal processing at that time.

Four years later in 1994, Dockery and Pope [48] discussed the (then current) state of the art, precisely demonstrating the issues facing researchers in this area.

Epidemiologic studies suffer from the weakness that observed associations with a specific exposure may result from an unmeasured association with an unknown or uncontrolled factor correlated with both exposure and disease, i.e., from a confounder. These time-series studies have the advantage that many major causes of increased mortality (such as smoking, hypertension, or even age) cannot confound the observed associations with particulate air pollution because these factors do not vary with daily pollution exposures. This is not to say that response may not differ by these factors. Indeed, the mortality effects of particulate air pollution are most strongly seen in the elderly in all those studies in which age is considered as an effect modifier.

The authors also discussed the lag structure inherent in time-domain time series studies, saying:

Another interesting feature of these studies is the consistent finding of lagged associations between particle exposure and increased mortality [...]
[Research suggests] that mortality effects of particulate air pollution may be lagged by several days.

The same authors, continuing what would become a decades-long endeavour, continued [170] studying the methodology of these epidemiologic models, saying:

In all of epidemiology, a basic issue in modelling is to control properly for potential confounding […] Since any two variables that show a long term trend must be correlated, searches for correlations that are more likely to be causal must exclude these trends. These trends may not be linear […] A second common attribute of many variables that evolve over time is seasonality […] it is necessary to remove these patterns. They are often described as long wavelength patterns […]

The authors continue discussing the removal of these long wavelength patterns, which they use as a synonym for low-frequency, proposing use of either a moving average (specifically using a 15-day span) or weighted moving average (kernel smoothers). The language in the discussion makes it clear that the researchers are interested in the residuals (deviation from [the] moving average) as the primary variate of interest. As such, while not explicitly framing the problem as that of filtering, it seems clear that a high-pass filter is what is being attempted. Certainly both 15-day moving average and weighted average smoothers are equivalent to linear filters [74] in a broad sense, although filters are usually formed as causal.

As this second paper was published in 1996, Dockery and Pope took advantage of the then-recent publication of Hastie and Tibshirani’s monograph on Generalized Additive Models (GAMs) [74] and proposed the use of these models. This formulation allowed them to treat temperature and other continuous variables as smooth functions. They also advocated use of a smooth function of time as a way of accounting
for long-timescale variation in the mortality. They further suggest the use of regression splines as smoothers, and go so far as to begin the discussion regarding what cutoff is appropriate for removal of low-frequency unmeasured confounding structure. Finally, the authors conclude with a discussion of which weather variables are likely to be plausibly associated with mortality from an epidemiology perspective. Their conclusion is that both temperature and humidity are important, but they leave open the question of what metric should be used, i.e., whether maximum, minimum, mean, or some other measure of geographically-local daily temperature/humidity. We note that there was other significant research during this time period which came to many of the same conclusions (e.g., [26]), and it is difficult to determine where many of the ideas truly originated. Further, it is difficult to determine how best to include temperature and humidity, as they are highly correlated (and in fact humidity is a function of temperature under most metrics).

The Health Effects Institute (HEI) was founded in 1980, and is dually funded by both the United States Environmental Protection Agency (EPA) and world-wide automotive manufacturers. They have funded a number of critical studies, including both the Particle Epidemiology Reanalysis Project (with which the paper [27] and report [102] are associated) and the National Morbidity and Mortality Air Pollution Study (to be described below).

In the paper by Burnett et al. [27], the authors used GAMs to model the relationship between environmental factors and daily hospital admissions. Rather than include a semi- or non-parametric smoother in their GAM, they instead pre-filtered the time series of hospital admission counts using a 31-day LOESS. Interestingly, the authors also pre-filtered the air pollution time series to remove temporal trends and day-of-week effects. These pre-filtering ideas will reappear later in this work.

We now examine two papers by Kelsall, Zeger, and Samet from 1997 and 1999.

---

8Read: low-frequency.
that examined log-linear frequency domain models. These papers were the inspiration for several of the innovations we will review below. We begin with the 1997 paper [87], which examined a data series for Philadelphia.

In the 1997 paper [87] the authors first formally advocated the use of log-linear regression using GAMs. They additionally split the covariates into parametric (linear) and non-parametric (non-linear smooth) relations. To control for unmeasured time variables they used smoothing splines of time to represent trends in mortality that were mainly related to either cardiovascular diseases or respiratory infections.

The 1999 Kelsall et al. paper [88] continues with a Poisson regression model, modeling the logarithm of the mean mortality count as an additive combination of environmental and time covariates. This paper introduced frequency-domain methods to this area of research. Unfortunately, the authors estimated the coefficients for a log-linear Poisson model in the frequency domain using a severely biased power spectrum estimate, the periodogram. The conclusions of the paper are therefore suspect as it is extremely difficult to obtain reasonable results when using an estimator such as the periodogram (see Section 5.4 for more details).

The HEI began funding a new project in 1999 (mentioned above) called the National Morbidity and Mortality Air Pollution Study (NMMAPS), which eventually led to the release of the NMMAPS database[142, 140]. The principal investigators on this study included Scott Zeger, Francesca Dominici and Jonathan Samet, all (at that time) researchers affiliated with Johns Hopkins Bloomberg School of Public Health. The first paper published by these researchers was published in 1999 [235], and examined the idea of harvesting. The key idea was whether increased mortality associated with higher pollution levels could be restricted to very frail persons who might be expected to die shortly whether or not they were exposed to the air pollution, a possibility labeled as the harvesting hypothesis. The authors extended the ideas of [87, 88] and created harvesting-resistant estimates by excluding the short-term information
that would be affected by harvesting.

Zeger et al. made no suggestion that one time scale component of pollution would (or could) have any effect on another time scale; that is, there are no cross-correlations computed. Additionally we note that in the authors’ attempt to find a harvesting-resistant estimate of the effect due to air pollution, they used a weighted average of the time-scale specific estimates, each of which was already biased significantly due to bias in their technique for band-pass filtering the spectrum estimates. Thus their final conclusions should be considered carefully.

We now turn to the seminal paper of the decade, written by Dominici, Samet, and Zeger and published in 2000 [52]. In this paper the authors incorporate most of the developments of the previously mentioned papers, excepting only the frequency-domain log-linear modeling of [87, 88]. They carefully detail a GAM for measuring geographically-localized risk due to air pollution, and modelled daily expected deaths as a function of the pollution levels on the same or immediately preceding days, not of the average exposure for the preceding month, season or year as might be done in a study of chronic effects, and included smooth functions of time as predictors so as to avoid confounding by both influenza epidemics (which are seasonal) and by other longer-term factors.

The key paragraph in the paper for the purposes of understanding the reasoning of Dominici et al. for inclusion of the smooth function of time follows.

To protect the pollution relative rates $\beta^c$ from confounding by longer-term trends due, for example, to changes in health status, changes in the sizes and characteristics of populations, seasonality and influenza epidemics, and to account for any additional temporal correlation in the count time series, we estimated the pollution effect using only shorter-term variations
in mortality and air pollution. To do so, we partial out the smooth fluctuations in the mortality over time by including arbitrary smooth functions of calendar time $S^c_{\text{time}, \lambda}$ for each city. Here, $\lambda$ is a smoothness parameter which we prespecified, on the basis of prior epidemiological knowledge of the timescale of the major possible confounders, to have 7 degrees of freedom per year of data so that little information from timescales longer than approximately 2 months is included when estimating $\beta^c$. This choice largely eliminates expected confounding from seasonal influenza epidemics and from longer-term trends due to changing medical practice and health behaviours, while retaining as much unconfounded information as possible. We also controlled for age-specific longer-term and seasonal variations in mortality, adding a separate smooth function of time with 8 degrees of freedom for each age group.

The degrees-of-freedom referred to in this quote are, in context, the number of knots used in the natural cubic regression spline smooth of the mortality with respect to time. As such, the conclusion that this spline preserves little information from timescales longer than 2 months is not actually correct, as we will show later in this chapter. Dominici et al. attempted to determine the efficacy of their approach by examining the Bartlett autocorrelation function of the standardized residuals:

To test whether the log-linear generalized additive model has taken appropriate account of the time dependence of the outcome, we calculate, for each city, the autocorrelation function of the standardized residuals. Fig. 4 displays the 20 autocorrelation functions; they are centred near zero, ranging between -0.05 and 0.05, confirming that the filtering has removed the serial dependence.

[...] The standardized residuals are $r_t = \frac{y_t - \hat{y}_t}{\sqrt{\hat{y}_t}}$ where $y_t$ is the response
series (mortality counts) and $\hat{y}_t$ is the prediction series from the model fit.

Unfortunately, estimating the autocorrelation of the residuals using a Bartlett autocorrelation estimator is known to be severely biased in the presence of low-frequency power \[7, 202\], so it is uncertain if these results are actually meaningful. A better approach might be to examine a consistent estimate of the power spectrum of the residuals and attempt to classify its nature. We will discuss this further in the following sections.

There was a large-scale “crisis” in the field in 2002 when Dominici, Samet and Zeger realized that the standard errors estimated by \texttt{SPlus} were biased due to computational concerns. They published two papers on this topic; the first detailing the problem \[50\], and the second proposing a solution \[49\]. The problem with the errors was due to default convergence parameters in the \texttt{SPlus} statistical function \texttt{gam} (written by Trevor Hastie) being too lax. This paper also noted that GAMs may fail to detect concurvity\[^9\] in some situations, leading to underestimation of the standard error of the coefficients for the pollutant. A huge amount of effort went into re-examination of previously completed analyses, with the final conclusion that the original conclusions had been correct, although the exact risk estimates were corrected by the updated methodology.

Exploring the problems with concurvity in these GAM risk models, a paper by Ramsay, Burnett and Krewski \[149\] was published in 2003. This paper presented results showing that the inability of GAM to detect concurvity can lead to misleading statistical inferences. Specifically, estimates of standard errors can be biased quite badly, leading to overstatement of the significance of associations (risk estimates).

These findings were further developed by Dominici, McDermott and Hastie. \[49\] The authors highlighted confounding bias as an important issue, and showed how the control for this bias (smooth functions of time and temperature) led to the convergence

\[^9\]The nonparametric analogue of collinearity.
issues first detailed in [50]. The paper further developed a new routine `gam.exact` which gave a closed-form solution for the asymptotically exact covariance matrix of the linear component of the GAM. This routine was further integrated into the R package `gam` [76], ported from SPlus by Trevor Hastie.

In 2004 an additional researcher (Michelle Bell of Yale) joined the Hopkins team, and Bell, Samet and Dominici [10] presented further analysis with additional descriptions of the problem of unmeasured confounding. This work was continued with the further addition of Roger Peng [51], a computational biostatistician who would contribute significant efforts towards the creation of the NMMAPS database and associated R packages [142, 140].

The work mentioned so far was primarily concerned with estimation of risk over many years, and essentially took one coefficient for the risk over the span. Careful reading of previous papers (including [48, 27] etc.) show that there was considerable interest in breaking down the risk estimates to shorter time scales. In addition, several proposals had been made to estimate risk under the assumption that the risk itself varied with season. The primary paper of interest to us in this context is that of Peng, Dominici and Louis [141] which is almost entirely concerned with the problem of unmeasured confounders, especially on seasonal and long-term time scales. The simulations in this paper appear to show that the use of increased degrees of freedom for semi-parametric regression splines is not necessary. As we will discuss later in this chapter, there are elements of this analysis which do not hold up under careful frequency-domain inspection. We include here several quotes from this paper:

Welty and Zeger (2005) developed a rich class of distributed lag models that were specifically targeted at adjusting for temperature in multicity time series studies of air pollution and mortality. This class of models includes a variety of predictors such as running means of temperature,
non-linear functions of running means, multiple lags of temperature and interactions between temperature at different lags. They applied their models to the NMMAPS database and found that the national average estimate of the effect of PM$_{10}$ (PM with aerodynamic diameter less than 10 $\mu$m) on total non-accidental mortality is robust to a large class of statistical models that are used to adjust for potential confounding by temperature and dewpoint temperature. Building on these findings, in this paper we focus on the problem of controlling for unmeasured confounders, i.e., seasonal and long-term trends.

With respect to implementation, they say

A common approach to adjusting for seasonal and long-term trends is to use semiparametric models which incorporate a smooth function of time [...]. This approach can be thought of as regressing residuals from the smoothed dependent variable on residuals from the smoothed regressors. In this setting, the smooth function of time serves as a linear filter on the mortality and pollution series and removes any seasonal or long-term trends in the data. Several alternatives for representing the smooth functions have been applied including smoothing splines, penalized splines and parametric (natural) splines (Dominici et al., 2002; Ramsay et al., 2003; Schwartz et al., 2003; Touloumi et al., 2004; Health Effects Institute, 2003). The smooth function of time naturally accounts only for potential confounding by factors which vary smoothly with time. Factors which vary on shorter timescales may also confound the relationship between air pollution and mortality and controlling for them is an important concern.

These quotes provide much of the inspiration for the work that follows in this chapter.
The review presented in this section is not by any means complete. In particular, we have not explicitly discussed the second reason for including the smooth function of time. By their nature, the mortality counts that we use as response are correlated, yet in the regression model formulation, we assume that they are distributed as independent Poisson random variables. This is patently incorrect, and the secondary reason for including the smooth function of time is to filter these observations so that they retain little-to-no autocorrelated structure. Of course, as we will see in the next section, this does not fully happen, and in fact even when filtering the counts to remove all low-frequency power, there remains high-frequency structure which negates the assumption of independence.

This area of consideration is incredibly rich, with hundreds of researchers contributing toward the goal of accurate estimates of health risk due to air pollution. We do note that there are numerous papers that have been published in the last decade which do not add appreciably to the statistical framework. Many of these papers take for granted the results of, e.g., [52, 50, 49, 141], and simply apply the models to new datasets. We hope that by presenting the problems with this framework (in Section 5.4) and our proposed solution (in Section 5.5) we have added somewhat to the level of rigor and understanding in the field.

5.3 Cubic Splines as GAM Smoothers

As reviewed in the previous section, the standard model used for epidemiologic time series analyses includes a smooth function of time to account for unmeasured confounding terms. The default choice for this smooth function of time has become parametric natural cubic splines with \( \sim 7 \text{ dof/yr} \) [52, 49, 141]. This corresponds to 7 simple knots per year for a natural cubic regression spline (B-spline in implementation) basis. In \( \mathbb{R} \) the computation of basis vectors for these cubic splines is done
through the `ns()` routine which is included in the base package `splines`, and largely leverages the algorithms of de Boor [45]. More details on this topic are given below.

### 5.3.1 Spline Theory

*The material in this section is drawn from Wood [226], de Boor [45], Wahba [207] and Green and Silverman [67].*

The use of cubic splines as smoothers or interpolators is not a new idea, going back to the age of draftsmen using thin, flexible strips of metal to draw curved lines. Extensive theoretical work in the mid- to late-20th century moved splines to the new world of computing. This work culminated in the classic book by Carl de Boor [45] with much of the mathematical derivations behind this book given in Schoenberg [164]. Many of the algorithms contained in that book have yet to be significantly improved and much of the underlying code in modern spline fitting still relies on them.

For simplicity we begin with a parsimonious model that has only one covariate, a smooth function of a single variable:

\[ y_i \mid \beta = f(x_i) + \epsilon_i, \]  

(5.1)

where \( y_i \) is a response variable, \( x_i \) a covariate, \( f \) an arbitrary smooth function and \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \). We use \( \mid \beta \) to indicate “modeled by”\(^{10}\). To estimate \( f \) requires that it be represented in such a way that (5.1) becomes a linear model. To do this we choose a basis such that \( f \) (or an approximation to it) is an element of the vector space spanned by the basis. We choose this basis, then \( f \) is assumed to have the \( q \)-dimensional representation

\[ f(x) = \sum_{j=1}^{q} b_j(x) \beta_j, \]  

(5.2)

for some scalar values of \( \{ \beta_j \} \) and basis functions \( \{ b_j(x) \}_{j=1}^{q} \).

\(^{10}\)In R, this is expressed by using \(~\).
One formulation that is of particular importance is the cubic spline. Such a spline is made up of segments of cubic polynomials joined together so the segments are continuous up to second derivative. The points at which the segments join are called knots. Given a set of points \( \{x_i, y_i\} \) for \( i = 1, \ldots, n \), and \( x_i < x_{i+1} \), the natural cubic spline \( g(x) \) that interpolates these points is a function with support \([x_1, x_n]\) made up of \( n - 1 \) cubic polynomials that are joined so the entire spline is continuous up to second derivative, \( g(x_i) = y_i \) and \( g''(x_1) = g''(x_n) = 0 \). This last condition is what gives the natural modifier to the name. The reason the natural cubic spline is important is because of all functions that: a) are continuous on \([x_1, x_n]\), b) have absolutely continuous first derivatives, and c) interpolate \( \{x_i, y_i\} \), \( g(x) \) is the function that is smoothest in the sense of minimizing

\[
J(f) = \int_{x_1}^{x_n} f''(x)^2 \, dx. \tag{5.3}
\]

A proof for this can be found in [67, Ch. 2]. In addition, cubic spline interpolators have a number of additional desirable properties, see [45, Ch. 5].

Now, restricting ourselves to \( g(x_i) = y_i \) (under the formulation above) leaves little room for change in the structure of the interpolation. If instead we remove the restriction \( g(x_i) = y_i \), and treat the \( g(x_i) \) as \( n \) free parameters, estimating them in order to minimize

\[
\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 \, dx, \tag{5.4}
\]

with \( \lambda \) a tuneable parameter, results in a smooth \( g(x) \) that is a smoothing spline [157]. The only apparent downside to smoothing splines is that they have as many free parameters as data points to be smoothed. As \( \lambda \) will constrain the resulting function to be smoother than \( n \) degrees-of-freedom would suggest, an obvious compromise [226] is to use penalized regression splines instead.

A penalized regression spline is the obvious extension of the natural cubic regression spline with the difference that the total number of knots being used is chosen as...
The knots are typically chosen to be evenly spaced between \( x_1 \) and \( x_n \), although variations occur depending on the particular form of cubic regression spline basis sought. Note that in this formulation we retain the \( g(x_i) = y_i \) restriction of the natural cubic spline. One simple formulation uses the non-orthogonal (under the \( L^2 \) norm) basis

\[
\{ b_1 = 1, b_2 = x, b_3 = x^2, \ldots, b_n = x^{n-1} \}.
\]

Other formulations include those which give simple interpretations of expansion coefficients or which are especially suited for computational efficiency and stability [45].

The *B-spline basis* is an appealing representation for cubic (and higher order) regression splines because the basis functions have strictly local support, each basis function being non-zero for only the intervals between 5 adjacent knots. This generalizes to higher order splines, so that the basis functions are non-zero for \( p + 2 \) adjacent knots, for \( p \) the order of the spline. However, to define a \( k \) parameter B-spline basis we require \( k + p \) knots \( x_1 < x_2 < \cdots < x_{k+p} \), where the spline evaluation support is only \([x_{p+1}, x_k]\). This allows the first \( p \) and last \( p \) knots to be essentially arbitrary [226], and gives representation

\[
f(x) = \sum_{i=1}^{k} B_i^m(x) \beta_i. \tag{5.5}
\]

Originally, B-splines were invented to provide a stable basis for very large scale spline interpolation [45], and they continue to be used as their derivation and computation are reasonably simple, and they have too many optimal properties to be easily replaced.

There are incredible depths to the theory of splines as smoothers for GAMs which are not even touched on in this limited overview. For a full exposure to the theory, we suggest beginning with Hastie and Tibshirani [74] (or [75]), Wood [226, Ch. 4] and de Boor [45], then progressing to Wahba [206, 207] and finally, for the determined, to Green and Silverman [67], Kim and Gu [91], and Wang [210].
5.3.2 GAM Theory

The previous section gave an overview of the structure of natural cubic regression splines. In this section we will discuss the practical aspects of implementation of these splines, both with respect to the generation of the basis vectors, and also their inclusion in GAMs, especially with respect to the convergence behaviour of the GAM solvers as implemented in the R packages `gam` and `mgcv`.

We begin with the necessary framework that allows a GAM to be set up as a penalized Generalized Linear Model (GLM). In general, a GAM models some response $y_i$ using a model structure similar to

$$g(\mu_i) = X_i^*\theta + \sum_{j=1}^{k} f_j(x_{ji})$$  \hspace{1cm} (5.6)

with $\mu_i = E\{y_i\}$, $y_i$ having either an exponential family distribution or a more relaxed “quasi-likelihood” mean-variance relationship, and the $f_j$’s smooth functions of the $x_j$’s. Note that we do not know the $\mu_i$ a priori – in fact, the estimation of this expected value is the point of the modeling process. To estimate such a model we begin by defining a basis for each smooth function $f_j$, in form similar to the B-spline basis defined in Eqn. 5.5:

$$f_j(x_j) = \sum_{i=1}^{q_j} \beta_{ji} b_{ji}(x_j)$$  \hspace{1cm} (5.7)

with the $\beta_{ji}$ the coefficients of the smooth. Given these bases, a model matrix $\tilde{X}_j$ is created for each smooth.

The form given in Eqn. 5.6 is typically not an identifiable model\(^{11}\). Loosely, this lack of identifiability stems from the smooths having non-zero mean; these means can be effectively moved between parameters at will. Identifiability is obtained if each smooth is subjected to a “centering constraint” [226, Ch. 4.2], where the sum (or mean) of the elements of $f_j = \tilde{X}_j\beta_j$ (i.e., a vector of the smooths) should be zero.

\(^{11}\)Identifiable is used in the statistical sense to indicate that it would be possible to learn the true values of the underlying model parameters if an infinite number of observations were gathered.
Note that $f_i = f_j(x_{ji})$ and $\tilde{X}_{jik} = b_{jk}(x_{ji})$, that is, the subscripts are themselves vectors. Typically this constraint is absorbed into a re-parameterization – this is explicitly the form that mgcv implements in practice. This re-parameterization is reasonably straightforward, and is represented by a single Householder matrix.

Given centered model matrices for each smooth term, $j = 1, \ldots, k$, we can rewrite Eqn. 5.6 as

$$g(\mu_i) = X_i \beta$$

where $X = [X^*|X_1|X_2|\cdots]$ and $\beta^T = [\theta^T, \beta_1^T, \ldots, \beta_k^T]$, where the $[x|y|\cdots]$ notation indicates a block concatenated matrix. Then this form is clearly a simple GLM, and it is thus possible to write its likelihood. One can also see that if the $q_j$ (of Eqn. 5.7) are large and $\beta$ is estimated using ordinary likelihood maximization, the chances of overfitting are high. For this reason one of two alternative approaches to the standard GLM form is typically used, namely (i) Penalized Iteratively Reweighted Least Squares (P-IRLS), or (b) Backfitting. These two approaches are discussed in the following.

5.3.2.1 Solving GAMs

Penalized Iteratively Re-weighted Least Squares, as described by Wood [226] and more historically derived in O’Sullivan et al. [126], is an extension of the original IRLS algorithm detailed by Nelder and Wedderburn [122], and extended in McCullough and Nelder [116]. Begin with a penalized likelihood for the model given in Eqn. 5.8,

$$l_p(\beta) = l(\beta) - \frac{1}{2} \beta^T S \beta,$$

where $S = \sum_j \lambda_j S_j$, the $\lambda_j$ are assumed to be known and the $S_j$ are matrices of known parameters for each smoother. Note that the quadratic form in the penalty is due to $f$ being linear in the parameters $\beta_j$ [226, pp.128,142]. Maximizing $l_p$ with respect to $\beta$ through the usual derivative technique leads to equations which are identical to
those that would be solved for the maximization of the penalized linear least-squares problem [226, pp.63–66], i.e.

\[ \mathcal{S}_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\text{var}(Y_i)} + \beta^T S \beta. \] (5.10)

where the \( \text{var}(Y_i) \) are assumed to be known. Continuing in the penalized linear least-squares vein, in the vicinity of some vector estimate of the parameter, \( \hat{\beta}^{[k]} \), where the superscript \([k]\) denotes the iteration, the first term of \( \mathcal{S}_p \) can be approximated

\[ \mathcal{S}_p = \left\| \sqrt{W^{[k]}} (z^{[k]} - X \beta) \right\|^2 + \beta^T S \beta, \] (5.11)

with \( z^{[k]} \) a vector of pseudodata (the current best estimate) and \( W^{[k]} = \text{diag}(w^{[k]}_i) \).

Then

\[ w^{[k]}_i = \left( V(\mu^{[k]}_i) g'(\mu^{[k]}_i)^2 \right)^{-1} \]

\[ z_i = g'(\mu^{[k]}_i) \left( y_i - \mu^{[k]}_i \right) + X_i \hat{\beta}^{[k]} \] (5.12)

In this setup \( V(\mu^{[k]}_i) \) is a link function defined as \( V(\mu) = b''(\theta)/\omega \) such that \( \text{var}(Y) = V(\mu) \phi \) in the standard GLM formulation [116]. As such, \( V(\mu^{[k]}_i) \) is obtained from the current estimate of \( \beta \). The maximum penalized likelihood estimates \( \hat{\beta} \) are obtained by iterating

1. Current \( \mu^{[k]} \), calculate \( V(\mu^{[k]}) \) and then calculate \( z^{[k]} \) and \( w^{[k]}_i \).

2. Minimize Eqn. 5.11 with respect to \( \beta \) to obtain \( \hat{\beta}^{[k+1]} \). Evaluate \( \eta^{[k+1]} = X \hat{\beta}^{[k+1]} \) and obtain fitted values \( \mu^{[k+1]}_i = g^{-1}(\eta^{[k+1]}_i) \).

These steps are iterated to convergence.

*Backfitting*, a second approach to solving GAMs, is detailed in Hastie and Tibshirani [74] and reviewed in [75, 226]. The idea is to estimate each smooth component individually by iteratively smoothing partial residuals from the additive model with respect to the covariate(s) associated with the component. The partial residuals result from the subtraction of all current model term estimates (excepting the current
component’s) from the response variable.

Formally, suppose we are estimating an additive model

\[ y_i = f_0 + \sum_{j=1}^{k} f_j(x_{ji}) + \epsilon_i, \]  

(5.13)

with the \( f_j \) being smooth functions, the \( x_{ji} \) covariates, and \( y_i \) the response variable. If \( \hat{f}_j \) is a vector of estimates of \( f_j(x_{ji}) \), i.e., \( \hat{f}_j = [\hat{f}_j(x_{j1}), \hat{f}_j(x_{j2}), \ldots, \hat{f}_j(x_{jn})] \), then the backfitting algorithm begins by setting \( \hat{\alpha} = \overline{y} \) as the estimate of the mean response, and \( \hat{f}_j = 0, \ j = 1, \ldots, k \) and iterating the following:

1. Update the estimate \( \hat{\alpha} \).

2. For \( j = 1, \ldots, k \), repeat:

   (a) Calculate the partial residuals

   \[ e^j_p = y - \hat{\alpha} - \sum_{m \neq j} \hat{f}_m \]

   (b) Set \( \hat{f}_j \) equal to the result of smoothing \( e^j_p \) with respect to \( x_j \).

until the estimates \( \hat{f}_j \) converge. This algorithm does not make explicit the means of actually smoothing \( e^j_p \) with respect to covariate \( x_j \), which adds to the flexibility and leaves the choice up to the user. Several downsides include the inability to efficiently include generalized cross-validation techniques (for penalization parameters), high computational burden in general, and high computational burden for direct estimation of the “hat” matrix of the whole model.

In practice, \texttt{mgcv} actually uses a combination of backfitting and P-IRLS. The \textit{interior performance iteration} uses P-IRLS for estimation of parameters in the usual way, but the \textit{outer iteration} uses a form of backfitting to move between smooths\(^ {12} \).

\(^ {12} \)Interestingly, the interpolation algorithms derived in previous chapters also use a form of backfitting – it is a very general technique!
5.3.3 Practical Code Notes for spsmooth

\texttt{mgcv} is a complicated software package, including many routines and techniques which are outside the focus of the present work. There are two things that we need for the following discussion. The first is the idea of \textit{partial residuals} mentioned briefly in the previous section. If the \texttt{gam} solver of \texttt{mgcv} is iterated to convergence, and one wishes to examine the partial residual response that one covariate of interest was fit to, the partial residual is computed as the sum of the overall model residual and the \textit{prediction} for the covariate of interest. This makes some intuitive sense, but is not documented explicitly in \texttt{mgcv}; the solution was provided by the package author, Dr. Simon Wood, in an email sent to the \texttt{R} help list.

The second needed item is a brief note about extension of the package. \texttt{mgcv} was explicitly written to provide support for user-supplied smoothing methods. The “hook” that allows these smoothers to be included does a namespace search in \texttt{R} for generator and predictor methods for the smoothing method, and then integrates them into the overall outer/inner iteration loop seamlessly. For further information about the implementation of \texttt{spsmooth}, see Appendix A or the package help files \cite{mgcv_manual}.

5.4 Problems with Use of Cubic Spline Smoothers

As a specific demonstration, we consider data from the city of Chicago, Illinois. The data set consists of 14 years of daily measurements and is contained in the NMMAPS database \cite{NMMAPS}. For simplicity we consider only three variates: daily all-cause mortality\textsuperscript{13}, daily mean aggregate temperature, and daily (trimmed) mean aggregate PM\textsubscript{10} measurements. These variates are included in the model of Dominici \textit{et al.}\textsuperscript{13}

\textsuperscript{13}As mentioned in an earlier section, there are three types of mortality that are recorded and can be considered: all-cause (AC), cardio-pulmonary (CP), and non-cardio-pulmonary (NCP). All three exclude accidental death, and CP mortality is obtained by aggregating all deaths which fall under the ICD-9 or ICD-10 classification system categories for deaths relating to the cardiorespiratory system.
[52, 141, 140], although they are not the full complement of those models. This particular example was also considered by Wood in his 2006 book [226, pp. 247–253].

Using the arguments detailed in Section 5.2 and at the start of this chapter, we assume that the observed number of deaths are Poisson random variables with a time-varying mean that is the product of a time-varying death rate and several pollution or environmental effects. We further assume for simplicity that the covariates are uncorrelated with one another. Although this is almost certainly false, it is necessary for the construction of the basic model, and much of the work that follows attempts to work around the inherent problems with this assumption. Write

\[
\log(\text{death}_t) = \beta \cdot \text{pm10tmean}_t + f_1(\text{temperature}_t) + f_2(\text{time}_t) + \text{DOW},
\]

where we additionally assume that temperature is related to mortality in a nonlinear

Figure 5.1: Daily All-Cause Mortality for Chicago, 1987–2000. Red line shows the fitted values from model in Eqn. 5.14.
(yet smooth) way, as discussed above in the introductory comments of the chapter. In the above equation, death is the daily counts of mortality, pm10tmean is the PM$_{10}$ observational variate, and DOW refers to a factor variable representing the day-of-week. As discussed in Section 5.2, we use a cubic regression spline with a fixed number of degrees-of-freedom per year as our choice for $f_2$; in this case, we use 7 $dof$/year (knots/year) to match the choice made in [52] and confirmed by [141]. We additionally use the Peng et al. [141] choice of 6 (fixed) degrees-of-freedom (knots) for the smooth function of temperature.

There are a number of missing values in the pm10tmean series (253 out of 5114 days). Unfortunately, as briefly reviewed in Chapter 3, missing values prevent accurate estimation of power spectra. We linearly interpolate the missing values$^{14}$, and compute the model from (5.14) for both the interpolated and non-interpolated series. The model coefficients are shown in Table 5.1, and demonstrate that while there is a small difference in the results obtained, they are close enough that graphical demonstrations using the interpolated series should provide intuition about the behaviour of the non-interpolated series. A graphical presentation of the fitted values from Eqn. 5.14 using the non-interpolated series is shown in Figure 5.1. The bulk of the fit comes from the smooth function of time.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.7281878</td>
<td>0.0037498</td>
<td>4.7271804</td>
<td>0.0038722</td>
</tr>
<tr>
<td>pm10tmean</td>
<td>0.0003968</td>
<td>0.0000893</td>
<td>0.0003634</td>
<td>0.0000952</td>
</tr>
</tbody>
</table>

Table 5.1: Examination of coefficients for Chicago using (5.14).

Now, recall what the smooth function of time was supposed to accomplish. We desire a term which, included in the model, accounts for long timescale (low frequency)

$^{14}$We could also have used the interpolator from Chapter 3, but as this is merely for demonstration, the additional speed of linear interpolation made it a simpler choice.
variation in the mortality that would be otherwise confounded with both unmeasured (influenza epidemics, demographic shifts, etc.) and measured (temperature, etc.) confounding variates. Although papers in the field imply that the choice of timescale cutoff was due to epidemiologic concerns, this does not seem to be the case\textsuperscript{15}, and the timescale cutoff was actually chosen as a minimization procedure using sensitivity analysis. Having chosen this cutoff as 7 cycles/year (or approximately 7 weeks), if we examine the residual effective response after accounting for time only, we hope to see a series which contains very little low-frequency power. In fact, the ideal is to have none at all, although it is impossible to achieve a perfectly rectangular high-pass filter response for a finite data sequence. There is a fundamental problem with the timescale assumption, which is that assuming that a cut-off of $1/7$ years corresponds to a regression spline smoother with 7 degrees-of-freedom (knots) per year. As we see in the following, a cubic regression spline smoother using 7 knots per year has an approximate cut-off of $2/7$ years, or $\sim 3.5$ months!

Using the multitaper spectrum estimate, we examine the residual effective mortality after accounting only for the smooth function of time; that is, what is jointly fit to by the other terms in the model of Eqn. 5.14. A plot of the spectra thus generated appears in Figure 5.2. The blue vertical dashed line is set at a cutoff of 7 cycles/year, or $\approx 52$ days. In an optimal case, the red curve should follow the black (GAM) curve from the right of the plot through to the cutoff (blue dashed line), and then drop dramatically and stay “low” through to the origin at the far left of the plot. This behaviour would be appropriate for a high-pass linear filter, which is what this natural cubic spline smoother is intended to be. Note that the light grey background curve is the periodogram estimate corresponding to the identical time series as the red curve. The exceptionally high variance of this curve demonstrates in one figure why the periodogram is not suitable for serious analysis. We have included the periodogram in

\textsuperscript{15}Personal communication with Dr. Mark Goldberg
Figure 5.2: Demonstration of multitaper power spectra for Chicago using Eqn. 5.14. The black curve is the multitaper spectrum of the log mortality as seen by the GAM. The red curve shows the multitaper spectrum of the residual effective mortality as seen jointly by the temperature, pollutant, and day-of-week term in the model. For both multitaper spectrum estimates, \(NW = 4\) and \(K = 7\). The light grey curve in the background is the periodogram of the same data as the red curve. The blue dashed line is the cutoff commonly thought to correspond to a 7 dof/yr natural cubic regression spline smoother.

Two significant problems are clear from Figure 5.2: (i) the residual effective mortality shown (red line) does not have a cutoff at 7 cycles/year, and in fact the black and red curves demonstrate near-equality between 4 and 7 cycles/year, and (b) the residual effective mortality has extremely strong periodic structure at 1 and 2 cycles/year. The periodicity at 1 cycle/year (due to the Earth’s orbit around the Sun) is also pronounced in environmental covariates such as temperature and pollution,
and is also part of the “long timescale variation” that was discussed in Section 5.2. The periodicity at 2 cycles/year is an artifact of the nonlinear filtering that is done when smoothing temperature against residual mortality, the $f_1(\text{temperature}_t)$ term in Eqn. 5.14. If anything, these periodicities (structure below 7 cycles/year) are what the smooth function of time was included in the model to remove. It is clear from this figure that the smooth function of time has *not* removed all of the periodic low-frequency variation it was intended to eliminate.

A *transfer function* is a frequency-domain representation of the characteristics of a filter. The *magnitude-squared transfer function*, or magnitude response, shows the behaviour of the pass-band (the frequency region in which the original series structure is preserved) and the attenuation-band (in which the power is decimated). Most interestingly for our purposes, the magnitude response function also demonstrates the frequency cut-off point between the pass and attenuation bands, and provides information about the strength of the out-of-band attenuation provided by the filter. As a note, typical “naive” linear filters provide several orders of magnitude attenuation outside the pass-band, while excellent examples of filters can provide 10 or more orders of magnitude [57]. Figure 5.3 shows the empirical transfer function that is theoretically equivalent to the unweighted cubic regression spline used in Figure 5.2. There are two differences between the effective transfer function implied by Figure 5.2 and that in Figure 5.3. The first is that the P-IRLS solver uses adaptive (variance-based) weighting on the data points available, so that the final filter is non-linear. The second is that the P-IRLS solver is iterative, so that the filter that is applied is not being applied to the original data each time, but rather to a residual series equivalent to the difference between the log mortality (response) and the current “copies” of the other covariates. Despite these differences, the empirical (unweighted) transfer function in Figure 5.3 is similar enough to the actual initializing transfer function to be a useful guide for interpretation.
Figure 5.3: Empirical magnitude response function for a natural cubic regression spline defined with 7 dof/yr, applied to 14 years of data. This curve is the average of 20 realizations of the empirical result of applying the projection operator to simulated noise $\sim N(0, 2)$. Note the high primary sidelobe magnitude and the extremely slow drop-off beginning at 3 cycles/year. The shaded region represents the magnitude response of an ideal rectangular low-pass filter with cut-off frequency $\sim 7$ cycles/year. The red curve is the high-pass filter corresponding to the effective transfer function after removing the output of the low-pass filter demonstrated as the black curve.

The empirical magnitude response transfer function shown in Figure 5.3 is that of the low-pass filter associated with a 99 dof (knot) natural cubic regression spline smoother. The effective filter applied to the mortality data, as seen by the pollutant, is the converse of this (in theory, ignoring the P-IRLS implications). It may be difficult from Figure 5.3 to visualize what this would look like, so we demonstrate this converse as the red curve in the same figure. This curve provides insight into
the shape of the red curve in Figure 5.2, showing the decrease beginning at $\sim 4$ cycles/year. The cut-off frequency of the red curve appears to be 4 cycles/year with a slow transition from the stop-band to the pass-band.

![Figure 5.4: Comparison of the spectrum of the log mortality (in black) and the residual after accounting for the smooth function of time (in red).](image)

The key to understanding the difference between the red curve of Figure 5.3 and the demonstrated effective residual mortality after accounting for the smooth function of time shown in Figure 5.2 is the iterative process inherent in the P-IRLS solver. If we assume that the first term (of the four) that is fit is the smooth function of time, then the solver will initialize with the transfer function shown in Figure 5.3. While the inner iteration loop introduces weighting that will modify the transfer function somewhat, the majority of the difference demonstrated occurs, not due to the weighting, but due to the inclusion of the other terms. As a quick demonstration
of this, consider a simplified model:

$$\log(\text{death}) \sim s(\text{time}, df = 7 \cdot 14) \quad (5.15)$$

with $s(\cdot)$ being a cubic regression spline with fixed degrees of freedom, and the residual effective mortality being simply the residuals. The spectrum of this residual is shown in Figure 5.4. We see that the inclusion of the weights has modified the effective transfer function only marginally, preserving the shape between 1 and 7 cycles/year. This is unfortunate but not directly a target for improvement as the weighting is an inherent part of the P-IRLS scheme, and the impact is not severe.

### 5.4.1 Bias in the Coefficients

Due to the iterative nature of the solver in mgcv, it is possible to examine more closely the effect of the residual effective magnitude response transfer function demonstrated in the previous section. Upon convergence, the addition of the overall model residuals and any one prediction\(^{16}\) provides the residual effective response specific to the corresponding predictor. This allows us to examine what the pollutant is fit to (in a model sense). Further, as the pollutant is a parametric term in the overall model, we can reproduce the obtained coefficient by using a very simple linear model. Denote the residual effective mortality as fit to by the pollutant as $\text{death}_{\text{resid}, t}$, the overall model residuals as $r_t$ and the coefficient corresponding to the pollutant $\text{pm10tmean}$ as $\beta$. Then

$$\text{death}_{\text{resid}, t} = r_t + \beta \text{pm10tmean}_t \quad (5.16)$$

and we estimate the model

$$\text{death}_{\text{resid}, t} \mid \text{pm10tmean}_t, \quad (5.17)$$

with the model having mean constrained to be zero\(^{17}\).

---

\(^{16}\)The individual predictions being the solved-for coefficients expanded on the covariate of interest.  
\(^{17}\)In R, write this as $\text{lm}(\text{death}_{\text{resid}, t} \sim \text{pm10tmean} - 1)$.  

119
If we apply this model, Eqns. 5.16 and 5.17, using the standard linear model approach and being careful to use the same model weights as were obtained in the \texttt{gam} solver fit from \texttt{mgcv}, we obtain the identical coefficient for the pollutant. This is not entirely surprising (as it is by construction), but it is reassuring. A demonstration of this is given in Table 5.2 together with the next topic, below.

As we have access to such a simple model (Eqns. 5.16 and 5.17), we can easily consider what the results would be if we manually high-pass filter the response (\texttt{death}_{\text{resid},t}) to ensure that it has minimal power below the cutoff of 7 cycles/year. This can be done with any of an enormous family of linear filters, but we specifically choose to use a Slepian projection filter for its simplicity of form (as an orthogonal family, its projection filter form is particularly clean). As was mentioned in Chapter 2.1, prolate spheroidal wave functions (sequences) were developed in work at Bell Labs [173, 95, 96, 171, 172] and are now referred to as \textit{Slepians}. Among all families of finite length (indexlimited) sequences they are unique in having the most concentrated spectrum in the band $-W \leq f \leq W$. They are doubly orthogonal, over both $0,1,\ldots,N$ and $-\infty,\ldots,-1,0,1,\ldots,\infty$, and have an odd/even symmetry. In addition, the simply projection filter formed from the Slepian sequences has zero phase distortion and an excellent transition from pass-band to stop-band. For references on filter design, see [203, 128, 57].

Selecting the cutoff $W \in (0,1/2)$ and with a fixed data length $N$, the projection subspace corresponding to the projection filter has $\sim 2NW$ concentrated orthogonal basis vectors [172]. Formally, if we have all $N$ Slepian sequences $\left\{u_i^{(k)}\right\}_{k=0}^{N-1}$ for a given $W$ and $N$, and we truncate this series to the $\sim 2NW$ sequences with the largest eigenvalues, we create the projection operator which acts on a sequence $X_t$, also of length $N$. Begin by writing the expansion

$$S^T X_t = (y_0, y_1, \ldots, y_{2NW-1})^T = \sum_{k=0}^{2NW-1} u_i^{(k)} \cdot X_t.$$  (5.18)
and then using the expansion coefficients, write the projection as

$$SS^T X_t = \begin{pmatrix} v_t^{(0)} & v_t^{(1)} & \cdots & v_t^{(2NW-1)} \end{pmatrix} \begin{pmatrix} y_0 & y_1 & \cdots & y_{2NW-1} \end{pmatrix}^T$$

$$= \begin{pmatrix} v_t^{(0)} & v_t^{(1)} & \cdots & v_t^{(2NW-1)} \end{pmatrix} \begin{pmatrix} v_t^{(0)} & v_t^{(1)} & \cdots & v_t^{(2NW-1)} \end{pmatrix}^T X_t.$$  \hfill (5.19)

This can be written in matrix form by taking

$$S = \begin{bmatrix} \mid & & \mid \mid & v_t^{(0)} & v_t^{(1)} & \cdots & v_t^{(2NW-1)} \mid & \mid \end{bmatrix},$$ \hfill (5.20)

and forming the $N \times N$ matrix

$$\mathcal{H} = S \left(S^T S\right)^{-1} S^T.$$ \hfill (5.21)
Note that in the case of all data being available, as the Slepian sequences are orthogonal, the center \( S^T S \) term is the \( 2NW \times 2NW \) identity matrix. Finally, returning to our residual effective mortality (5.16), write the filtered copy as

\[
(death_{\text{resid},t})_{\text{HP}} = death_{\text{resid},t} - \mathcal{H} \cdot death_{\text{resid},t}.
\]  

(5.22)

This quantity is then the difference between the original residual effective mortality and the projection and expansion of it using the Slepian sequences. The projection filtered copy is a low-pass filter of the residual effective mortality, and subtracting it gives the corresponding high-pass filter result. Figure 5.5 shows the projection of the residual effective mortality of Eqn. 5.16 as the black curve plotted over the light grey.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Interpolated Coefficient</th>
<th>Std. Err.</th>
<th>Not Interpolated Coefficient</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original ( \text{gam} ) ( \text{mgcv} )</td>
<td>3.968</td>
<td>0.893</td>
<td>3.634</td>
<td>0.952</td>
</tr>
<tr>
<td>Linear Model - (5.17)</td>
<td>3.968</td>
<td>0.719</td>
<td>3.634</td>
<td>0.752</td>
</tr>
<tr>
<td>Linear Model - (5.23)</td>
<td>3.432</td>
<td>0.699</td>
<td>3.141</td>
<td>0.731</td>
</tr>
</tbody>
</table>

Table 5.2: Examination of coefficients for Chicago using Eqn. 5.14, Eqn. 5.16 and Eqn. 5.23. All coefficients and standard errors \( Z \) should be read as \( Z \times 10^{-4} \), i.e. they have been multiplied by \( 10^4 \). The reader should notice the identical coefficients between the original \( \text{gam} \) fit and the (5.17), and the change when (5.23) is estimated.

Applying Eqn. 5.22, we obtain a new response which is regressed on the pollutant in a linear model formulation:

\[
(death_{\text{resid},t})_{\text{HP}} \parallel \text{pollutant}
\]

(5.23)

under the constraint that the mean is zero. The results obtained by using this implementation are given in Table 5.2 along with the results from Eqns. 5.16 and 5.17. The reader will note that in both the interpolated and non-interpolated cases, the coefficient obtained from the manually pre-filtered response is somewhat smaller than from the unmodified response. Recall the interpretation of this: the GAM solver gives
one coefficient (top row of table). The explicit linear model formulation of Eqn. 5.17 gives the same coefficient and different standard errors (as the errors are computed in context with the other covariates, this is expected). Manually high-pass filtering the response used in Eqn. 5.17, we re-estimate the coefficient (third row of table), and observe a difference. This tells us that the portion of the residual effective mortality below 7 cycles/year does have an impact on the estimation of risk – it was possible that there would not have been a significant difference even with the poor filtering performance of the natural cubic regression spline smoother.

Now that we have shown that there is a difference, we ask the logical next question: how much power is actually in the frequency range 0–7 cycles/year in these two implementations? Applying spectral techniques we obtain estimates of this power, and present the findings in Table 5.3. These results indicate that while there may only be $\sim 5\%$ of the power remaining below the frequency cut-off, this power has a disproportionate effect on the coefficients obtained from the model, as can be seen in Table 5.2 (as the change in coefficient is larger than the relative percentage of power). In order to ensure the interpretation of the coefficient is valid, we would like to ensure this power is as small as possible. This will additionally eliminate this potential source of bias. Recall the discussion of Section 5.2 for more details on the interpretation.

We have shown that inclusion of low-frequency power in the residual effective mortality, even at levels as low as $\sim 5\%$, changes the coefficient obtained for risk due to air pollution. Note that this is not necessarily a bad thing – we present it as a problem due only to the epidemiologic rationale given for including the smooth function of time. If we wish to account for long timescale variation in the response in order to prevent confounding from unmeasured variates, then leaving small proportions of variation is unwarranted. The presentation thus far has not provided a solution to this problem. More will be said on this in Section 5.5. We further note that we have
Table 5.3: Estimation of the percentage of power (variance) below the frequency cut-off of 7 cycles/year using Eqns. 5.16 and 5.23. For the non-interpolated case we have estimated the power in the time domain as a percentage of variance, using a gap-enabled high-performance low-pass filter. Comparing the percentage power remaining with the coefficients of Table 5.2 we see that going from a small percentage of power remaining below 7 cycles/year to an insignificant amount causes a significant change in the coefficient.

5.4.2 One Further Problem

Thus far we have seen two interrelated problems with the use of natural cubic regression splines as smoothers for time-based variates. The first problem was the effective cut-off of the high-pass filter not corresponding to the timescale selected for epidemiologic reasons. The second was simply another way of looking at the same problem, showing that this un-attenuated power was having a significant influence on the risk estimates obtained. A third problem, as mentioned at the end of the last section, is the question of how the 1 and 2 cycles/year periodicities appeared in Figure 5.2.

The answer to the third question is simple enough when we consider the P-IRLS iterative solver (see Section 5.3.2.1). Both the temperature and the pollutant contain extremely strong low-frequency structure. These are demonstrated in Figures 5.6 and 5.7. In both cases the frequencies corresponding to 1 and 2 cycles/year have been noted with vertical red dotted lines, and the frequency cut-off at 7 cycles/year with vertical blue dashed lines. The interesting “decaying” shape to the left of 1 cycle/year in the pm10tmean series appears to be an artifact of the averaging process used to
Figure 5.6: Multitaper power spectrum of daily temperature for Chicago, 1987–2000, variable name tmpd from the NMMAPS database. No missing values. NW = 4 and K = 7. The extremely strong peak is the annual cycle, 1 cycle/year.

create the aggregated series from multiple data stations [142, 140].

Now, we consider how the P-IRLS solver acts on these predictors. Consider a thought experiment where we have a smooth function of time as the first (non-parametric) predictor (call this $A$), and the second (parametric) predictor (call this $B$) is an additive combination of two pure sinusoids at (say) 1 cycle/year and 30 cycles/year. Formally, assuming the two predictors are measured daily,

$$A = f(t) = \sum_{j=1}^{q} \beta_j b_j(t)$$  \hspace{1cm} (5.24)

for $f$ expanded as a spline basis, and

$$B = \alpha_1 \sin \left(2\pi \frac{t}{365}\right) + \alpha_2 \sin \left(2\pi \frac{30t}{365}\right).$$  \hspace{1cm} (5.25)
Further assume that the response $R$ is composed of two pure sinusoids at 1 and 30 cycles/year together with a noisy background,

$$R = \gamma_1 \sin \left(2\pi \frac{t}{365}\right) + \gamma_2 \sin \left(2\pi \frac{30t}{365}\right) + \zeta_t$$

(5.26)

for $\zeta_t$ some arbitrary noise sequence.

Now, referring back to the standard epidemiological model (see Section 5.1), use of a 7 $dof$/year smoother is thought to account for the variation in the response below $\approx 7$ cycles/year. For the purposes of this discussion, we will assume that we have an ideal high-pass filter as a smoother, so that the smoother perfectly accounts for all variation in the response below 7 cycles/year. In this case, the parametric fit of $R$ regressed on $B$ should be largely driven by the 30 cycles/year periodicity. Now,
step through a model of the P-IRLS algorithm, where $p(K)$ represents the current prediction of $K$:

1. Initialize solver: both $A$ (all spline bases) and $B$ coefficients set to 0.
   - $A$ is smoothed against $R$, obtaining a series containing the strong 1 cycle/year periodicity from $R$ together with the low-frequency noise.
   - $B$ is fit to the residuals of $R - p(A)$; largely, the fit will be due to the coherent structure at 30 cycles/year.

2. Iterate the following until convergence:
   - $A$ is smoothed against $R - p(B)$. Because $p(B)$ is a parametric term, for non-zero coefficient, $p(B)$ will contain multiples of both periodic components. The idealized filter (smoother) will then capture only the part of the annual cycle that is not accounted for by the multiple of $B$.
   - $B$ is smoothed against $R - p(A)$; the coefficient will be based on a combination of some part of the annual cycle (as $A$ cannot capture all of it) and the complete 30 cycle/year structure, as well as the noise.

As we see from the above example, due to the parametric nature of the pollutant, any non-zero inclusion of this predictor will account for some of the low-frequency power, in particular any periodic structure that is common between the two. The smooth function of time cannot correct for this as it is smoothed on the residual effective mortality after accounting for the pollutant (and other terms). In other words, it can only smooth (and account for) what is actually left after fitting the other terms, and parametric terms will be spread-spectrum.
5.5 Proposal: Slepian Projection Filters as Smoothers

As a best first attempt, we tried to correct only for the first cause of bias listed in the previous section. Our goal was to use a more optimally concentrated linear filter so as to better capture the low-frequency structure of the log mortality. We use Slepian projection filters for this purpose as well – the discussion of the previous section and the simplicity and optimality properties of the Slepian sequences makes them an excellent choice. We elect to focus on a very simple form of filter that uses the \( K = 2NW \) most concentrated Slepian sequences as a projection filter. This is the exact form mentioned in the previous section to examine the amount of low-frequency power remaining in the residual effective mortality.

For information on the implementation of these filters as smoothers we refer the reader to Appendix A. For this discussion, all we need is the knowledge that the Slepian projection smoother class behaves similarly to the natural cubic regression spline smoother class, and the two are interchangeable for data analysis. By default, as mentioned in the previous section, the hat matrix for a Slepian projection filter is only \( SS^T \) as the \( (S^T S)^{-1} \) term is the identity. This does not hold when using weighted regression (as the P-IRLS algorithm can be shown to be equivalent to). In the weighted formulation, the regression “hat” matrix [116] can instead be written as

\[
\mathcal{H} = \sqrt{W} S (S^T W S)^{-1} S^T \sqrt{W},
\]

where the \( W \) is the weight matrix in the standard GLM setup. This equation demonstrates that the Slepian projection filter may be vulnerable to the same nonlinearity from the weights as was the natural cubic regression spline filter. However, the advantage of \texttt{spsmooth} is that it begins from a higher frequency concentration point than \texttt{ns}. In addition, we remind the reader that \textit{this} nonlinearity is typically not that significant, as hinted at by Figure 5.4.

Returning to the model of Eqn. 5.15 we replace the basis for the smooth function
Figure 5.8: Comparison of the empirical transfer functions of Equations 5.28 and 5.15. Note how the \texttt{spsmooth} transfer function has a steep drop-off immediately after the 7 cycles/year point, and has minimal sidelobes. The extreme quantization in the red curve, especially around 8 cycles/year, is due to the \textit{empirical} representation and the performance of the smoother on the extremes of the data series (start and end).

of time with the new \texttt{spsmooth} smoother:

\[
\log(\text{death}) = \text{pm10tmean} + s(\text{tmpd}, df = 6) + s_2(\text{time}, W = 7 \text{cycles/yr}) + \text{as.factor(DOW)}
\]  \hspace{1cm} (5.28)

where \(s_1\) is a cubic regression spline, and \(s_2\) is a Slepian projection smoother. Note that by using \(W = 7 \text{ cycles/year}\) we are using more degrees-of-freedom (basis vectors) in the \texttt{spsmooth} case than the cubic regression smoother case. For the same assumed cut-off frequency \(f_0\) in cycles/year and same number of years \(N_y\), a cubic regression basis will use \(f_0N_y + 1\), while a Slepian projection smoother will use approximately
$2N_yf_0 - 1$. For our test case of Chicago and the NMMAPS database, this is the difference between $7 \cdot 14 + 1 = 99$ $dof$ and $2 \cdot 14 \cdot 7 - 1 = 195$ $dof$. This does point out an interesting problem: the drastic difference in $dof$ seems to make a direct comparison of these two methods invalid. The point we are trying to make is that the direct comparison is being done from a philosophical perspective: how many degrees-of-freedom (infer: under a specific smoother formulation) are necessary to protect our estimate from unmeasured confounding? In the case of the cubic regression smoother, the effective transfer function is so far from optimal that the “saved” degrees-of-freedom are meaningless in the face of unwarranted inference.

Figures 5.8 and 5.9 compare the theoretical transfer function of the `spsmooth` filter with 195 $dof$ (for $N = 14$ years and $W = 7$ cycles/yr) to the cubic regression
filter with 99 \textit{dof}.

\begin{table}
\centering
\begin{tabular}{|l|cc|cc|}
\hline
Parameter & Interpolated & & & Not Interpolated \\
 & Coefficient & Std. Err. & & Coefficient & Std. Err. \\
\hline
Original \textit{gam} (\textit{mgcv}) & 3.968 & 0.893 & & 3.634 & 0.952 \\
Linear Model - (5.16) & 3.968 & 0.719 & & 3.634 & 0.752 \\
Linear Model - (5.23) & 3.432 & 0.699 & & 3.141 & 0.731 \\
\hline
Updated \textit{gam} using \textit{spsmooth} & 3.115 & 0.899 & & 3.111 & 0.960 \\
Linear Model - (5.16) & 3.115 & 0.698 & & 3.111 & 0.729 \\
Linear Model - (5.23) & 2.678 & 0.698 & & 2.680 & 0.729 \\
\hline
\end{tabular}
\caption{Examination of coefficients for Chicago using Eqn. 5.14, Eqn. 5.16 and Eqn. 5.23 and contrasting with the same applied using Eqn. 5.28. All coefficients and standard errors \textit{Z} should be read as \textit{Z} \times 10^{-4}, i.e. they have all been multiplied by 10^4. The change between the 5\textsuperscript{th} and 6\textsuperscript{th} rows indicates that there is still something missing.}
\end{table}

\begin{table}
\centering
\begin{tabular}{|l|c|c|}
\hline
Equation & Interpolated & Not Interpolated \\
\hline
Linear Model - (5.16) & 6.21 \times 10^{-2}\% & 5.40 \times 10^{-2}\% \\
Linear Model - (5.23) & 4.52 \times 10^{-7}\% & 3.83 \times 10^{-4}\% \\
\hline
\end{tabular}
\caption{Estimation of the percentage of power (variance) below the frequency cut-off of 7 cycles/year using Eqn. 5.28 and applying Eqns. 5.16 and 5.23. This table should be compared to Table 5.3 to see that introducing the \textit{spsmooth} formulation lowered the base power remaining from \approx 5\% to 0.062\% - despite this, applying the manual filter lowers the power still further and \textit{also} changes the coefficient, see Table 5.4.}
\end{table}

We now reproduce the model results of Table 5.2 with \textit{spsmooth} replacing the cubic regression smoother in Table 5.4. These results show that although the inclusion of the \textit{spsmooth} smoother changes the coefficient, the linear model formulation of Eqn. 5.23 which removes the remaining low-frequency power still has a significant influence on the coefficient. Examination of Table 5.5 shows that the residual power remaining below the cut-off of 7 cycles/year is quite small: less than 0.1\%. Despite this, there is a large amount of change in the coefficient of risk due to air pollution.
This was originally puzzling, but the thought experiment of Section 5.4.2 provided us with the solution. The updated \texttt{spsmooth} projection filter/smoother may be doing a much better job than the original cubic regression spline smoother, but any parametric terms (e.g. the pollutant) can still introduce low-frequency structure into the residual effective mortality that the smoother can never access.

![Power Spectrum](image)

**Figure 5.10:** Demonstration of multitaper power spectra for Eqn. 5.28. The black curve is the spectrum of the log mortality, i.e. the original response for the GAM. The red curve shows the spectrum of the residual effective mortality that is left after fitting the smooth function of time; this residual is then regressed on the temperature, pollutant, and day-of-week terms. \textbf{Note}: the periodicity at 2 cycles/year is entirely spurious, and is due to the nonlinear filter applied to the temperature series!

The change in the coefficient despite the minimal amount of low-frequency power remaining is due primarily to the persistence of the 1 and 2 cycles/year periodicity. Figure 5.10 shows the residual effective mortality from Eqn. 5.28 as jointly seen by the
temperature, pollutant, and day-of-week term. Note that although the residual power between 3 and 7 cycles/year has been attenuated by several orders of magnitude, the periodicities at 1 and 2 cycles/year persist. This returns us to the discussion at the end of Section 5.4.2 where we stepped through a model of the P-IRLS algorithm. The conclusion from this discussion was that due to the parametric inclusion of the pollutant in the model formulation, the smooth function of time will not be able to prevent the residual effective mortality from having these periodicities.

Taking inspiration from the parametric form, our final suggestion for fixing this model formulation is to pre-filter the pollutant. If we can remove the low-frequency periodicities from the pollutant before the algorithm initializes, then this will prevent the “masked” low-frequency power from modifying the smooth function of time, and the eventual fit should be that much more precise. To pre-filter the pollutant, we use the same filter formulation we have been using all along: a Slepian high-pass projection filter with frequency cutoff 7 cycles/year. Again, the high-pass pollutant is written as

\[
\text{pm10tmean}_{HP} = \text{pm10tmean} - \mathcal{H}_S \cdot \text{pm10tmean},
\]

where \(\mathcal{H}_S\) is the Slepian projection filter operator, and defined in Eqn. 5.21. The GAM formulation is then identical to Eqn. 5.28 except for the change in pollutant term:

\[
\log(\text{death}) = \text{pm10tmean}_{HP} + s(\text{tmpd}, df = 6) + \text{s}_2(\text{time}, W = 7/\text{yr}) + \text{as.factor(DOW)}
\]

We conclude with repetition of the two tables and figure first shown for Eqn. 5.28 and now shown for Eqn. 5.30.

The results of Tables 5.6 and 5.7 are excellent. They show that the residual power below 7 cycles/year is quite low: while there is some difference between the model using \text{spsmooth} and high-pass filtering the residual effective mortality, it is the
Table 5.6: Examination of coefficients for Chicago using Eqn. 5.14, Eqn. 5.16 and Eqn. 5.23 and contrasting with the same applied using Eqn. 5.30. All coefficients and standard errors $Z$ should be read as $Z \times 10^{-4}$, i.e. they have all been multiplied by $10^4$. The change between the 5th and 6th rows (or lack thereof) indicates success, and should be interpreted with the aid of the blue curve in Figure 5.11.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Interpolated</th>
<th>Not Interpolated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Std. Err.</td>
</tr>
<tr>
<td><strong>Original $\text{gam}$</strong></td>
<td>3.968</td>
<td>0.893</td>
</tr>
<tr>
<td>Linear Model - (5.16)</td>
<td>3.968</td>
<td>0.719</td>
</tr>
<tr>
<td>Linear Model - (5.23)</td>
<td>3.432</td>
<td>0.699</td>
</tr>
<tr>
<td><strong>Updated $\text{gam}$ using $\text{spsmooth}$</strong></td>
<td>3.115</td>
<td>0.899</td>
</tr>
<tr>
<td>Linear Model - (5.16)</td>
<td>3.115</td>
<td>0.754</td>
</tr>
<tr>
<td>Linear Model - (5.23)</td>
<td>3.114</td>
<td>0.754</td>
</tr>
</tbody>
</table>

Table 5.7: Estimation of the percentage of power (variance) below the frequency cut-off of 7 cycles/year using Eqns. 5.16 and 5.23.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Interpolated</th>
<th>Not Interpolated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model - (5.16)</td>
<td>$9.19 \times 10^{-3}%$</td>
<td>$6.07 \times 10^{-3}%$</td>
</tr>
<tr>
<td>Linear Model - (5.23)</td>
<td>$4.48 \times 10^{-7}%$</td>
<td>$9.34 \times 10^{-5}%$</td>
</tr>
</tbody>
</table>

difference between $10^{-3}$ and $10^{-7}\%$: insignificant percentages. This is confirmed by Table 5.6, which shows that the coefficients obtained from the $\text{spsmooth}$ formulation are almost identical to those obtained from the manually high-pass filtered copy. In Figure 5.11, we examine the residual effective mortality power spectrum.

The spectra exhibited in Figure 5.11 are illustrative. The red spectrum is very similar to that shown in Figure 5.10. The periodicities at 1 and 2 cycles/year are still present, and are effectively as strong as before. Trying to reconcile their presence with the minimal changes (rows 5 and 6) in the coefficients of Table 5.6 and the residual power in Table 5.7 forces us to examine the remaining terms. Recall that the red curve is the residual effective mortality that is regressed on the temperature, high-pass pollutant, and day-of-week terms. Given that the pollutant has no low-frequency
Figure 5.11: Demonstration of the multitaper power spectrum relating to Eqn. 5.30. The black curve is the spectrum of the log mortality used as the original response in the GAM. The red curve shows the spectrum of the residual effective mortality that is regressed on the temperature, pollutant, and day-of-week terms in the model. The blue curve shows the spectrum of the residual effective mortality that is regressed only on the pollutant and day-of-week terms; notice the strong difference at 1 and 2 cycles/year. The difference indicates that the bulk of the problematic power is due to the temperature term. All spectra are presented with the mean left intact, which produces the rectangular shapes at the left of the frequency domain.

power (due to our prefiltering), it is difficult to see how it could have any influence on the dynamics of the P-IRLS algorithm described at the end of Section 5.4.2. This leads us to the smooth function of temperature. The blue curve presented in Figure 5.11 is the residual effective mortality that is regressed on the high-passed pollutant and day-of-week terms only: note how the power drops by another order of magnitude across the 0.5–7 cycles/year band, and the periodicities are almost completely removed. This
leads us to the final development of this chapter.

Figure 5.12: Demonstration of multitaper power spectrum of daily temperature from Chicago, 1987–2000. The black curve demonstrates the (scaled) log spectrum of observed temperature. The red curve shows the smoothed (prediction) series as obtained from the model of Eqn. 5.30. Note that the smooth has introduced a strong (completely spurious!) 2 cycle/year periodicity that is not in the original series and which is caused by the nonlinear smooth of temperature on residual mortality.

Temperature is an environmental covariate with strong influences from solar effects, in particular, the yearly orbit. As such, the time series of temperature contains strong yearly and bi-yearly periodicities, as seen in Figure 5.12. The red curve shows the smooth of temperature against log mortality while accounting for a smooth function of time and both high-pass pollutant and day-of-week. This figure shows that the majority of the periodic structure observed in Figure 5.11 is actually due to temperature, as shown by the blue curve. Thankfully the blue curve also indicates that most
of the spurious power that is introduced through the nonlinear smooth of temperature on residual mortality is self-contained in the temperature term, so that it does not excessively effect the risk estimate (coefficient of the pollutant). We will examine the smooth of temperature further in the next chapter: it appears to be a necessary evil for now, although a future project is to revisit the entire issue and apply time series techniques to the mortality-temperature relationship and try to avoid the entire nonlinearity issue.

5.6 Summary

In this chapter we have reviewed the philosophy and literature behind estimation of population health risk due to air pollution. We have further explored the use of cubic regression splines as smoothers in Generalized Additive Models, and demonstrated problems with these smoothers. By appealing to linear filter design theory, we formulated a replacement “sp” smoother, implemented an extension package for \texttt{mgcv} which allows for use of this smoother in practice, re-estimated the models, and determined one further possible source of confounding. Pre-filtering the pollutant allowed for removal of any additional confounding possibility due to long timescale variation, and further showed that the coefficient obtained from the \texttt{spsmooth} model without the pre-filtering is actually very close to that obtained by the pre-filtered version. Finally, we examined the role that temperature plays in the model process, and showed that it is responsible for the 1 and 2 cycles/year structure that is found in the residual effective mortality as regressed jointly on the temperature, the pollutant, and the day-of-week. As part of this structure is entirely spurious we will revisit the issue of temperature in the following chapter.
Chapter 6

Population Health Risk due to Air Pollution

The developments in Chapter 5 were specifically aimed at the problem of smoothing mortality with respect to time. In this chapter we will take a broader view of the problem of measuring national risk due to air pollution. Some of the work presented in this chapter has previously been presented to Health Canada in a series of technical reports [188, 28, 29, 187, 189, 190, 182, 184, 185, 186] and at a number of scientific meetings [32, 30, 33, 183].

The material of Chapter 3 is used in the creation of the database for the AHI. Chapter 4 is not used for this database, but was used (as demonstrated in that chapter) for gapfilling the daily noon solar flux series, which was intended as a possible covariate in the level 1 models, although this unfortunately was not completed for this thesis. The results of Chapter 5 come into play in Section 6.3 and the following discussion.
6.1 The Air Health Indicator Project

Health Canada has recently developed a new methodology, the Air Health Indicator, for assessing the effects on daily mortality of short-term exposure air pollution as they may vary dynamically over space and time in response to changes in air quality. This methodology has been assessed in simulation studies and by applying it to 20 years of data from major Canadian cities. Hundreds of time-series studies of daily mortality have been published worldwide (some of which are reviewed in Sections 5.1 and 5.2), and are critical components of the scientific evidence supporting a causal relationship between air pollution and public health. The AHI provides time trends in annual risks at city, regional and national levels; the hope is that these trends will demonstrate a trend over the available time period. The AHI can be used in policy analysis, with potentially important applications to the assessment of the public health impacts of air quality regulation.

For the AHI, health outcome (daily mortality) data from 24 Canadian cities is used and estimation of between-city heterogeneity has been problematic due to the small number of cities involved. Applying least squares estimation (LSE) occasionally results in negative estimates that are meaningless from a scientific viewpoint. To improve the estimate, a Bayesian Markov Chain Monte Carlo (MCMC) [63] approach has been employed to estimate the heterogeneity as well as the national risks. In a previous contract, Dr. Glen Takahara [181] investigated the effects of prior distributions on hyperparameters of the risk estimation and found the use of non-informative prior distributions to be desirable. Both research scientists at Health Canada and Dr. Takahara have also experimented with incorporating demographic variables such as population and population density into the national risk estimation and have found that a strong cyclic pattern is apparent in the national risk estimates, especially when data has been restricted to mortalities occurring during warm weather months and
related to cardio-respiratory causes. Such a pattern, if due to environmental or other explainable causes, will be important for its scientific and policy implications. The causes of this pattern are not understood – this is an important problem both for its own sake and to determine how it influences the AHI.

Estimation done prior to 2011 assumed that the annual risks were independent. In order to relax this assumption, we have worked toward a method allowing for correlated data. We have applied time series spectral methods to analyze the frequency-domain structure of the mortality and environmental covariates. Previous models used either NO$_2$ or Ozone as primary air pollutant, but there is strong interest in using the Canadian PM$_{2.5}^1$ data as well. Unfortunately, the PM$_{2.5}$ data is not fully available and contains many gaps. Development of the interpolation techniques in Chapters 3 and 4 was additionally undertaken to aid in reconstruction of PM$_{2.5}$ data, although this remains a work-in-progress.

### 6.2 Database

Data is integral to any successful modeling endeavour. Unfortunately, since we were the first to work with many of the data sets in question, there were a number of issues caused by having to clean, organize, sort, and analyze the data before we could begin. The raw “database” consisted of hundreds of Excel spreadsheets containing hourly or daily measurements from 24 Canadian urban centers; the measurements were of Ozone, NO$_2$, SO$_2$, Temperature, Humidity and Mortality. The environmental covariates were all recorded hourly, and the mortality recorded as counts, each count corresponding to a single human death.

In a process that is too long to go into here, but which can be reviewed in [28], these spreadsheets were amalgamated into large data structures and aggregated to a

---

$^1$Similar to the PM$_{10}$ data from the NMMAPS database, used in Chapter 5.
daily measurement scale. The metrics used for combining the discrete measurement
stations and instruments into one single “daily” measurement were quite complicated,
and are fully detailed in [28] as well. After the station aggregation, the data sets for
Ozone and NO$_2$ were interpolated using the methods of Chapter 3. Finally, these sets
were combined in R to form a single AHIdat database – it is at this point that various
lags and categorizations were applied to form different variables.

6.2.1 Particulate Matter (PM$_{2.5}$)

We had hoped to interpolate the Particulate Matter (PM$_{2.5}$) data as well, to provide
an additional pollutant for analysis. Unfortunately, the scope of the interpolation was
too broad to be fully completed for this thesis. There are two observational PM$_{2.5}$
data sets available, the dichot series which was recorded by a dichotomous air flow
receiver tube with a fine particle filter [60], and the Tapered Electrode Oscillating
Microbalance (TEOM) series which was recorded by a number of monitors that use
direct intertial micro-balance measurements of particulate matter from the air [152].
Environment Canada is most interested in results from the TEOMs data set rather
than the dichot data. As such, we have partially interpolated the TEOMs data that
is available for the 1998-2009 time period. Unfortunately, this series does not cover
the full time span (1981–2007), and as such is less valuable for the results of the rest
of this chapter.

It may be possible to impute further data values for the 1980-2000 period using the
dichot data set, which is measured only once every 6 days and contains further gaps.
However, this imputation is beyond the scope of this thesis and the technique (so far)
remains only a notion. As far as we know, it is near-impossible to actually reconstruct
a reasonable daily series from one sample per six days, although we feel confident that
the data values can at least be doubled so as to provide one measurement every three

141
days using spectral methods and leveraging the sampling theorem.

6.3 Initial Model Formulation

The core model of the AHI is designed to estimate geographically localized population risk due to air pollution. The time resolution is set to daily scale to both allow for sufficient deaths to occur in the time span so as to avoid excessive numbers of zeros and also to correspond to the typical reporting interval for environmental data. The spatial resolution is restricted to subsets of Canadian urban city limits, defined with the help of Environment Canada based on location of pollutant sensors so as to ensure that the deaths observed and the pollution levels observed are at least plausibly related. The AHI uses GAMs in the style of [52] (and many others), restricted specifically to year-at-a-time, and occasionally limiting time resolution to single seasons. Beginning with the GAM used in [52], the log of the mean response is modeled as the additive sum of smooth functions of some number of predictors and the additional inclusion of parametric terms for a pollutant of interest and the day-of-week (as a factor variable). Formally:

\[
\log(\mu_{at}^c) = x_{at}^c \beta^c + \gamma DOW + S_{1}^c(\text{time}, 7/year) + S_{2}^c(\text{temp}_0, 6) \\
+ S_{3}^c(\text{temp}_{1-3}, 6) + S_{4}^c(\text{dew}_0, 3) + S_{5}^c(\text{dew}_{1-3}, 3) \\
+ \text{intercept for age group } a \\
+ \text{separate smooth functions of time (8 dof) for age group } a.
\] (6.1)

For the AHI model, we remove the age-specific intercept and smooth functions of time and additionally remove the dewpoint and focus on a single pollutant covariate – for the purposes of this thesis, we will focus on NO$_2$. Initially, the pollutant used was the 3-day moving average of NO$_2$, that is, the average of the current, lag-1, and lag-2 days’ measurements. We further include two smooth functions of temperature: one at lag 0 and one at the average of lags 1, 2 and 3. These two temperature
covariates are included because model fitting work performed by the researchers at Health Canada found that a two-temperature model with these choices lowered the overall AIC model-fitting criterion. Thus, as a simplified model, we begin with

\[
\log(\mu_t) = \beta (\text{NO}_2)_{1-3} + S_1(\text{temp}_0, df = 3) + S_2(\text{temp}_{1-3}, df = 3) + S_3(\text{time}, df = 7/\text{year}) + \gamma \text{DOW},
\]

(6.2)

where \((\text{NO}_2)_{1-3}\) is the 3-day moving average measure of NO\(_2\), DOW is a factor variable for day-of-week, and the degrees of freedom for the three smoothers were chosen on the basis of epidemiologic understanding of the relationships. We will revisit these degrees of freedom in the following sections, and the reader is encouraged to recall the discussion from Chapter 5 as well.

### 6.4 Pollutant Choices

For the univariate pollution model, we use only NO\(_2\) as our choice of covariate. There are obviously any number of choices which are possible for pollutant: Ozone, NO\(_2\), SO\(_2\), and (most importantly) PM\(_{10}\) and PM\(_{2.5}\), among others. As discussed in Section 6.2.1, we do not have access to significant PM databases for our Canadian cities, and must regretfully exclude them from the discussion. SO\(_2\) is also quite limited in its data availability, so our focus for the remainder of the thesis will be on NO\(_2\) for the rest of the univariate model building discussion. Ozone is interesting for the multivariate case, which unfortunately is beyond the scope of this thesis, but is an obvious extension to this work. Note that similar effects to the NO\(_2\) choice are also observed in the Ozone case, so there is nothing lost by restricting ourselves in this way.

As mentioned in the previous section, the original choice of lag for the NO\(_2\) pollutant was a 3-day moving average of the zeroth, first and second order lags. This
Figure 6.1: Demonstration of the multitaper spectrum of a 3-day moving average of 
NO$_2$ from 1981 to 2007, taken from CD 3520, Toronto. The light grey 
spectrum in the background is the original power spectrum of NO$_2$, the 
black curve is the 3-day moving average, and the red curve overlaid on 
top is the theoretical magnitude response transfer function of the linear 
filter represented by the moving average. The blue vertical lines indicate 
periods of 1, 2 and 3 cycles/week that are clearly visible against the 
background.

was seen as a compromise between the individual lags that seemed to allow for amal-
gamation of the information contained in them. However, when viewed from a time 
series perspective, a 3–day moving average is simply a crude linear filter, and thus 
may not be a sensible choice. Consider Figure 6.1. In this figure we show the original 
log multitaper power spectrum (in grey) with the 3–day moving average plotted in 
black over top. Note the extremely strong zero response at a period of 3 days – as 
is obvious from examining the red magnitude response transfer function also on the 
plot, this is due to the filtering operation. When we recall that the risk estimate we
seek is intended to be high-frequency (i.e., of short timescale), distorting the very structure we seek to model seems counterproductive.

If we eliminate multiple-day averages due to this argument, we are left with singular lags of the pollutant – the zeroth lag, first lag, and so on. From a causal point of view, it also seems unreasonable to expect air pollution from many days in the past to have any influence on mortality today [170, 235]. For these reasons, and also for ease of analysis, we restrict ourselves for the remainder of this work to the 0, 1 and 2 lags only. There is a possible argument for the third-order and higher lags being of importance as well, but space is limited.

6.4.1 Choosing the Lag

Begin with the model of Eqn. 6.2 and consider a data set consisting of Census Divisions (CD) 3520 (Toronto), 5915 (Vancouver), 2466 (Montreal), 3521 (Peel), 4806 (Edmonton) and 4811 (Calgary). These are the census divisions with a population of at least 1 million as of 2007, which is the last year of data available. These 6 CDs provide enough data to allow us to examine the lag choices, and as they are distributed nationally, should generalize reasonably to the country as a whole.

We apply the model (6.2) to 27 years\(^2\) of All-Cause and Cardio-Pulmonary mortality data, 1981–2007, for each of the 6 cities and obtain estimates of risk with error bounds. For the smooth functions of temperature we use cubic regression splines, and for the smooth function of time we use the \texttt{spsmooth} package and a Slepian projection smoother basis, as detailed in Chapter 5. A quasi-likelihood family is used to allow for overdispersion. The results for lags 0, 1 and 2 are presented in Table 6.1.

As this table shows, for the All-Cause mortality, 5 of the 6 cities have the strongest

\(^2\)This is true for 5 of the 6 cities: Montreal has data available only up to 2001 (21 years) due to the mortality data having not been released by the Institut de la Statistique Québec.
Table 6.1: Coefficients and Standard Errors for the model of Eqn. 6.2. Computed on three lags of NO$_2$ (0, 1 and 2) and for 6 Census Divisions, using the Slepian projection smoother for time and a quasi-likelihood link function. Both All-Cause (AC) and Cardio-Pulmonary (CP) models are computed.

<table>
<thead>
<tr>
<th>CD</th>
<th>Lag-0 AC</th>
<th>Coef</th>
<th>StdE</th>
<th>Lag-1 AC</th>
<th>Coef</th>
<th>StdE</th>
<th>Lag-2 AC</th>
<th>Coef</th>
<th>StdE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3520</td>
<td>-7.67e-8</td>
<td>2.40e-4</td>
<td></td>
<td>6.26e-4</td>
<td>2.44e-4</td>
<td></td>
<td>5.47e-4</td>
<td>2.31e-4</td>
<td></td>
</tr>
<tr>
<td>5915</td>
<td>3.19e-4</td>
<td>4.56e-4</td>
<td></td>
<td>1.50e-3</td>
<td>4.50e-4</td>
<td></td>
<td>7.39e-4</td>
<td>4.53e-4</td>
<td></td>
</tr>
<tr>
<td>3521</td>
<td>-5.57e-4</td>
<td>6.00e-4</td>
<td></td>
<td>4.65e-4</td>
<td>6.26e-4</td>
<td></td>
<td>6.26e-5</td>
<td>5.90e-4</td>
<td></td>
</tr>
<tr>
<td>2466</td>
<td>-5.34e-4</td>
<td>3.23e-4</td>
<td></td>
<td>3.56e-4</td>
<td>3.32e-4</td>
<td></td>
<td>5.77e-4</td>
<td>3.13e-4</td>
<td></td>
</tr>
<tr>
<td>4811</td>
<td>-4.43e-4</td>
<td>5.10e-4</td>
<td></td>
<td>1.22e-4</td>
<td>5.27e-4</td>
<td></td>
<td>3.84e-4</td>
<td>5.13e-4</td>
<td></td>
</tr>
<tr>
<td>4806</td>
<td>9.09e-4</td>
<td>5.45e-4</td>
<td></td>
<td>5.82e-4</td>
<td>5.66e-4</td>
<td></td>
<td>-5.84e-5</td>
<td>5.46e-4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD</th>
<th>Lag-0 CP</th>
<th>Coef</th>
<th>StdE</th>
<th>Lag-1 CP</th>
<th>Coef</th>
<th>StdE</th>
<th>Lag-2 CP</th>
<th>Coef</th>
<th>StdE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3520</td>
<td>-7.67e-8</td>
<td>2.40e-4</td>
<td></td>
<td>5.40e-4</td>
<td>3.65e-4</td>
<td></td>
<td>9.96e-4</td>
<td>3.46e-4</td>
<td></td>
</tr>
<tr>
<td>5915</td>
<td>9.48e-4</td>
<td>6.29e-4</td>
<td></td>
<td>1.70e-3</td>
<td>6.21e-4</td>
<td></td>
<td>6.55e-4</td>
<td>6.26e-4</td>
<td></td>
</tr>
<tr>
<td>3521</td>
<td>-1.56e-3</td>
<td>9.37e-4</td>
<td></td>
<td>3.05e-4</td>
<td>9.74e-4</td>
<td></td>
<td>8.24e-5</td>
<td>9.17e-4</td>
<td></td>
</tr>
<tr>
<td>2466</td>
<td>-6.23e-4</td>
<td>4.49e-4</td>
<td></td>
<td>2.54e-4</td>
<td>4.62e-4</td>
<td></td>
<td>4.94e-4</td>
<td>4.35e-4</td>
<td></td>
</tr>
<tr>
<td>4811</td>
<td>-4.02e-4</td>
<td>7.19e-4</td>
<td></td>
<td>-4.46e-4</td>
<td>7.41e-4</td>
<td></td>
<td>8.33e-4</td>
<td>7.21e-4</td>
<td></td>
</tr>
<tr>
<td>4806</td>
<td>7.33e-4</td>
<td>7.76e-4</td>
<td></td>
<td>6.71e-4</td>
<td>8.06e-4</td>
<td></td>
<td>-7.69e-4</td>
<td>7.78e-4</td>
<td></td>
</tr>
</tbody>
</table>

relation at either lag-1 or lag-2 days, rather than lag-0. The same is true for Cardio-Pulmonary mortality, again for all cities excepting Calgary, which is slightly stronger at lag-0. Note here that negative results are disregarded – there are currently no indications that air pollution can prevent human mortality, so these coefficients are taken simply for their statistical meaning, which is to say, anti-correlation. Further note that these models were computed using NO$_2$ which has not been prefiltered, as was suggested in Chapter 5. We examine this point in the next section.

### 6.4.2 Prefiltering

In Chapter 5, Eqn. 5.30 and the discussion surrounding it explained why we might want to pre-filter the pollutant of interest. Implementing this for the three lags of
NO\textsubscript{2} considered, we run a similar model to Eqn. 6.2, changing only the pollutant to its high-pass filtered version:

\[
\log(\mu_t) = \beta (\text{NO}_2)_{\text{HP},l} + S_1(\text{temp}_0, df = 3) + S_2(\text{temp}_{1-3}, df = 3) + S_3(\text{time}, df = 7/\text{year}) + \gamma \text{DOW},
\]

for \(l \in \{0, 1, 2\}\). The risk estimates for the pollutant at lags 0, 1 and 2 are shown in Table 6.2 below, and tell the same story as Table 6.1 – the first and second order lags are more related to the mortality than the zeroth order lag.

<table>
<thead>
<tr>
<th>CD</th>
<th>Lag-0 AC Coef</th>
<th>StdE</th>
<th>Lag-0 CP Coef</th>
<th>StdE</th>
<th>Lag-1 AC Coef</th>
<th>StdE</th>
<th>Lag-1 CP Coef</th>
<th>StdE</th>
<th>Lag-2 AC Coef</th>
<th>StdE</th>
<th>Lag-2 CP Coef</th>
<th>StdE</th>
</tr>
</thead>
<tbody>
<tr>
<td>3520</td>
<td>3.30e-07</td>
<td>2.40e-04</td>
<td>6.27e-04</td>
<td>2.44e-04</td>
<td>5.48e-04</td>
<td>2.31e-04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3521</td>
<td>-4.95e-04</td>
<td>6.00e-04</td>
<td>5.28e-04</td>
<td>6.26e-04</td>
<td>1.15e-04</td>
<td>5.90e-04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4806</td>
<td>9.26e-04</td>
<td>5.46e-04</td>
<td>5.96e-04</td>
<td>5.66e-04</td>
<td>-5.03e-05</td>
<td>5.46e-04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD</th>
<th>Lag-0 CP Coef</th>
<th>StdE</th>
<th>Lag-0 CP Coef</th>
<th>StdE</th>
<th>Lag-1 CP Coef</th>
<th>StdE</th>
<th>Lag-1 CP Coef</th>
<th>StdE</th>
<th>Lag-2 CP Coef</th>
<th>StdE</th>
</tr>
</thead>
</table>

Table 6.2: Coefficients and Standard Errors for the model of Eqn. 6.3 using the pre-filtered pollutant. Computed on three lags of NO\textsubscript{2} (0, 1 and 2) and for 6 Census Divisions, using the Slepian projection smoother for time and a quasi-likelihood link function. Both All-Cause (AC) and Cardio-Pulmonary (CP) models are computed.

For the remainder of this work we will use the lag-1 copy as a simplification step – additional work has shown that the second order lag has similar properties to much of what follows, so we felt this simplification was a reasonable choice.
6.5 Weekly Periodicities and the Day-of-Week Term

It is common for air pollution GAMs to include a term for the day-of-week. In the implementation of the model solvers, this term essentially maps to a series of means on offset lag-7 sequences, that is

\[(\text{DOW})_1 = \frac{1}{N/7} \sum_{j=1,8,15,\ldots} x_j\]
\[(\text{DOW})_2 = \frac{1}{N/7} \sum_{j=2,9,16,\ldots} x_j\]
\[
\quad \vdots
\]
\[(\text{DOW})_6 = \frac{1}{N/7} \sum_{j=6,13,20,\ldots} x_j\]
\[(\text{DOW})_7 = -\sum_{k=1}^{6} (\text{DOW})_k\]

with the additional constraint that \((\text{DOW})_1 + (\text{DOW})_2 + \cdots + (\text{DOW})_7 = 0\) due to the centering constraint of the GAM solver. The final (7th) day-of-week factor variable is used to enforce the zero-sum constraint. Thus, the implementation of the day-of-week essentially acts as a filter on the mortality, damping the effects of periodicities at 1,2 and 3 cycles/week. It would go further with respect to harmonics of 1 cycle/week, but the Nyquist frequency for our data is 5.787µHz, or a 2 day period; 4 cycles/week is 1.75 days, and thus outside the bound. However, this 4 cycles/week periodicity does alias back through the Nyquist bound – it actually lands exactly on top of the 3 cycles/week periodicity. This helps explain the somewhat curious behaviour of the 3 cycles/week periodicity, which has erratic behaviour (in comparison to the 1 and 2 cycles/week periodicities).

Periodicities that cycle on the order of the week or harmonics of it are clearly human in origin. There is nothing special in the universe about 7 rotations of our world about its axis – certainly 1 rotation is important, and shows up as daily harmonics in almost every long-run observational series in existence, but 7 of these rotations put
together has no particular importance to our world. We do not see these periodicities in temperature data or other large-scale meteorological series, but we *do* see them in ground-level ozone, NO\textsubscript{2} and other nitrate series, and human mortality. The tie between the week and ground-level air pollution is clear: the work-week and weekend interact to produce traffic patterns which induce patterns on the air pollution series, and many of these patterns cycle with period 1 week. Similarly, mortality series track the week due to a number of causes; see [55, 170, 106] as randomly chosen examples from a wealth of literature showing these effects.

![Figure 6.2: Possible causal links between pollution, mortality, and the day-of-week. The directional arrows indicate causality.](image)

The inclusion of the DOW term in a GAM effectively masks the weekly periodicity and its harmonics from influencing the estimation of risk due to air pollution. The question we ask is whether this is too extreme a stance to take. The causal links between the day-of-week and pollution seem clear, and *some* of the causal links
between the day-of-week and mortality also are clear, but we do not know if there are cross-relations or not. A future work is the examination of partial coherences in an attempt to clear up this lack of understanding. The possible paths are summarized in Figure 6.2. We know from atmospheric chemistry work [43] that NO$_x$\(^3\) in the atmosphere contributes to the generation of Ozone, which is the reason for the causal link between NO$_2$ and Ozone. We further know that NO$_2$ and Ozone have *some* causal link to human mortality through the wealth of research touched on in the review of Section 5.1. Finally, we know that the human calendar week has causal links to both the production of air pollution and human mortality. The question is: does the week’s influence on NO$_2$ and Ozone carry through to their further influence on mortality, or does it die there, and any influence the pollutants have on mortality is independent of the week?

Thus, there are two possible extreme positions. The first is that the day-of-week influences mortality and also influences air pollution, but does not influence mortality through air pollution. This is the stance taken by most common models, and the use of the DOW term masks the day-of-week periodicities so they do not influence the estimation of risk. The second position is that the day-of-week does not influence mortality at all except through air pollution, i.e., that the lowest arrow in Figure 6.2 does not exist. We feel the truth lies between these two extremes, and thus we must consider how to attack the problem of estimation of risk in the presence of weekly periodicities.

### 6.5.1 Line Component Removal

As mentioned in Chapter 3 in the development of the $T_t$ estimator, it is possible to remove line components from a time series using a multitaper spectrum estimate and

\[^3\text{Nitric oxide, NO, and Nitrogen dioxide, NO}_2.\]
an inverse FFT. This technique allows us to examine the components of the pollutant and mortality series and determine their precise characteristics without requiring high levels of understanding for frequency-domain time-series techniques.

When the weekly harmonics of the pollutant and mortality series are separated out, they are typically not in phase, i.e., there is a lag offset in these components. For a tutorial examination of phase, the reader should refer to Appendix D. For prediction of a periodic component in the mortality series from the corresponding periodic component in the pollutant series (at a given frequency) we should time shift one of the components (most logically the pollutant) so that they are lined up. This time shifting is a way to use lags of individual components of the pollutant, but is a more focused and specific way of taking lags than simply taking a lag of the entire pollutant series (which would still leave at least some of the periodic components of the pollutant out of phase with the corresponding periodic components of the mortality).

We note here that significant research has been completed on on distributed-lag models. For an overview of the early history, see [2, 70]. With specific application to the estimation of risk due to air pollution, the more recent work of [233, 62, 66] is more interesting, especially the developments of Gasparrini et al. [62]. Unfortunately the scope of this thesis does not allow for examination of our newly developed techniques in the context of distributed lag or distributed lag non-linear models (dlnlms), but we would be amiss to not mention them briefly. We intend this to be a significant area of future work, and will apply our new smoother methodology to these models to confirm (or modify, if need be) the previously determined understanding of the response relationships between causal factors and health effects. Note further that dlnlms are mostly applied to temperature and mortality, as briefly discussed in the previous chapter and again taken up in Section 6.6.1.
6.6 Coherence and Time Delay Estimation

The tutorial of Appendix D demonstrates two toy examples and also shows the behaviour of the coarse lag-at-a-time GAM structure, which unfortunately is a reality for these models. We were able to explicitly model the structure of the line components and thus easily obtain phase offset estimates from them without any tools more complicated than a multitaper spectrum estimate. In this section, we will briefly examine more powerful signal processing tools that give us additional insight into the problem.

In the analysis of multivariate time series, we often estimate the individual power spectra (or *autospectra*) for the individual series, say $S_{xx}(f)$ and $S_{yy}(f)$. We can additionally estimate the cross-spectra $S_{xy}(f)$ in the obvious fashion. All of these estimates will be computed using the multitaper spectral estimator. Following the lead of Thomson [192] (who further was following Tukey), define coherency as

$$\tilde{\gamma}(f) = \frac{\hat{S}_{xy}(f)}{(S_{xx}(f)S_{yy}(f)))^{1/2}}$$

with coherence the magnitude of coherency, $\gamma = |\tilde{\gamma}(f)|$. Further define the magnitude-squared coherence (MSC) as $\gamma^2$.

Coherency, being a complex variate, is characterized by both magnitude (coherence) and phase, with phase often being more sensitive than MSC as a parameter. Phase is circular (being defined on $(-\pi, \pi]$, $(0, 2\pi]$, or a variation thereof) and computed through use of the four quadrant complex arctangent. The hardest part of phase estimation (when done with respect to frequency) can be determining when the phase series “wraps” around the $\pm \pi$ boundaries and unwrapping the series correctly. The examples that follow will demonstrate that this is a non-trivial task.

It is possible to jackknife both the coherence (or MSC) estimate and the phase estimate from this formulation by using delete-one-jackknifing. The details of this will not be given here, and we direct the reader to [192, 197, 200, 148] for full derivations.
and additional distributional assumptions. All plots in the remainder of this section were produced in R using the \texttt{multitaper} package with jackknifing enabled and using $NW = 5.0$ and $K = 8$. The jackknife is applied to obtain $\pm 1\sigma$ bounds on the estimates, and will be presented in the plots that follow.

### 6.6.1 Mortality and Temperature

Figure 6.3: Magnitude-Squared Coherence between All-Cause Mortality and Temperature for Toronto for the full 1981–2007 time span. The red vertical dashed line is at 1 cycle/year. Both series have had arithmetic means removed from them. The light dashed line is the $-1\sigma$ confidence bound, found by jackknifing.

We begin by examining all-cause mortality and daily mean temperature for Toronto, CD 3520, between 1981 and 2007. Removing the arithmetic mean from both series across the entire span, we compute multitaper spectral estimates for each and then compute the cross-spectra and use this to find the MSC and phase\footnote{This functionality is included in the \texttt{multitaper} package for R [148] as functions \texttt{spec.mtm}, \texttt{mtm.coh} and \texttt{plot.mtm.coh}.}. Figures 6.3 and
show the [0, 10] cycles/year frequency band of the MSC and phase respectively.

Figure 6.4: Between-series phase for All-Cause Mortality and Temperature for Toronto for the full 1981–2007 time span. The red horizontal dashed lines are at ±180°. Note the behaviour from 0 to ≈ 0.5 cycles/year and the drop near 1 cycles/year down to ≈ −180° phase. This behaviour then persists through to almost 5 cycles/year, and then shows changes. The grey dashed lines are the ±1σ confidence bounds, found by jackknifing.

The two plots demonstrate a combined behaviour near the origin: almost perfectly incoherent, and effectively zero phase. This behaviour is due to the analysis of the full 27-year data set which has mean set to zero. Recall that the zeroth frequency represents the term

\[ S(0) = \frac{1}{K} \left\| \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} x_n v_n^{(k)} c_0 \right\|^2 \]

(6.6)

the average of the square of the dot products of the Slepian sequence tapers and the data series. As the data series sums to 0, this value will be quite small (although
not exactly zero); examining the spectrum of a series which has been zero-meaned will typically show a “hole” at the origin due to this procedure. In summary, this relationship at the origin is simply a symptom of the zero-meaning process, and is not terribly interesting.

However, moving slightly to the right in frequency in Figure 6.4, there is a blip at almost 180° phase. This is followed by a block from ≈ 1 to 5 cycles/year where the phase hovers around −180° (or 180° equivalently). This is quite interesting: this low-frequency structure which demonstrates a 180° phase reflects the nature of the seasonality of mortality and temperature. We further examine this in Figures 6.5

![Figure 6.5: Magnitude-Squared Coherence between All-Cause Mortality and Temperature for the summer (days 120 to 240) of 1983, for Toronto, CD 3520. The bandwidth $W = 15.2$ cycles/year and is the approximate width of the first rectangular coherence block. The dashed line is the $-1\sigma$ confidence bound.](image_url)
and 6.6 for 1983. This year was chosen as a particularly clean example of our effect. The figures show an expanded low-frequency range, from 0 to 50 cycles/year (corresponding to periods as low as 7 days).

Figure 6.6: Between-series phase for All-Cause Mortality and Temperature for the summer (days 120 to 240) of 1983, for Toronto, CD 3520. The bandwidth $W = 15.2$ cycles/year and is the location of the red dashed line. Note how the phase of the two series is $\approx 180^\circ$ from 0 to 15 cycles/year, and then approximately zero out to 50 cycles/year. The grey dashed lines are the $\pm 1\sigma$ confidence bounds, found by jackknifing.

Figures 6.5 and 6.6 present an interesting picture of the relationship between mortality and temperature. The coherence plot (Figure 6.5) shows strong coherence near the origin, corresponding to the anti-seasonality, as mentioned before. It also shows almost no coherence between 15 and 25 cycles/year, but then climbs to a reasonable level (exceeding 98%) for the 25 through 50 cycles/year band. Examining the phase plot (Figure 6.6) we see that throughout the band from 15 to 50 cycles/year there is effectively zero phase between the two series. This corresponds to timescales from 7 to $\approx 25$ days, which are the periods that Gasparrini et al. [62, 66] found to have some structure, leading them to conclude that a response relationship persisted out
to 30 days lag for warm or hot temperatures. This result is therefore intriguing, and will inform our future work on this topic. Further note that the jackknifed confidence bounds on the phase estimate in Figure 6.6 are quite tight, indicating a reasonable amount of credibility in the shape of the phase plot.

### 6.6.2 Mortality and NO$_2$

We further examine the coherence and phase between mortality and NO$_2$, again for Toronto, and again for both full timescales (1981–2007) and for year-at-a-time summer months. These results are presented in Figures 6.7 and 6.8.

![Magnitude-Squared Coherence for All-Cause Mortality and NO$_2$ for the 27 years from 1981–2007 for Toronto, CD 3520. The dashed line is the $-1\sigma$ confidence bound, found by jackknifing. Note that the coherence at 1 cycle/year is less strong in this case than in the Mortality-Temperature case – this is understandable given the less coherent yearly periodicity in NO$_2$.](image)

Figure 6.7 shows that there are a number of moderately coherent periodicities
between Mortality and NO₂, although only one of high significance in the [0, 10] cycles/year band. The region below 1 cycle/year is incoherent as before. In Figure 6.8, there are a number of regions which can be considered to be either approximately 0 phase (near the origin, 6–6.5 cycles/year, 8.5–9.0 cycles/year, etc.) or approximately ±180° phase (1.5–2.0 cycles/year, 3–4 cycles/year, etc.). It is difficult from this to generalize anything with respect to an appropriate bandwidth for the time-based smoother, so we move to an examination of the MSC and phase for summer only.

The summers, being very short series, can exhibit more erratic behaviour. However, in aggregate they present an interesting picture. Figures 6.9 and 6.10 show the coherence and phase between All-Cause Mortality and NO₂ for the summer of 2006 for Toronto, CD 3520. These two figures clearly show a coherent low-frequency
structure (largely the mean and trend) which has zero phase, and then incoherent structure from 15 to 50 cycles/year which has wandering phase. A large number of the summers exhibit this same behaviour, indicating that it is a common response. From this we see that on small timescales, estimation of the mean and trend are very important for ensuring lack of coherence in low-frequencies. This is of particular interest to us because of our interest in yearly AHI risk estimates for trend estimation.

It can be quite difficult to interpret phase plots between two series, especially series as complicated as environmental pollutants. One of the more valuable applications is in examination of specific small bandwidths, e.g. the region between 0 and 7 cycles/year. With careful use of trend estimators, it may be possible to tease out the
response relationship for mortality and an air pollutant and use this to better inform our choice of smoother bandwidth, as initially discussed in Chapter 5.

6.6.3 Future Work

As briefly shown in this section, it is possible to examine the coherence and phase between two series and to gain some insight. The choice of $W$ bandwidth that has traditionally been selected based on the sensitivity of the estimate of the coefficient

Figure 6.10: Between-series phase for All-Cause Mortality and NO$_2$ for the summer (days 120 to 240) of 1983, for Toronto, CD 3520. The bandwidth $W = 15.2$ cycles/year and is the location of the red dashed line. Note how the phase of the two series is $\approx 0^\circ$ from 0 to 15 cycles/year, and then is erratic from there to 50 cycles/year. Contrast this with the MSC, which shows the two series are coherent for the first 15 cycles/year and then largely incoherent from 15 to 50 cycles/year. The grey dashed lines are the $\pm 1\sigma$ confidence bounds, found by jackknifing.
of an air pollutant may actually be selecting for the phase transition point of temperature and mortality. Both are certainly included in both the classic model and the sensitivity simulations of [141], and we further know that temperature has a much stronger relationship with mortality than air pollution (see, e.g., [62, 66] for further discussion). Further results from the tools presented in this section must be left for future work, but we believe the use of these tools may enable significant gains in interpretation and understanding, especially with respect to the (now continuous) choice of smoother bandwidth.

6.7 Analysis of 6 Canadian Cities

Using the ideas introduced in the previous two sections, and the tutorial presentation in Appendix D, we remove the three weekly harmonic periodicities from lag-1 NO$_2$ and daily All-Cause and Cardio-Pulmonary mortality. Lag-1 was chosen for simplicity, as discussed previously. We align the phase of the NO$_2$ weekly harmonics to match precisely with those of the two mortality series, and then re-embed them in the residual NO$_2$ series. We then run four models for each city and mortality type: a) the standard model of Eqn. 6.3 with no line component correction; b) Eqn. 6.3 with the DOW term removed; c) Eqn. 6.3 with line component correction and with the DOW term present; and d) Eqn. 6.3 with line component correction and with the DOW term removed. Models (a) and (d) are effectively the two extremes mentioned earlier in the introduction of Section 6.5.

The results for these four models, six cities, and two mortalities are presented in Tables 6.3 – cases a) and c) – and 6.4 – cases b) and d). These results show that when the line components are phase-aligned and the day-of-week term left in the model (Table 6.3) that the resulting risk estimates are largely unchanged. In the case of the All-Cause mortality this is almost entirely true – the changes, if they exist, are
in the 8\textsuperscript{th} decimal and are insignificant in the extreme. For the Cardio-Pulmonary mortality there are some larger changes, but still nothing of magnitude $1e - 6$ or larger. Effectively, the change went unnoticed by the model.

\begin{table}
\centering
\begin{tabular}{|c|cc|cc|}
\hline
\text{CD} & \text{AC - Model (a)} & \text{AC - Model (c)} & \\
\text{} & \text{Coef} & \text{StdE} & \text{Coef} & \text{StdE} \\
\hline
4806 & 5.2377e-04 & 5.5638e-04 & 5.2379e-04 & 5.5638e-04 \\
5915 & 1.8412e-03 & 4.6168e-04 & 1.8412e-03 & 4.6168e-04 \\
\hline
\text{CD} & \text{CP - Model (a)} & \text{CP - Model (c)} & \\
\text{} & \text{Coef} & \text{StdE} & \text{Coef} & \text{StdE} \\
\hline
5915 & 2.1158e-03 & 6.3760e-04 & 2.1158e-03 & 6.3760e-04 \\
\hline
\end{tabular}
\caption{Results for two models based on Eqn. 6.3: a) the standard model of Eqn. 6.3 with no line component correction; and c) Eqn. 6.3 with line component correction and with the DOW term present. The first table is All-Cause mortality as response, and the second Cardio-Pulmonary. The lag-1 NO\textsubscript{2} pollutant was used.}
\end{table}

Table 6.4 represents the second extreme where all of the power in the weekly harmonics for the pollutant are assumed to be causally linked to those same harmonics in the mortality. In this case, there are more significant changes to the risk estimates as the mortality’s harmonics are unmasked and can be fit to. Four of the six cities show positive increases in risk for All-Cause mortality, while Peel shows a decrease and Calgary stays roughly the same. Positive increases in the risk estimate are understandable when adding a perfectly correlated periodic term, as this will pull a linear model fit slightly toward the term’s coefficient, which in this case is positive and
Table 6.4: Results for two models based on Eqn. 6.3: b) the standard model of Eqn. 6.3 with no line component correction and the DOW term removed; and d) Eqn. 6.3 with line component correction and with the DOW term removed. The first table is All-Cause mortality as response, and the second Cardio-Pulmonary. The lag-1 NO\textsubscript{2} pollutant was used.

<table>
<thead>
<tr>
<th>CD</th>
<th>AC - Model (b) Coef</th>
<th>AC - Model (b) StdE</th>
<th>AC - Model (d) Coef</th>
<th>AC - Model (d) StdE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2466</td>
<td>5.1214e-4</td>
<td>3.0845e-4</td>
<td>5.2532e-4</td>
<td>3.0840e-4</td>
</tr>
<tr>
<td>3520</td>
<td>7.8286e-4</td>
<td>2.2872e-4</td>
<td>8.6018e-4</td>
<td>2.2875e-4</td>
</tr>
<tr>
<td>3521</td>
<td>8.7495e-4</td>
<td>5.8597e-4</td>
<td>8.4090e-4</td>
<td>5.8605e-4</td>
</tr>
<tr>
<td>4806</td>
<td>6.9706e-4</td>
<td>5.2629e-4</td>
<td>7.8924e-4</td>
<td>5.2622e-4</td>
</tr>
<tr>
<td>4811</td>
<td>4.2797e-4</td>
<td>5.0790e-4</td>
<td>4.2727e-4</td>
<td>5.0790e-4</td>
</tr>
<tr>
<td>5915</td>
<td>1.8693e-3</td>
<td>4.3666e-4</td>
<td>1.8718e-3</td>
<td>4.3653e-4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CD</th>
<th>CP - Model (b) Coef</th>
<th>CP - Model (b) StdE</th>
<th>CP - Model (d) Coef</th>
<th>CP - Model (d) StdE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2466</td>
<td>-1.4451e-4</td>
<td>4.3078e-4</td>
<td>2.0613e-4</td>
<td>4.3172e-4</td>
</tr>
<tr>
<td>3520</td>
<td>4.5698e-4</td>
<td>3.4336e-4</td>
<td>7.0908e-4</td>
<td>3.4300e-4</td>
</tr>
<tr>
<td>3521</td>
<td>2.3036e-4</td>
<td>9.1214e-4</td>
<td>2.3942e-4</td>
<td>9.1203e-4</td>
</tr>
<tr>
<td>4806</td>
<td>6.4506e-4</td>
<td>7.4966e-4</td>
<td>8.6165e-4</td>
<td>7.4946e-4</td>
</tr>
<tr>
<td>4811</td>
<td>-3.9981e-4</td>
<td>7.1527e-4</td>
<td>-3.5862e-4</td>
<td>7.1518e-4</td>
</tr>
<tr>
<td>5915</td>
<td>1.6554e-3</td>
<td>6.0411e-4</td>
<td>1.9377e-3</td>
<td>6.0500e-4</td>
</tr>
</tbody>
</table>

greater than any of the risk estimates. In the case of Cardio-Pulmonary mortality, all six cities show positive increases – this mortality data set is less noisy and known to be more strongly related to pollution, so the effect coming through is not surprising.

In conclusion, we have argued that the true risk estimate (as far as the weekly harmonics are concerned) likely lies somewhere between two extremes: the results of model (a), Table 6.3, and the results of model (d), Table 6.4. Where the truth lies is unfortunately a topic for further research and beyond the scope of this thesis – we will continue our analysis by estimating these two extremes and reporting both risk estimates under the understanding that the truth lies somewhere between the two.
6.8 Temperature

The final untouched predictors in the model of Eqn. 6.3 are the smooth functions of temperature. In the AHI model we began with, the two covariates were lag-0 temperature and the 3-day moving average of lags 1, 2 and 3. As we noted in Section 6.4, a 3-day moving average is effectively a crude low-pass filter, which seems an odd choice in the context of trying to fit high-pass mortality as a predictor. We will discuss this further below. The discussion in Section 6.6.1 and 6.6.3 is also pertinent for temperature, but less for an understanding of temperature as a predictor, and more for understanding of the relationship between temperature and mortality.

The primary reason for including temperature as a predictor for mortality is its indisputable relationship as a causal link. The rationale for including it via a low-degree-of-freedom smooth is the classic “U shaped” relationship between the two [170]. This U shape comes from three factors. The first is that at very low temperatures mortality is typically increased – both due to health issues directly, and due to confounding with flu epidemics which tend to occur during cold seasons. The second is that at moderate temperatures mortality decreases to its minimum – essentially, it neither helps nor harms human health. Then, at high temperatures mortality increases due to the conflux of extreme heat and high humidity. This nonlinear relationship can typically be seen clearly in a smooth of temperature against raw mortality; as an example, see Figure 6.11 where we demonstrate a number of natural cubic regression spline smooths of daily mean temperature against daily mortality for 27 years of data from Toronto, CD 3520. The form of the smooths in this figure are typical of other data sets from both the NMMAPS and AHI databases.

Leveraging the additional insight we gained from Chapter 5, we must question the basic assumption behind the smooth function of temperature\(^5\). As we see clearly

\(^5\)Note that we do not question the inclusion of temperature in the model. If anything, temperature is the most important predictor in the model.
Figure 6.11: Plots of lags 0, 1, and the 3-day moving average “123” against All-Cause mortality for 27 years (1981–2007) of mortality for Toronto, CD 3520. Across the plot are three different choices of degrees-of-freedom for the natural cubic regression spline smoother.

From Figure 6.11, there is a nonlinear relationship between mortality and temperature. However, when we setup the GAM to immediately filter the low-frequency variation in the mortality, it is less clear what our smooth function of temperature is actually smoothing against. In Figure 6.12 we display smooths of temperature against residual
Figure 6.12: Plots of the smooths of lags 0, 1, and the 3-day moving average “123” temperature against temperature. The smooths are obtained by smoothing temperature against All-Cause mortality for 27 years (1981–2007) of mortality for Toronto, CD 3520. Model includes an additional Slepian projection smoother function of time, see Eqn. 6.7. Across the plot are three different choices of degrees-of-freedom for the natural cubic regression spline smoother. Note that this smoother is nonlinear.
effective mortality from a simplified two-covariate model:

\[
\log(\mu_t) = S_1(\text{temp}_t, df = 3) + S_2(\text{time}, W = 7/365) \tag{6.7}
\]

where \( l \in \{0, 1, 123\} \), \( S_1 \) is a cubic regression spline, and \( S_2 \) is a Slepian projection smoother. The models are again run on 27 years of Toronto data, using All-Cause mortality. From Figure 6.12, we see that for lag-0 and lag-1 mortality, once the low-frequency portion of the mortality is taken care of, the resulting smooth is much closer to a monotonically increasing function, possibly even linear. This is not the case for the 3-day moving average, which retains some semblance of the previous U shaped nonlinear relation.

Given the shapes seen in the top six sub-plots of Figure 6.12, we must question the rationale for including the \textit{smoothed} function of temperature (as compared to a linear or distributed-lag function). Recall from Chapter 5 that we have already discovered spurious (high power) structure being introduced by the non-linear smoother. The main reason for the inclusion of temperature in this fashion was to access the non-linear “bend” that occurred in the temperature-mortality relationship, but as we have demonstrated, if the mortality is effectively filtered in the GAM, the resultant smooth does not “bend” at all, and in fact is approximately monotonic. Thus, and we leave this as an open question for future work, as it begins to pass beyond the scope of this thesis, should lag-0 or lag-1 temperature: a) be included in the model at all; and b) if one or both is to be included, should it be smoothed or instead included in some parametric form? The next section evaluates a sub-question related to (a), which is: should multiple lags of temperature be included in the model? Note that even beyond the zeroth and first lags of temperature, evidence has been discovered which suggests that lags out to 10 or more days can have statistically significant impacts on mortality [62, 66]. Further work is definitely needed in this area.
6.8.1 Multiple Lags of Temperature

The next model that we investigated was one which included both the zeroth and first-order lags of temperature, both linked via low-order smooth functions. Begin with the model thus far, ignoring the weekly harmonics for this example as they do not influence the results:

$$
\log(\mu) = \beta_1 (NO_2)_{t-1,HP} + S_1(\text{temp}_0, df = 3) + S_2(\text{temp}_1, df = 3) + S_3(\text{time}, df = 7/yr) + \text{as.factor(DOW)},
$$

where the smooth function of time uses the Slepian projection smoother, and the other two smooths are cubic regression splines with low degrees-of-freedom.
Consider Figure 6.13. Here we show the two fitted smooths of lag-0 and lag-1 temperature over two yearly cycles. Note how the fits are out of phase with one another (one peaks where the other troughs, and vice-versa). This leads to the overall effect of the first being reduced, as much of the second fit goes to canceling out the effect of the other term. The effects for the combination of lag-0 and the 3-day moving average of lags 1, 2 and 3 is even more pronounced, with the amplitudes of the two predictions being comparable and the resultant sum largely cancelling. We suspect (but cannot prove) that the reduction in deviance observed by Health Canada researchers when first including the second lag-temperature term came from the reduction in low-frequency power confounding through the two terms cancelling – see Section 5.4 for further details on the confounding problem.

We have run extensive tests comparing models with/without smoothing effects, with/without low-frequency temperature power, and with one of the two lags for temperature considered. In the end, there was no clear answer as to which was the best. For interpretation reasons, we advocate using the same lag for temperature as for the pollutant, which for our current model gives us lag-1 only. Additionally, once the changes from the following section are taken into account, there is no low-frequency power remaining in our GAM, so including low-frequency temperature seems counter-intuitive. Further study is warranted to determine the precise nature of the relationship between temperature and mortality under the condition that the mortality response seen is modified by an effective smooth function of time filter and one or more pollutants.

6.9 Finalized Single Pollutant Model

We did not explicitly examine the smooth function of time in this section, as the entirety of Chapter 5 was concerned with its behaviour and our proposed replacement,
spsmooth. Including the Slepian projection smoother, a single smooth function of lag-1 temperature, lag-1 high-pass filtered NO\textsubscript{2} as the pollutant with the weekly harmonics aligned, and alternating inclusion of the DOW term, we have our final equation. Write

\[
\log(\mu_t) = \beta (\text{NO}_2)_{t-1, \text{HP}} + S_1(\text{time}, W = 7/365) + S_2(\text{temp}_{t-1}, df = 3) \\
+ I_j (\text{DOW}),
\]

where \(S_1\) is a Slepian projection smoother, \(S_2\) is a cubic regression spline smoother, and \(I_j\) is an indicator function allowing us vary the inclusion of DOW. These two models are compared in the following to the original Health Canada AHI model of Eqn. 6.2 with the difference that the pollutant in Eqn. 6.2 is replaced with lag-1 NO\textsubscript{2}.

### 6.10 Full Database Analysis

In this section we use the finalized single pollutant model of the previous section, Eqn. 6.9, and apply it to 24 census divisions from the AHI database. Each model is run on 27 years\(^6\) of data, 1981–2007, using All-Cause and Cardio-Pulmonary mortality. A variation on the original AHI model, Eqn. 6.2 is also run:

\[
\log(\mu_t) \mid= \beta (\text{NO}_2)_{t-1} + S(\text{temp}_0, df = 3) + S(\text{temp}_{1-3}, df = 3) \\
+ S(\text{time}, df = 7/year) + I_j \gamma \text{DOW},
\]

with \(S\) a natural cubic regression spline. Thus, between the two mortalities, the inclusion/ exclusion of DOW, and the two models (Eqns. 6.10 and 6.9), there are 8 models in total that we apply to the database. The full set of plots of the coefficients with \(2\sigma\) confidence intervals is presented in Appendix C.

Referring to Table C.1, recall that in the previous section we found that the inclusion/exclusion of the DOW in the case of the weekly harmonic phase-alignment led to the two extremes. Thus, in these results, we can think of Models 5 and 6, and

\(^6\)21 years in the cases of Quebec City and Montréal, 1981–2001.
Models 7 and 8 as being the two extremes for the risk estimate of our finalized model for All-Cause and Cardio-Pulmonary mortality, respectively. As should be clear from Figures C.7 and C.8, there is not a significant amount of variation between the two, so the conclusions drawn from previous analyses should remain valid.

In general, viewing Figures C.1–C.4, we see that the blue and green confidence intervals (corresponding to our updated model, Eqn. 6.9) are consistently more positive than the black and red intervals (corresponding to the original AHI model, Eqn. 6.10). This holds regardless of which mortality series is used, and across almost all of Canada – 24 census divisions separated by up to 5800 km (Vancouver to Halifax). This is not a firm demonstration, but as the data remained the same and the models remained similar excepting our updates, it is indicative of a possible improvement in the underlying estimation of risk.

6.11 Discussion and Further Work

This chapter has barely scratched the surface with respect to analysis that can be done on this large-scale data set. We showed (from Chapter 5) that the use of natural cubic regression splines as time-based smoothers leads to undesirable long timescale variation, and proposed a replacement smoother that corrected this problem. We further (in Chapter 6) redeveloped a model for population health risk due to air pollution that: a) demonstrated spectrally why the 3-day moving average of pollution is not desireable; b) determined that lag-0 NO\textsubscript{2} has a much less significant relationship to mortality than lags 1 and 2; c) went into some detail on phase-alignment of weekly harmonic periodicities and the same for mortality and predictor series as a whole; and d) examined smooth functions of temperature and showed that the traditional “U shaped” nonlinear relationship between temperature and mortality does not hold under the time-based smoother constraint. We finally formalized a model and ran
this model against all 24 census divisions contained in the AHI, showing that the use of our updated model, Eqn. 6.9, gave consistently more positive risk estimates than the original model, Eqn. 6.10.

Further work for this topic includes an intensive examination of the temperature covariates and their smooths, and hopefully a determination of the best relationship to use for a model which has no low-frequency variation. It is possible that pre-filtering the temperature will also contribute slightly to reduced long timescale power, which in turn will make the “local in time” assumption stronger for inference purposes. In general, the nonlinear smooth that is used for temperature in the GAM structure does not seem interpretable – it is not clear what its purpose in the model is when the nonlinear “bend” relationship does not hold.

We further plan a full sensitivity analysis in the style of Peng et al. [141] to determine how the new \texttt{spsmooth} time-based smoother behaves when presented with synthetic data and with varying parameters. In this work we did not even touch on the point that varying the parameter of the Slepian projection smoother can be done continuously: \( W \) is a real variable and can be varied from 0 through 5.5\( \mu \)Hz (Nyquist for daily data), eliminating entire timescales with minimal effort. It may be the case that the previously determined 7 cycles/year cutoff (which was actually \( \approx 3.5 \) cycles/year) is not optimal\(^7\), and we may be able to more rigorously define what happens as the timescale is varied.

Extending the univariate pollutant to a multivariate case is an interesting problem, and although we have completed some initial tests and “proof-of-concept” work, the subtle details remain to be worked out. Extensive work remains in the estimation of \textit{national} and \textit{regional} risk estimates – hierarchical Bayesian models have had some success in the past, and running these models against our new estimates of risk may

\(^7\)In personal discussions with epidemiologists working in this area, the point was raised that there is \textit{no} epidemiological reason for choosing 7 \textit{dof}/year, and that this choice was made purely on the basis of statistical sensitivity analysis.
lead to new understanding. Finally, recall that the AHI was intended to be computed on smaller data sets, down to length 1 year. Generalization of our techniques and model from the full data span down to shorter data spans is possible, and early tests are promising, but the work remains incomplete.
Chapter 7

Conclusions and Directions for Future Work

In the first part of this thesis we demonstrated two new methods for interpolation of time series. These methods proved to be as efficient as the best least-squares linear prediction methods of Wiener [220] for flat spectra, while allowing for generalizations to the time series, in particular the inclusion of periodic line components. Code was developed for both of the interpolation methods and will be freely available on CRAN, the Comprehensive R Archive Network. Examination of the interpolator performance on an ARMA model example showed less promising results, largely driven by the autoregressive nature of the prediction. The target for the interpolators is series which have approximately white backgrounds and embedded line components, an area which covers many scientific data sets used in modern statistical exploration.

Further work in this area will require the implementation of a number of other methods which are available in the literature, but not readily available in code. A number of the papers referenced in Chapter 2.2 have interesting ideas in them, but in all cases code is not readily accessible, or when it is, is largely written to work on only a single example. Expanding the available codebase, perhaps through extension
of the tsinterp package, is a future research goal. This will also allow for the comparison between the different methods and possible improvements through hybridized combinations of them. We would also like to extend the underlying $W_t$ estimator to allow for cyclostationary series, as this allows for more interesting examples.

In the second part of this thesis we demonstrated an improved family of smoothers for Generalized Additive Models as used for estimation of risk due to air pollution. This development is a significant contribution to the field in three parts. The first part was the introduction of frequency-domain techniques to the examination of residuals and partial residuals in the Generalized Additive Model structure. This contribution leads to significantly improved intuition regarding the model performance, and led directly to the second part. The second contribution was in the introduction of a continuous concentration-based formulation for the smooth function of time included in these models. This formulation allows for significantly improved low-frequency residual power in the estimation of risk due to air pollution, giving more weight to the interpretation of the risk as “local in time”. Finally, the third contribution was in the demonstration of the spread-spectrum feature introduction inherent in parametric terms. This leads to the pre-filtering proposal for the pollutant, a technique which was previously used by researchers but which fell out of common usage in the last decade. These three contributions give an improved model structure with access to a number of new ways of thinking, and should be viewed as the primary accomplishment of this thesis.

In this area, there remains significant work ahead. The sensitivity of the new technique is an open question, and a paper in the vein of [141] will be necessary to prove to researchers in the field that the new method is appropriate for use on their data. We further plan to write a paper demonstrating the use of the spsmooth package to encourage further use of it. As of the time of defense of this thesis, two papers were under preparation, the first presenting these new modeling techniques
to epidemiology researchers (written in a more tutorial manner), and the second presenting the statistical justifications for their development and aimed at a more statistical audience. Significant further work is needed to cement this approach in the standard framework, but the proof-of-concept and initial results are complete and appear highly promising.

Continuing in the second part of the thesis, a new phase-aligned pollutant model was derived and tested. Comparisons with earlier models show that the risk estimates for the new model are consistently more positive, which may indicate a stronger relationship between air pollution and mortality than was previously thought. We further explored the use of smooth functions of temperature and showed that the conventional “U shaped” interpretation of the relationship between temperature and mortality does not hold when temperature is smoothed against residual mortality (after accounting for low-frequency variation). The primary concern with the work in this thesis was demonstrating that the risk estimates remain similar but that correcting the long timescale variation bias allows the interpretation to be more precise and reliable. Further work in this area includes development of satisfactory pooling estimates for national risk, further analysis of the structure of the relationship between mortality and temperature, and especially examination of the distributed-lag relationship between temperature and mortality with the possible application of our new smoothing methodology.


Bibliography


[32] W.S. Burr, A. Moghtaderi, G. Takahara, and H.H. Shin. Estimation and Extrac-
tion of Harmonic Features from Pollution and Mortality Time Series. Statistical

[33] W.S. Burr, G. Takahara, and H.H. Shin. Towards Developing an Air Health
Indicator using Spectral Decomposition of GAMs. Statistical Society of Canada

[34] C. Chatfield. Inverse autocorrelations. *Journal of the Royal Statistical Society:*

[35] R.B. Cleveland, W.S. Cleveland, J.E. McRae, and I. Terpenning. STL: A
seasonal-trend decomposition procedure based on LOESS. *Journal of Official

[36] W.S. Cleveland. The inverse autocorrelations of a time series and their appli-

[37] E.W. Cliver. Solar activity and geomagnetic storms: The first 40 years. *EOS,  

Transactions on Audio and Electroacoustics.

Fourier Transform Algorithm to the Estimation of Spectra and Cross-Spectra.
In *Computer Processing in Communications*, volume 19, pages 5–20. Polytech-
nic Institute of Brooklyn Microwave Research Institute Symposia Series, 1969.


[73] F.J. Harris. On the Use of Windows for Harmonic Analysis with the Discrete


188


[188] G. Takahara, A. Moghtaderi, and W. Burr. Investigation of Cyclic Patterns...


Appendix A

spsmooth: An Extension Package for mgcv

spsmooth is a small R package written to provide access to the developments of Chapter 5. The package integrates into the mgcv [227] package and gives analysts access to the alternative Slepian projection smoother. This appendix will present a small vignette on how to use the Slepian projection smoother, and the meaning of the user-accessible parameters. All code snippets provided are intended to be run from within R, and were tested on version 2.15.1 of the same. The versions of spsmooth and mgcv used were 1.0—0 and 1.7—19, respectively. The built-in dataset chicagoNMMAPS from the spsmooth package is used in this vignette, and was originally sourced from the NMMAPSData package [142].

Begin by running R and loading the packages:

> library("mgcv")
> library("spsmooth")

The help file for spsmooth contains most of the information included in this appendix:

> ?spsmooth
As mentioned in the help file, the only function included in `spsmooth` that is intended to be called by the user is `summary()`, which is a wrapper function for `summary.gam()` from the `mgcv` package.

We assume that the reader is familiar with the basic R interface for model building, including the `formula` class. Whether a linear model (`lm`), a generalized linear model (`glm`), a generalized additive model (`gam`) or any of R's other supported model types, the basic function call is the same. The following is a toy linear model example.

```r
> x <- 1:100
> y <- rnorm(n=100)
> MyFormula <- y ~ x
> MyData <- cbind(x,y)
> fit <- lm(MyFormula, data=MyData)
```

In the case of `mgcv` and its implementation of GAMs, there are a few other parameters that should always be defined: the `family` and the `na.action`. These are demonstrated in the following:

```r
> fit <- gam(MyFormula, data=MyData, family=poisson, na.action=na.exclude)
```

The allowable families can be explored through use of the `?family` help file. One of the primary uses of GAMs over GLMs is for the inclusion of smooth functions of one or more covariates. In `mgcv`, this is accomplished by use of the `s()` wrapper function in the `formula` object.

```r
> MyFormula <- y ~ s(x)
```

In this case, `s()` defaults to a thin-plate regression spline. There are a number of possible options for choosing which type of smoothing spline you would like for your covariate. From the list of defaults we are primarily interested in the natural cubic regression (`cr`) splines:
> MyFormula <- y ~ s(x, bs='cr')

Once the user has chosen a basis, there are additional parameters that can be set. The three that interest us for this vignette are k, fx and xt. The k parameter allows the user to define the number of knots – for epidemiologic time series analyses, this is often prespecified based on prior (scientific) reasoning. For example, we might write:

> MyFormula <- y ~ s(x, bs='cr', k=floor(7/365*1000))

to define a cubic regression spline smooth with 7 dof/yr. The fx parameter allows us to move from the default state of penalized regression splines (that is, where the degrees-of-freedom is flexible and can be optimized by one of a number of criteria) to fixed regression splines. This parameter is Boolean, so

> MyFormula <- y ~ s(x, bs='cr', k=floor(7/365*1000), fx=TRUE)

Finally, the xt parameter is used to pass extra parameters and settings to specific spline smoothing classes via a list object.

The spsmooth package implements several constructor functions which allow for sp (Slepian projection) smooth functions to be defined. For sp, the k parameter allows for choice of the number of basis vectors (Slepian sequences) used, but it is not recommended that the default value be changed. Instead, we encourage users modify the W parameter, which is passed via xt. The fx parameter also has no purpose for the sp smoother, as penalized regression using this basis is not a supported feature.

> MyFormula <- y ~ s(x, bs='sp', xt=list(W=7/365))

Finally, there is one additional parameter that must be set for sp in the case of missing data – a Boolean (true/false) mask array indicating the joint status of all lags of the covariates. This is admittedly clumsy, but is required due to the way mgcv deals with missing data values. Slepian sequences are meaningless if arbitrarily
subsampling (i.e. taking only the first 2/3 of $N$ points), and the significantly improved properties of the smoother do not persist if (say) they are created to be length $M \ll N$ for a series with $M$ available values out of $N$ total. The sequences must be created of length $N$ and then appropriately sampled to obtain the $M$ available values. This does eliminate the orthogonality of the sequences, but preserves the projection filter structure as much as possible.

Now, we demonstrate a simple analysis of the built-in Chicago dataset. Begin by defining two models: the first is a simplified version of the model by Dominici et al [52], the second is a similar version that replaces the natural cubic regression spline smooth of time with a Slepian projection smooth.

```r
> data("chicagoNMMAPS")
> str(chic)
> MyFormula1 <- death ~ pm10tmean + s(tmpd, bs='cr', k=6, fx=TRUE) +
  s(dptp, bs='cr', k=3, fx=TRUE) + as.factor(dow) +
  s(time, bs='cr', k=98, fx=TRUE)
> MyFormula2 <- death ~ pm10tmean + s(tmpd, bs='cr', k=6, fx=TRUE) +
  s(dptp, bs='cr', k=3, fx=TRUE) + as.factor(dow) +
  s(time, bs='sp', xt=list(W=7/365.25*14, mask=chic$mask))
> fit1 <- gam(Model1, data=chic, na.action=na.exclude, family=quasi(log,mu),
             control=list(epsilon=1e-6, maxit=50))
> fit2 <- gam(Model2, data=chic, na.action=na.exclude, family=quasi(log,mu),
             control=list(epsilon=1e-6, maxit=50))
```

From this point on, the models behave similarly – the overloaded `summary()` function included in `spsmooth` ensures that summaries and predictions of `sp` objects proceed as the user would expect. To conclude, we show how the user might obtain the residual effective mortality mentioned in Chapter 5.
For mgcv, to obtain the residual response as a certain covariate “saw” it (upon convergence), we add together the overall model residuals and the prediction for the covariate of interest. For example, to obtain the residual effective mortality as seen by the pm10tmean pollutant:

```r
> pred <- predict(fit1, type="terms")
> pm10p <- pred[, "pm10tmean"]
> resid <- rep(NA, length(chic[, 1]))
> resid[chic[, "mask"]==TRUE] <- fit1["residuals"]
> resEffMort <- resid + pm10p
```

Finally, to fit the residual effective mortality by the pm10tmean to obtain the same coefficient as was obtained in the mgcv model, as was described and used in Chapter 5, we need to pull the weights at convergence from fit1 and include them in the call to lm.

```r
> wts <- rep(NA, length(chic[, 1]))
> wts[chic[, "mask"]==TRUE] <- fit1["weights"]
> MyDat <- cbind(resEffMort, chic[, "pm10tmean"], wts)
> fit3 <- lm(resEffMort ~ pm10tmean, data=chic, weights=wts)
```
Appendix B

tsinterp: A Time Series Interpolation Package

tsinterp is a mid-sized R package that was written to encapsulate the algorithms of Chapters 3 and 4. It is still under development as of September, 2012, with the latest version available on github at http://www.github.com/wesleyburr/tsinterp/. The package provides wrappers for the univariate and bivariate interpolators developed in this thesis as well as utility routines for finding and categorizing gaps. This appendix will demonstrate the use of some of these functions.

Begin by running R and loading the package:

> library("tsinterp")

The help file for tsinterp contains most of the information included in this appendix:

> ?tsinterp

The most important two functions in the package are interpolate and BiVarInt, which are wrapper functions for all of the others. Their syntax is as follows:

    interpolate (z, gap, maxit, progress, sigClip)
Parameters:

\( z \): Real contiguous input series; missing values can be NA or real (i.e. \(-999.99\))

\( \text{gap} \): Indexes of the missing values of \( z \) for the span passed to the function

\( \text{maxit} \): Maximum number of outer iterations to perform

\( \text{progress} \): Boolean variable, if set to TRUE will write extra progress information to the screen

\( \text{sigClip} \): Probabilistic significance clip for \( T \) line component finding

\( \text{delT} \): Time step for \( z \) series

Returns (as a list):

\( zF \): The interpolation series \( z \) with all missing values replaced

\( p \): The number of steps until iteration stopped; if \( p \mid \text{maxit} \) then converge=TRUE

\( \text{diffC} \): The metric taken between the \( p \)th and \( p-1 \)st steps

\( zA \): List containing interpolation steps; length \( p \)

\( \text{converge} \): Boolean variable indicating whether the outer loop converged

\( \text{BiVarInt} \ (z_1, z_2, \text{gap1}, \text{gap2}, \text{maxit}, \text{progress}, \text{sigClip}, \text{delT}) \)

Parameters:

\( z_1 \): Real contiguous input series 1; missing values can be NA or real

\( z_2 \): Real contiguous input series 2; missing values can be NA or real

\( \text{gap1} \): Indexes of the missing values of series \( z_1 \) for the span passed to the function
gap2: Indexes of the missing values of series \textbf{z2} for the span passed to the function

\textbf{maxit}: Maximum number of outer iterations to perform

\textbf{progress}: Boolean variable, if set to \textbf{TRUE} will write extra progress information to the screen

\textbf{sigClip}: Probabilistic significance clip for $T_i$ line component finding; applies to both series

\textbf{delT}: Time step for \textbf{z1} and \textbf{z2} – the two series must share this parameter

Returns (as a list):

(the same items as \textbf{interpolate}; \textbf{zF} is the fill for \textbf{z1})

Note that \textbf{BiVarInt} does not interpolate the second \textbf{z2} series at all – the function can be called with the two series swapped in order to accomplish this.

There are a number of functions which are user-accessible which may be useful to those implementing their own interpolation routines, such as \textbf{estimateMt}, \textbf{estimateTt} and \textbf{estimateWt}. The first two of these are used in both the univariate and bivariate cases, while the third is only used for the bivariate case. In the univariate case the $W_i$ component of the interpolation is simple enough that it is mainlined into the \textbf{interpolate} function.

Additional functions that may be useful for general data analysis on series with missing values are \textbf{dpssap} (computes the associated polynomials of the DPSSs, see Section 3.2.3 for more), \textbf{removePeriod} (removes line component at specified frequency from a contiguous series) and the pair of functions \textbf{SpecToACV} and \textbf{SpecToACVdual} that take an input \textbf{spec} object\footnote{Returned by a routine like \textbf{spec.mtm} from \textbf{multitaper}.} and inverts it to obtain a consistent estimate of the autocovariance sequence for the original series that created the \textbf{spec} object. As
before, further information on all of these functions can be found in the package
documentation files.

Now, we demonstrate the univariate interpolation of the daily noon solar flux
from Sagamore Hill, as discussed in Section 3.5. This data set is included with the
tsinterp package, so load it using the data command and examine it:

```R
> data(flux)
> str(flux)
```

The missing values in the SagOrig series are indicated by $-99.9$ flux values, and the $S$
column contains a TRUE/FALSE array with FALSE values corresponding to the missing
observations. Note that the interpolate function is not setup to work well on an
entire series (in this case, 15251 points) being passed to it. In the code on the following
page we determine where the gaps are and repeatedly call interpolate to fill portions
of the series. This also keeps the ACVF estimator in the routine from trying to fit
the 11-year cyclostationary variance that was discussed in Section 3.5. The choice
to not include a routine which does this in the package was deliberate: interpolating
time series is not something that should be done in a completely automated fashion
– computers are only as smart as their programmers, and “keeping an eye” on the
interpolation is necessary to ensure edge cases have not arisen.

We do not discuss the bivariate interpolator here for sake of space, but there is
a vignette in the tsinterp package which goes into more detail on its function and
form.
Hand-rolled code for interpolating the Flux series

```r
dif <- flux[, "S"
> sag <- flux[, "SagOrig"
> finish <- FALSE
> while(!finish) {
    gap <- which(ok==FALSE)
    diff <- gap[-1] - gap[-length(gap)]
    pos <- which(diff > 50)
    if(length(pos) > 0) {
        maxL <- gap[pos[1]+1] - 1
        minL <- max(1, gap[1] - 100)
        # subset the gap series to match the subset of sag[]
        gap <- gap[gap > minL & gap < maxL]
        gap <- gap - minL + 1
        zP <- sag[minL:maxL]
        y <- interpolate(zP, gap, maxit=20, progress=FALSE, sigClip=0.999)
        sag[minL:maxL] <- y[[1]]
        ok[minL:maxL] <- rep(TRUE, (maxL-minL+1))
        if(length(which(ok==FALSE))==0) { finish <- TRUE }
    } else { finish <- TRUE }

}
Appendix C

Coefficients and Standard Errors for AHI Models

In this appendix we present several variations on the presentation of the coefficients and standard errors for models run on the AHI. The models run are:

<table>
<thead>
<tr>
<th>Model</th>
<th>Mortality</th>
<th>Time</th>
<th>Pollutant</th>
<th>Temp. 1</th>
<th>Temp. 2</th>
<th>DOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All-Cause</td>
<td>ns</td>
<td>(NO$<em>2$)$</em>{t-1}$</td>
<td>Lag-0</td>
<td>Temp.123</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>All-Cause</td>
<td>ns</td>
<td>(NO$<em>2$)$</em>{t-1}$</td>
<td>Lag-0</td>
<td>Temp.123</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Cardio-Pulmonary</td>
<td>ns</td>
<td>(NO$<em>2$)$</em>{t-1}$</td>
<td>Lag-0</td>
<td>Temp.123</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>Cardio-Pulmonary</td>
<td>ns</td>
<td>(NO$<em>2$)$</em>{t-1}$</td>
<td>Lag-0</td>
<td>Temp.123</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>All-Cause</td>
<td>sp</td>
<td>(NO$<em>2$)$</em>{lc,t-1}$</td>
<td>Lag-1</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>All-Cause</td>
<td>sp</td>
<td>(NO$<em>2$)$</em>{lc,t-1}$</td>
<td>Lag-1</td>
<td>–</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Cardio-Pulmonary</td>
<td>sp</td>
<td>(NO$<em>2$)$</em>{lc,t-1}$</td>
<td>Lag-1</td>
<td>–</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Cardio-Pulmonary</td>
<td>sp</td>
<td>(NO$<em>2$)$</em>{lc,t-1}$</td>
<td>Lag-1</td>
<td>–</td>
<td>No</td>
</tr>
</tbody>
</table>

Table C.1: Breakdown of the 8 models run against the AHI database.

In the above table, Temp.123 refers to the 3-day moving average of temperature at lags 1, 2 and 3, and (NO$_2$)$_{t-1}$ to lag 1 NO$_2$. The $lc$ subscript indicates the phase-aligned weekly harmonic version of the series. We break up these models into sets based on common features – for Figures C.1–C.4 we display them grouped by mortality, that is, Models 1, 2, 5 and 6 together, and Models 3, 4, 7 and 8 together. For
Figures C.5–C.8, we display them pairwise, comparing the inclusion/exclusion of the DOW term. The census divisions are numbered according to the Statistics Canada enumeration scheme. For the reader’s convenience Table C.2 provides a conversion table.

<table>
<thead>
<tr>
<th>CD Code</th>
<th>Region</th>
<th>CD Code</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1209</td>
<td>Halifax</td>
<td>3530</td>
<td>Waterloo</td>
</tr>
<tr>
<td>1301</td>
<td>St.John</td>
<td>3537</td>
<td>Windsor</td>
</tr>
<tr>
<td>2423</td>
<td>Quebec</td>
<td>3538</td>
<td>Lambton</td>
</tr>
<tr>
<td>2466</td>
<td>Montreal</td>
<td>3539</td>
<td>London</td>
</tr>
<tr>
<td>3506</td>
<td>Ottawa</td>
<td>3553</td>
<td>Sudbury</td>
</tr>
<tr>
<td>3518</td>
<td>Oshawa</td>
<td>3557</td>
<td>Sault Ste.Marie</td>
</tr>
<tr>
<td>3519</td>
<td>York</td>
<td>4611</td>
<td>Winnipeg</td>
</tr>
<tr>
<td>3520</td>
<td>Toronto</td>
<td>4706</td>
<td>Regina</td>
</tr>
<tr>
<td>3521</td>
<td>Peel</td>
<td>4711</td>
<td>Saskatoon</td>
</tr>
<tr>
<td>3524</td>
<td>Oakville</td>
<td>4806</td>
<td>Calgary</td>
</tr>
<tr>
<td>3525</td>
<td>Hamilton</td>
<td>4811</td>
<td>Edmonton</td>
</tr>
<tr>
<td>3526</td>
<td>Niagara</td>
<td>5915</td>
<td>Vancouver</td>
</tr>
</tbody>
</table>

Table C.2: Conversion table for CD codes and region names.
Figure C.1: Models 1,2,5 and 6 from Table C.1 – AC Mortality, CDs 1-12.
Figure C.2: Models 1, 2, 5 and 6 from Table C.1 – AC Mortality, CDs 13-24.
Figure C.3: Models 3, 4, 7 and 8 from Table C.1 – CP Mortality, CD 1-24.
Figure C.4: Models 3, 4, 7, and 8 from Table C.1 – CP Mortality, CDs 1-12.
Figure C.5: Models 1 and 2 from Table C.1 – AC Mortality, CDs 1-24.
Figure C.6: Models 3 and 4 from Table C.1 – CP Mortality, CDs 1-24.
Figure C.7: Models 5 and 6 from Table C.1 – AC Mortality, CDs 1-24.
Figure C.8: Models 7 and 8 from Table C.1 – CP Mortality, CDs 1-24.
Appendix D

Tutorial on Phase and Lag

Material in this section is adapted from two technical reports written for Health Canada by the author of this thesis[185, 186], in conjunction with Dr. Glen Takahara. This material is written at a “tutorial” level, as the intended audience included researchers with no time series background.

The context of this appendix is the discussion of line components in Chapter 6.5, and can be further applied to discussion of lag offsets in relationships between pollution and mortality.

D.1 Simple Demonstration of Phase Offset

As an example, consider a very simple system where the response is a single sinusoid with additive noise, and the predictor is a similar single sinusoid, at the same frequency but with a phase offset, with independent additive noise. Formally,

\begin{align*}
\text{response: } y &= 10 \cdot \sin \left( \frac{2\pi t}{100} + 0 \right) + \xi_t \\
\text{predictor: } x &= 10 \cdot \sin \left( \frac{2\pi t}{100} + 90^\circ \right) + \zeta_t,
\end{align*}

where \( \xi_t, \zeta_t \sim \mathcal{N}(0, 1) \) and \( t = 1, 2, \ldots, 1000 \). In this situation, the period of each of the sinusoids is 100 time steps and the response lags the predictor by a quarter of
that period, or 25 steps. Figure D.1 shows a realization of this example; the reader can clearly see that the variates share a common frequency and amplitude, but differ by a phase factor.

Figure D.1: Realization of Eqn. D.1, with the predictor in red and the response in black. Note the shared amplitudes and periods, but the differing phases.

Now, consider a very simple linear model:

\[ y \sim x \]  

(D.2)

where we artificially constrain the intercept to be 0 on the basis of our prior knowledge. As one would suspect, using for our predictor the red curve from Figure D.1 results in a very low coefficient, despite the curves being similar in all but one way. We implement the above model, varying the predictor from a lag-0 copy of the second simulated series through to a lag-99 copy, one full period of the predictor, and exhibit the coefficients obtained in Figure D.2.

221
Figure D.2: Coefficients from linear model fitting response to lagged copies of predictor, lags ranging from 0 to 99. Note the maximum occurs at lag-25 (or 90°), exactly 1/4 of the total period, as the model was created.

The reader should notice that the coefficient can be negative, in fact strongly so, despite the curves being near identical (excepting the phase offset). This occurs in situations where two sinusoidal signals are at exactly opposite phases so they effectively are anti-correlated; in these cases the best-fit linear model coefficient will be negative (in this case as large as negative 1, as the amplitudes were constrained to be identical). Note how this negative 1 coefficient appears in Figure D.2 at lag-75, when the two series are at exactly opposite amplitudes in their 100-day period. This situation would also occur at lag -25.

We now examine a second toy example that is more specifically constrained to be representative of our periodicity problem. We set the period of the two periodicities to be 7 days, and sample once per day – this will represent the 1 cycle/week periodicity...
that is our current topic. The frequencies of the two cycles are constrained to be exactly 1 cycle/week, but we allow the amplitudes and phases to both differ. For ease of demonstration, we constrain the phase of the response to be 0 – hopefully it is obvious that this generalizes to arbitrary offsets without any real difficulty.

For this example, we shrink the variance of the response and its amplitude correspondingly, setting the amplitude to 1 and the variance to 0.1.

\[
\begin{align*}
\text{response: } y &= 1 \cdot \sin \left( \frac{2\pi t}{700} + 0 \right) + \xi_t \\
\text{predictor: } x &= 10 \cdot \sin \left( \frac{2\pi t}{700} + \phi \right) + \zeta_t, \\
\end{align*}
\]  

(D.3)

where $\xi_t \sim \mathcal{N}(0, 0.1)$ and $\zeta_t \sim \mathcal{N}(0, 1)$, and $\phi$ is an arbitrary phase offset between 0 and $360^\circ$; corresponding to a lag of 0 to 7 days. We allow $t = 1, 2, \ldots, 7000$, providing us with 10 cycles, each of period 700 samples (7 days), each sampled 100 times per
day. This extra precision will allow us to demonstrate the effect in the following. As a demonstration, we set $\phi = \frac{1.7}{7} \cdot 360^\circ = 87^\circ$. The series are exhibited in Figure D.3.

\[\text{Figure D.4: Coefficients resulting from regressing lagged copies of the predictor on the response from Eqn. D.3. Note how the maximum occurs at 1.7 days, as the model was constructed. Also note the red circles, which indicate the only coefficients we would actually have access to in a real-world situation, with samples taken once per day.}\]

We cheat slightly in the following demonstration: in the real example taken up at the end of the section, we do not have a finely divided sample set over our period of 7 days. For this example, however, we do, so we compute linear models at lags ranging from 0 to 699 samples, corresponding to 0 to 7 days. The results are shown in Figure D.4.

We close this section with a short note about the red circles shown in Figure D.4. These circles represent the only coefficients we would have access to in a real-world situation where we have samples taken once per day. If the true lag were indeed to be
1.7 days, we would see that the lag 1, 2 and 3 predictors were statistically significant in the linear model, and that the lag-4 through lag-6 predictors are inversely related to the response. We would not obtain the result that the maximum correlation occurs at a lag of 1.7 days, nor would we be able to directly access an estimate at that level due to a lack of resolution.

D.2 Estimation and Interpretation of Phase Offset

The actual computation of the phase offset between two line components (i.e. the harmonics of the week) from the mortality and the pollution is trivial: simply take the two phases and compute their difference, then scale by the period of the periodicity in order to obtain the offset in days. The complicated part of the estimation is in actually estimating the phase of each of the components\(^1\). Each series is sampled once per day, with a 7-day period, and we regress a sinusoidal model with appropriate period onto this series to obtain estimates of both the amplitude and phase. Once these estimates are obtained, we have perfect information about the line component, and we can easily obtain phase offsets from this.

If we consider the above toy examples where our response and predictor are both arbitrarily lagged sinusoids at a common frequency with additive noise, the interpretation of phase offsets becomes clear. The phase offset corresponds to a certain amount of period, as 360° is equivalent to a full period. For example, if our response and predictor are both cyclic with period 1 week, then a phase offset of 51° corresponds to a lag of approximately 1 day. We suggest that one of the reasons that there are statistically significant results when regressing different lag pollutants on mortality

\(^1\)In R, we obtain a line component series using routines included in the tsinterp package, with more details provided in Appendix B.
(as shown in Section 6.4.1) is that these lags are bringing major periodic components of the pollutant “into phase” with the corresponding periodic components of the mortality. This is explored in more detail in Chapter 6.5.

D.3 Real Data Example

In this example, we consider data from Toronto, CD 3520, between 1981 and 2007. As predictor we use Ozone due to its contiguous nature and strong periodicities, and as response we use All-Cause Mortality. From each of these series we remove the weekly periodicity, at frequency 1.653\(\mu\)Hz (period 7 days). In Table D.1, we list the various characteristics of these two periodicities.

<table>
<thead>
<tr>
<th></th>
<th>Mort.A</th>
<th>Ozone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amplitude</td>
<td>+0.2071</td>
<td>+2.1048</td>
</tr>
<tr>
<td>Phase (Degrees)</td>
<td>+87.13</td>
<td>−101.40</td>
</tr>
<tr>
<td>Phase (days)</td>
<td>+1.6942</td>
<td>−1.9717</td>
</tr>
<tr>
<td>Period (days)</td>
<td>+7.0000</td>
<td>+7.0000</td>
</tr>
</tbody>
</table>

Table D.1: Characteristics for weekly periodicities obtained from Mort.A and Ozone, Toronto. We note that the phase offsets may be (and are) different for differing census divisions. Also note that these offsets are only for the weekly (1 cycle/week) periodicity, not the entire series.

We note here that there is almost no significance to the phase measurements obtained for Mort.A and Ozone individually. All that really matters is their relative phase, i.e. the offset between the two of them. These measurements are entirely dependent upon where the time index was started from, which in our case was January 1, 1981. If we had started the time index on December 30, 1980, the phases would be different by exactly 2 days, or \(\approx 103^\circ\).

Again, as a demonstration we choose to cheat, generating two synthetic series with exactly the characteristics given here, excepting only that we set our synthetic
response series to have phase 0, and our synthetic predictor series to have phase $-101.4^\circ - 87.1^\circ = -188.5^\circ$, corresponding to $\approx -3.67$ days. This is equivalent to a phase offset (due to the cyclicity) of $+3.33$ days, or $+171.5^\circ$. The only difference between these two offsets is the interpretation: if the phase offset is negative, then mortality follows ozone, and a causal relationship is possible between then; if the phase offset is positive, then the converse is true. Since we can be confident that human mortality does not cause ozone, we choose the option that has ozone lagging occurring before mortality: the negative phase offset. It can be hard to tell with strong periodic terms – often we may be detecting the alignment between the current cycle of (say) ozone and the previous cycle of mortality, and thus see a positive phase offset. We interpret this in the physically sensible way.

We generate a single, simple model:

\[
\text{response: } y = 0.2071 \cdot \sin \left( \frac{2\pi t}{7} + 0 \right) \\
\text{predictor: } x = 2.1048 \cdot \sin \left( \frac{2\pi t}{7} - 188.5^\circ \right)
\] (D.4)

where for simulation purposes we set $t = 0.01, 0.02, 0.03, \ldots$. We generate these simulated series (with no noise), and fit varying synthetic lags using the GAM framework. The results are shown in Figure D.5. The vertical blue line is at $+3.33$ days, which corresponds to a lag of $3.33$ days, or a negative phase offset of $-188.5^\circ$. Note also that depending on where on the curve your chosen lag is, you may be at a correspondingly shallow or steep slope. Selecting lag-3 in Figure D.5 results in a very shallow slope and a coefficient estimate that is quite close to the maximum. But selecting lag-2 places your estimate on the steep slope of the sinusoid, and is a rapid decrease from the true maximum. This behaviour can help provide intuition for why some lagged coefficients appear to be quite similar, but others change rapidly.

Finally, we make a short note concerning the use of multiple lagged regressors in these epidemiological models. As we showed in Section 6.4.1, a typical procedure is to
try a number of models, varying only the lag of the primary pollutant predictor. As we have demonstrated above, in the case of a pure tone sinusoidal line component, this is effectively “exploring” the coefficient curve using coarse 1-day lag steps, and will not typically find the true maximum. It is, however, the only option when dealing with non-sinusoidal series, such as are typically encountered in these situations. Significant research has been done on the effects of including multiple lagged regressors of the same pollutant in a single model, e.g., [170, 87, 88, 235]. As we can infer from Figure D.5, doing so is likely to mis-estimate the relationship. Imagine that we were to include lags 3 and 4 of Ozone in a model that is predicting Mortality. The two are
obviously highly correlated, and neither is the true lagged regressor, so each contains a portion of the true relationship, along with a tremendous amount of correlation and confounded information. We believe that judicious use of this phase technique can help avoid some of these issues, allowing us to more precisely determine the lag relationship between the predictor and response, at least for line component elements of these series.