Novel Methods for Detecting Changepoints in Time Series Structures

Tessa Wilkie, M.A. (Cantab.), M.Sc., M.Res.

Submitted for the degree of Doctor of Philosophy at Lancaster University.

April 2025

STOR-i
excellence with impact

Abstract

This thesis makes three contributions. The first addresses the problem of finding a single changepoint in a time series of air quality measurements observed in and around Glasgow in the run up to, and aftermath of, measures introduced to improve air quality in the city. Since weather influences air quality and we have a time series of weather variables observed daily, we model this time series as a Vector Autoregressive process with exogenous variables (VAR-X). We adapt an existing changepoint method designed to detect a change in one or more series in an independent and identically distributed multivariate time series to finding a single change in a VAR or a VAR-X process.

Changepoint detection in the presence of missing data is relatively under-explored. Most existing methods assume that the data they are run on are complete. Some have explored replacing the missing values with imputed ones. The second contribution of this thesis is introducing a novel method for detecting changepoints in a linear regression model that is subject to missingness — in certain patterns — in the independent variables, without the need to impute separately.

The third contribution is the introduction of a novel likelihood ratio test based method for detecting a changepoint in a series of count data with trend. This is applied to interpolated count data: counts of infant mortality by year in selected local government districts in England between 1911 and 1973. We use the changepoint detection method to detect errors in the interpolation process, which manifest as abrupt changes, while allowing for historical trend which is present in all series.

Acknowledgements

I would like to thank, first and foremost, my supervisors: Distinguished Professors Idris Eckley and Paul Fearnhead of Lancaster University and Dr Jarno Hartog (Shell Research Ltd.) for their time, care and expertise throughout the undertaking of this work. As well as their input in to the work here, I am grateful for the encouragement they have given me on countless occasions.

Idris's wise counsel has helped me to grow as a researcher and his inspiring ideas and enthusiasm for changepoint detection applications has helped spark many of the directions that the work in this thesis has taken. A lot has been said, and will continue to be said, about Paul's scintillating statistics brain. I'm exceedingly thankful for his generosity in the time, thought and patience he has put into supervising me and into shaping this work. Many discussions with Jarno around the applications of changepoint detection helped to formulate the problems considered, and influence the approaches taken, in this thesis.

I would like to thank Edward Austin for his input into developing the work of Chapter 4, and for conversations about the work in Chapter 5. As well as his experience and expertise, his patience, enthusiasm and sense of humour has encouraged me enormously over the past few years.

The work in Chapter 5 would not have been possible without Ian Gregory, Distinguished Professor of Digital Humanities at Lancaster University, who generously shared his time, expertise and enthusiasm. His in-depth knowledge of the data and of the sub-

ject was instrumental in formulating the research question. He also contributed analysis of the findings in this chapter as well as helping to place them within the context of current historical scholarship.

I would also like to thank the examiners, for their time and thoughtful comments that have improved this thesis.

This thesis was completed while I was part of the EPSRC funded STOR-i Centre for Doctoral Training (EP/S022252/1). I am grateful to my fellow PhD students at STOR-i for their friendship and support, as well as Jen, Keilah, Kim, Nicky and Wendy in the administration team. Dr Daniel Grose helped me get set up to run large simulations on a super computer. I am enormously grateful to Distinguished Professor Jon Tawn, co-director of STOR-i, for his kindness, guidance and encouragement throughout my time at STOR-i.

I would like to acknowledge the financial support of Shell Research Ltd during my PhD. In addition to this, I'm very grateful for the friendly reception of Jarno and his colleagues on my occasions of visiting the Shell offices, and for their interest in and thoughtful comments on areas covered by this work.

The work on this thesis coincided with some difficult moments in my personal life. I would like to recognise the understanding, compassion and support of all three of my supervisors and the wider STOR-i community through this time.

Finally, I would like to thank my family, my friends, and my husband, David.

The findings, interpretations and conclusions expressed in this thesis are those of the author and do not necessarily reflect the views of Shell plc or any of its subsidiaries.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

The work in Chapter 3 is joint work by me and my supervisors: Idris Eckley, Paul Fearnhead and Jarno Hartog, while the work in Chapter 4 is by me, Edward Austin, Idris Eckley, Paul Fearnhead and Jarno Hartog.

The work in Chapter 5 is joint work by me, Idris Eckley, Paul Fearnhead and Ian Gregory. I am also grateful to Edward Austin for his input over many conversations on the subject of the work in this chapter.

This thesis contains 25611 words.

Tessa Wilkie

Contents

List of Figures

List of Tables

Chapter 1

Introduction

If we consider data that we observe over time, or in some other sort of ordered sequence, it is often of interest to know whether the characteristics of that data change at some point. This may tell us something valuable about the process that generates the data. For example, if we observe temperature in our home during cold weather and it dips suddenly, we may conclude that our boiler is broken or that we have left a window open. If we are looking to model the data, and there is a genuine change, we will need to update our model to reflect the difference in the data generating process before and after the change. Again, if we are modelling the temperature in our home, we would look to model periods when we have the heating on differently to when it is switched off. Detecting changes can be challenging. If our data has a simple structure and the change is obvious, it may be detectable by the human eye when plotted. However, it is not always clear whether an apparent change is genuine or down to random variation. This becomes even more difficult if we are dealing with complex data structures or high dimensions. The branch of statistics that deals with establishing the presence of a change, and its location, is known as changepoint detection.

Changepoint detection seeks to establish whether a statistical property of a data sequence changes one or more times; and, if so, where in the sequence this occurs. It

has a wide range of applications. The work in this thesis is motivated by the problem of changepoint detection in an industrial setting, though we will see the applications have taken some unexpected turns. A key application for companies that operate complex systems is condition monitoring: where identifying a change can flag a problem within the system, with the aim of allowing time to fix a fault before the fault causes the system to break. Another relevant application is pre-processing collected data ready for analysis: changepoint detection allows us to divide a data set where a system might be in different phases, by identifying the changes that signal the points at which it moves into a new phase.

This thesis addresses the problem of detecting a single changepoint in a data set that has already been collected — known as the offline setting. Our focus is on detecting changes in data sets where aspects of the data structure present challenges. While changepoint detection algorithms can be applied to any kind of ordered data, the focus of this thesis is on data that is observed over time.

In Chapter 3 we address the problem of detecting a change in a multivariate time series that is subject to a complex dependence structure: dependence over time — autocorrelation — and between series — cross-correlation.

In Chapter 4 we turn to considering changepoint detection in the presence of missing data. We introduce a novel method of finding changes in a linear regression with missing data in some of the independent variables. Finally in Chapter 5 we develop a changepoint detection method for a time series of counts that exhibits trend over time. The method is designed to detect abrupt changes, while ignoring changes in trend. It is applied to annual counts of infant mortality data in England and Wales in the 20th century. We then conclude in Chapter 6 with a discussion of directions for further research.

All computing done in this thesis was carried out in `R` (R Core Team, 2022). Plots are created using base `R` or `ggplot2` (Wickham et al., 2019).

Chapter 2

Literature Review

The work in this thesis addresses the problem of detecting a single change in a variety of time series structures: vector autoregressive (VAR) processes, linear regression models with missing data, and in a univariate sequence of count variables that exhibit trend. In this literature review we introduce key background concepts required for later chapters. Subsequent chapters contain more specific reviews of the literature pertinent to the contribution of the work therein.

We begin by introducing the changepoint problem, describing an established framework for detecting changes in univariate data in the offline setting. Initially we consider a change in the mean of a sequence of independent and identically distributed normal variables. We then turn to changes in the parameters of linear regression models, which is the focus of Chapter 4. We end the section with a discussion of the problem of finding abrupt changes in sequences where trend is present. In Section 2.2 we turn to review existing research which considers the detection of changepoints in multivariate data and follow with an introduction to VAR time series methods. We conclude this chapter by describing key concepts from the missing data literature and their application within the changepoint setting (Section 2.3). Literature introduced in these final sections will prove helpful in Chapters 3 and 4, respectively.

The methods in this thesis are based on penalised likelihood ratio tests, comparing the fit of a model with one or more changes to one with no change, with a penalty term to prevent overfitting. However, this is not the only approach that is popular in the literature: two others are Bayesian methods and Statistical Process Control. We give a brief overview of these areas here. More specific references to relevant parts of this literature are contained in the chapters that follow.

Bayesian methods for changepoint detection involve estimating the distributions of the number of changepoints and their location, as well as the distributional parameters for each segment of data between changepoints. This is done by deriving the posterior distributions of these unknown parameters, by specifying a prior and a likelihood (Eckley et al., 2011). Smith (1975) explores the case of a single change in mean in a sequence of observations — both in the case of normally distributed variables and those that have a binomial distribution. Carlin et al. (1992) and Stephens (1994) apply a Markov Chain Monte Carlo method — the Gibbs sampler — to the problem of estimating changepoints in a sequence of data, allowing numerical estimation of posteriors that are analytically intractable. For more details of Bayesian approaches to changepoint detection, see Eckley et al. (2011) and the references therein.

Detecting a change in the behaviour of a sequence of observations is also studied in Statistical Process Control, where methods include Shewhart control charts (Shewhart, 1926) and cumulative sum (Cusum) charts (Page, 1954). These charts are intended to monitor whether an industrial process, for example a manufacturing process, moves from an 'in-control' state to an 'out-of-control' state. They do this by calculating a statistic such as a grand average (used by, for example, Shewhart (1926)) and flagging where it crosses pre-determined limits. The use of the Cusum statistic for detecting the time or location of this departure, and its relation to the likelihood ratio test statistic, is explored by Hinkley (1971). We refer the reader to Qiu (2013) for a detailed overview of Statistical Process Control methods.

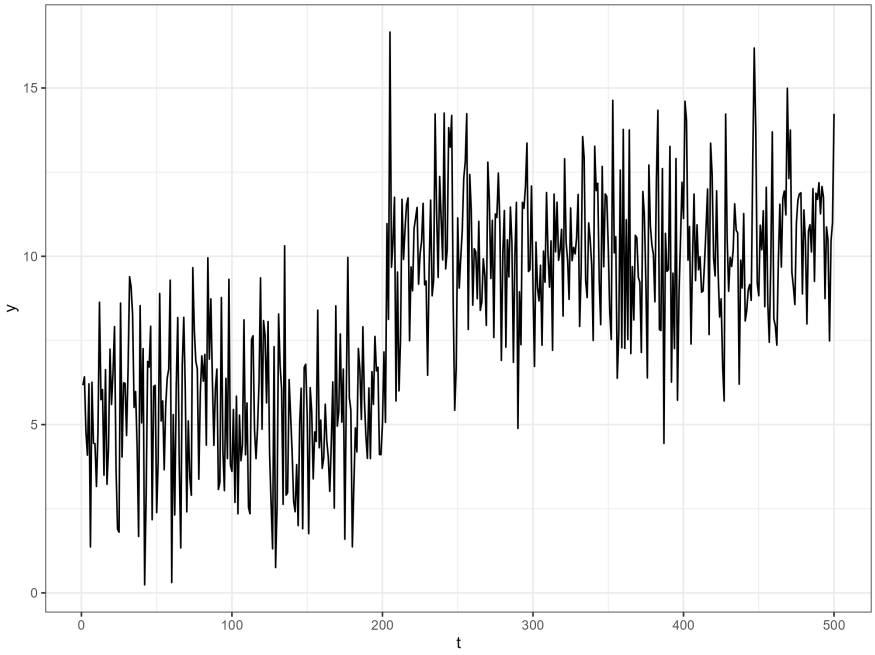


Figure 2.1.1: A change in the mean of a sequence of normally distributed variables. Changepoint at $t = 200$.

2.1 Introducing the changepoint problem: univariate case

Following the notation in Killick et al. (2012), let there be a series of sequential data of length n, denoted $y_{1:n}$, with a single changepoint, τ. The changepoint τ marks a change in a statistical property of the series. Some of the most commonly considered change types are changes in mean (Figure 2.1.1) or changes in variance (Figure 2.1.2). If, for example, we have a sequence of realisations of Gaussian data that are independently and identically distributed, and at some point τ in the sequence there is a change in mean (but the variance remains the same) then:

$$y_t \sim \begin{cases} N(\mu, \sigma^2), & \text{if } 1 \leq t \leq \tau, \\ N(\mu^*, \sigma^2) & \text{if } \tau + 1 \leq t \leq n. \end{cases}$$

To find a single changepoint, we define a cost function for a segment, which we denote

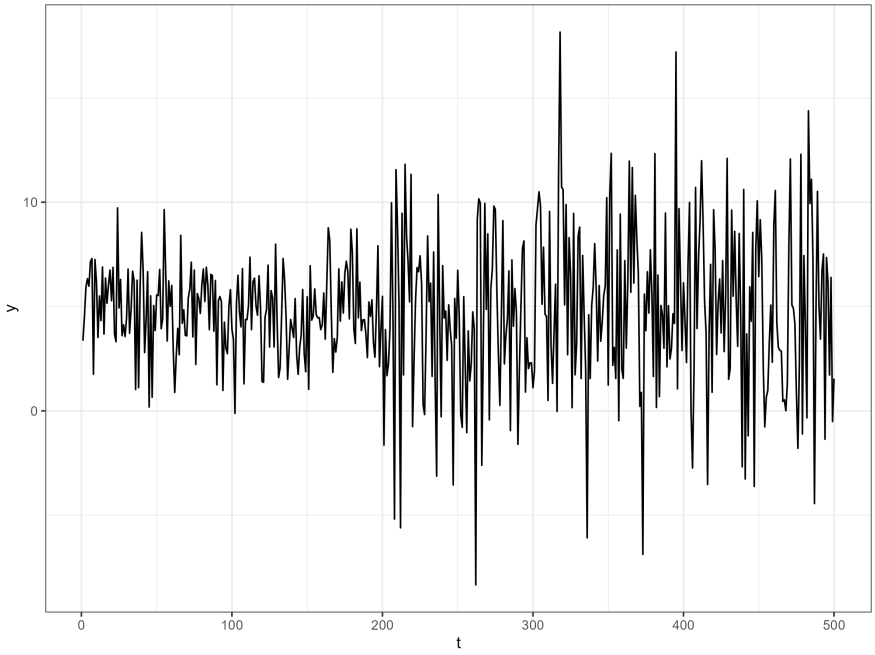


Figure 2.1.2: A change in the variance of a sequence of normally distributed variables. Changepoint at $t = 200$

$\mathcal{C}$. The cost for a segment containing observations $y_{s:t}$, where $s < t$, is then written $\mathcal{C}(y_{s:t})$. If we are using a parametric method, then this cost function should be minimised to find a good model approximation to the data for that segment. Specifically, we look to minimise the cost to find the best fitting segmentation of a series — indicating the most likely position of a changepoint (if one is indicated at all). The idea is that if we have a changepoint, then we will obtain the lowest costs by ending a segment at a changepoint, rather than having a segment with a changepoint in the middle. In the above situation, we would obtain the lowest cost by having two segments: one covering $y_{1:\tau}$ with μ_1 and one covering $y_{\tau+1:n}$ with μ_2.

For a single changepoint we look for a time, τ, that gives a segmentation with a lower cost than having no change (Killick et al., 2012, Equation 2):

$$\mathcal{C}(y_{1:\tau}) + \mathcal{C}(y_{\tau+1:n}) + \beta \leq \mathcal{C}(y_{1:n}).$$

In the above expression, β is a penalty function that controls the sensitivity of the

algorithm to identifying changepoints. When selecting β, a balance must be struck between identifying spurious changes and missing genuine changepoints. Under the null hypothesis, H_0, there is no change. Under H_1, there is a single changepoint at τ. Then the log-likelihood is maximised at $\log p(y_{1:n}|\hat{\theta})$ under H_0, while under H_1 the log likelihood reaches its maximum at the values of θ_1 and θ_2 that maximise the following:

$$\log p(y_{1:\tau}|\theta_1) + \log p(y_{\tau+1:n}|\theta_2).$$

Here θ represents the parameter/s of the series that are suspected of undergoing a change at τ, θ_1 is the parameter of interest before the changepoint, and θ_2 afterwards.

We can construct a test statistic, $\mathcal{L}$,

$$\mathcal{L} = 2\left[\max_{\theta_1}\left(\log p(y_{1:\tau}|\theta_1)\right) + \max_{\theta_2}\left(\log p(y_{\tau+1:n}|\theta_2)\right) - \max_{\theta}\left(\log p(y_{1:n}|\theta)\right)\right]$$

and we reject the null hypothesis that there is no change if $\mathcal{L} > c$, where c is a previously chosen threshold value (Eckley et al., 2011).

When searching for a single change, the penalised cost method is comparable to using a log likelihood ratio test: both compare the fit of a model with a changepoint to the fit of a model with no change, accepting a change if the difference in fit of the two models is sufficiently large (Eckley et al., 2011).

Finally, we note that the penalised cost method can be extended to searching for multiple changepoints. There are various approaches for searching for more than one change: approximate methods such as binary segmentation (Scott and Knott, 1974); or exact methods such as Pruned Exact Linear Time (PELT) (Killick et al., 2012). For a review of these methods, see, for example, Jandhyala et al. (2013) and Niu et al. (2016), and the references therein. The focus of this thesis is on approaches to detecting a single change, although we do extend the method introduced in Chapter 4 to search for multiple changepoints in one of the real data applications.

2.1.1 Changepoint detection in the presence of trend

Many methods exist that are suitable for finding changes in the mean of a sequence of observations (see Figure 2.1.1); and variance (see Figure 2.1.2). These methods assume that data is stationary between changepoints and do not perform well in the presence of trend, with trend provoking increased Type I errors. Existing methods that consider the problem of finding an abrupt change in the presence of data with trend are Romano et al. (2022) and Pein (2021). Neither of these methods is straightforwardly adaptable to count data with trend that is modelled with a Poisson distribution, such as the data we consider in Chapter 5.

Studies that do consider a change in Poisson data which is subject to trend do not, in the main, seek to find an abrupt change while allowing for trend. For example, Habibi (2021) detects a change in the mean and slope parameters of a Poisson regression. Additionally, Perry et al. (2006) and Assareh et al. (2013) aim to detect a changepoint in a Poisson process that is due to the introduction of trend. The only paper we are aware of that seeks to detect an abrupt change in Poisson data subject to trend — while not looking to detect a change in that trend — addresses Poisson processes, suited for inter-arrival times, not counts (Loader, 1992).

2.1.2 Changepoint detection in linear regression models

Having considered the problem of changepoint detection in a sequence of independent and identically distributed data earlier in this section, we turn to an extension to the univariate case that we consider in more depth in Chapter 4 of this thesis. This is the linear regression model, where there is a univariate dependent variable whose behaviour can be predicted by one or more independent variables. We follow the notation of Bai and Perron (1998). For simplicity, we consider the case of a single changepoint in one

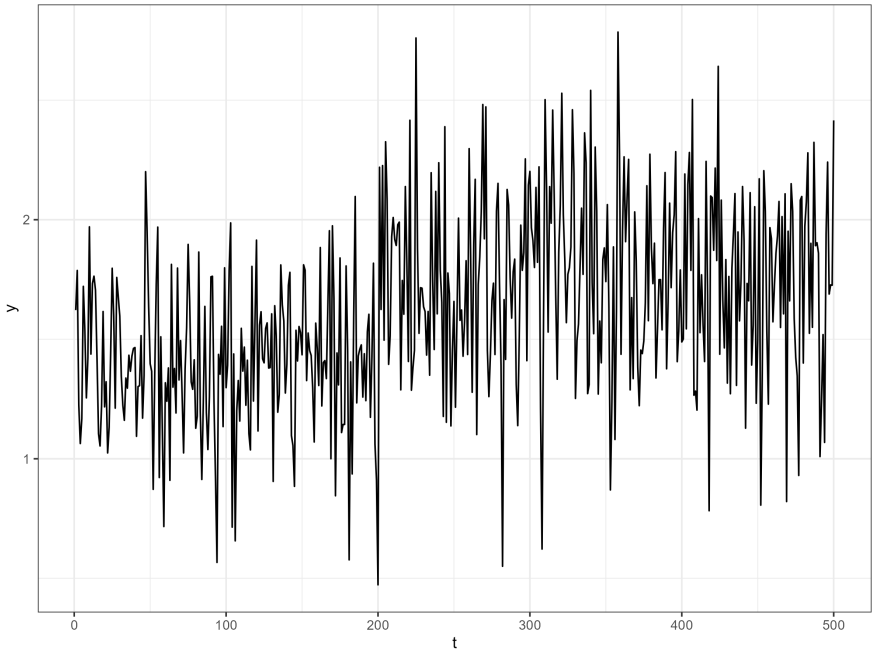


Figure 2.1.3: A change in the parameters of a linear regression model: independent variable plotted against time. Changepoint at $t = 200$

of the regression coefficients of a linear model:

$$y_t = x_t\beta + z_t\delta + \epsilon_t,$$

where $t = 1, \ldots, n$, $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{z}$ and ϵ are $n \times 1$ vectors, and y_t, x_t, z_t and ϵ_t are their values at time t. The coefficients of the regression, β and δ, are scalars. In this setting we consider t to indicate evenly spaced points in time, since the applications of this work are time series, but in other models t could be the index for any ordered sequence of observations. In this formulation we consider a change in the coefficient related to $\boldsymbol{z}$, δ, while the coefficient related to $\boldsymbol{x}$, β, does not change. This can be written as:

$$y_t = \begin{cases} x_t\beta + z_t\delta + \epsilon_t & \text{if } \ t \leq \tau \ , \\ x_t\beta + z_t\delta^* + \epsilon_t & \text{if } \ t > \tau \ , \end{cases}$$

where δ^* is a scalar coefficient and $\delta \neq \delta^*$. Bai and Perron (1998) consider the problem

of finding multiple changes in the parameters of a linear regression model, using least squares to estimate the parameters of the regression between changepoints, and to determine the optimal placement of changepoints. For a single change, with possible candidate changepoint locations $\tau = 1, \ldots, n$, we would use least squares to estimate $\hat{\beta}$, $\hat{\delta}$, $\hat{\delta}^*$, for all possible locations of $t = \tau$. By minimising the sum of the squares of residuals by solving the below equation, we can find a candidate changepoint $\hat{\tau}$:

$$\mathcal{S}_\tau = \min_{\tau=1,\ldots,n-1,\beta,\delta,\delta^*} \left(\sum_{1}^{\tau} [y_t - x_t\beta - z_t\delta] + \sum_{\tau+1}^{n} [y_t - x_t\beta - z_t\delta^*] \right),$$

where $\hat{\tau} = \arg\min_\tau \mathcal{S}_\tau$. Bai and Perron (1998) give test statistics and critical values for testing the hypothesis of no changepoint versus a specified number of changepoints, and for no change versus an unknown (but bounded) number of changes.

In Chapter 4 we will consider the detection of a single changepoint in the parameters of a linear regression model with the additional consideration that some of the dependent variables are subject to missingness. We will briefly describe key concepts in missing data in Section 2.3.

2.2 Changepoint methods for multivariate data

In this section we provide an overview of changepoint detection methods for the multivariate setting, including key challenges and recent developments, before describing the VAR process which is the focus of Chapter 3. Arguably the simplest and most studied setting is to consider a change in a multivariate time series of independently and identically distributed normal variables. This can then be extended to consider the multivariate setting, where there is a relationship between the variables. We refer the reader to Truong et al. (2020), for example, for a review of offline detection methods for multivariate series.

Multivariate data pose additional challenges for changepoint detection methods.

For example, not necessarily every variable in a multivariate series will experience a change. Moreover, among those that do experience a change, this need not necessarily occur at the same time. In addition to this, an eye must be kept on computational complexity, as the computational demands of analysing high dimensional data sets can quickly become troublesome. We can, of course, apply a univariate method to each variable of a multivariate series, such as the multivariate implementation of PELT (Killick et al., 2012). This approach loses the statistical power that comes with combining information across series: it is not set up to detect a common signal at a single time point across several series — reflecting the situation where a single changepoint affects some or all of a multivariate series (Bardwell et al., 2019). They also cannot account for correlation between variables. Many methods developed over the past 10 years for tackling multivariate data look at various ways to aggregate information across series. See Bardwell et al. (2019) and Tickle et al. (2021) for further discussion. The multivariate changepoint methods mentioned so far assume independence between series though some recent work allows for dependence across series: see for example, Tveten et al. (2022) and Wang and Samworth (2018).

The remainder of this section provides the background necessary to consider changepoints in VAR settings. We focus on the fundamentals of VAR time series models before reviewing recent changepoint contributions.

2.2.1 Vector Autoregressive processes

We now turn to discuss Vector Autoregressive (VAR) and Vector Autoregressive models with exogenous variables (VAR-X) models, which form the focus of the work in Chapter 3. A VAR process is very flexible: allowing for both autoregressive behaviour, where a variable in a multivariate time series is influenced by the past behaviour of that variable, and cross-correlation, where the behaviour of a variable is influenced by the past behaviour of other variables in a multivariate time series. For simplicity of notation,

we will assume throughout this section that processes have mean zero. Were the mean not zero, we would need to add a constant term to the least squares estimator described below, or re-adjust the observations to centre them.

2.2.2 VAR models and estimation

We first introduce autoregressive (AR) models (Box and Jenkins, 1970), since a VAR model is a multivariate extension of an AR model. An autoregressive process is defined as modelling the behaviour of the time series at time t as a linear aggregation of its behaviour at previous time lags, up to p, plus an error term (Box and Jenkins, 1970, Section 1.2):

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t.$$

Here y_t is a time series observed at time t, ϕ_i, $i = 1, \ldots, p$ are scalar coefficients representing the influence of the $t - i$th observation on y_t, and ϵ_t is white noise.

Moving to the multivariate setting, we introduce a time series of length n and dimension d, denoted $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{d,t})^T$ where $t = 1, \ldots, n$. We follow the notation of Tsay (2013) for VAR (p) models:

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t,$$

where each $\boldsymbol{\phi}_i$ matrix is a $d \times d$ matrix containing the coefficients of the VAR process corresponding to lag i. These coefficients describe the impact of the the time series' behaviour at time $t - i$ on its behaviour at time t. $\boldsymbol{\epsilon}_t$ is multivariate white noise. We assume the error terms have mean zero and their covariance matrix, which we denote $\boldsymbol{\Sigma}$, is positive definite (Tsay, 2013).

We give an overview of the least squares method for estimating the parameters of a VAR process as described in Tsay (2013, Section 2.5.1). We refer the reader to this

text for more detail. If we have a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from a d-dimensional process, then we can form the following equation with our data:

$$\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{\beta}^T + \boldsymbol{E},$$

where $\boldsymbol{Y}$ is an $n - p \times d$ matrix

$$\boldsymbol{Y} = \begin{pmatrix} y_{p+1,1} & \cdots & y_{p+1,d} \\ \vdots & \cdots & \vdots \\ y_{n,1} & \cdots & y_{n,d}, \end{pmatrix}$$

where $\boldsymbol{Z}^T$ is an $(n-p) \times dp$ vector of lagged observations of $\boldsymbol{y}_t$,

$$\boldsymbol{Z}^T = \begin{pmatrix} y_{p,1} & \cdots & y_{p,d} & \cdots & y_{1,1} \cdots & y_{1,d} \\ \vdots & & & & & \vdots \\ y_{n-1,1} & \cdots & y_{n-1,d} & \cdots & y_{n-p,1} & \cdots y_{n-p,d}, \end{pmatrix}.$$

and $\boldsymbol{\beta}^T = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p)^T$ is a $dp \times d$ matrix and $\boldsymbol{E}$ is an $n \times d$ matrix of error terms that are multivariate white noise. Tsay (2013) shows that the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(\boldsymbol{Z}^T\boldsymbol{Y}),$$

and the estimate of the covariance matrix of the vectors of error terms, $\boldsymbol{\Sigma}$, is estimated from the residuals,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - (d+1)p} \sum_{t=1}^{n} \boldsymbol{r}_t \boldsymbol{r}_t^T,$$

the residual vector at time t being found by subtracting the predicted values from the

observed:

$$\boldsymbol{r}_t = \boldsymbol{y}_t - \sum_{i=1}^{p} \hat{\boldsymbol{\phi}}_i \boldsymbol{y}_{t-i}.$$

2.2.3 VAR processes with exogenous variables

An extension of the VAR process is the VAR-X model (Engle et al., 1983), where a VAR process is influenced by one or more time series of exogenous variables. The relationship is one way — the exogenous variables influence the VAR process but the VAR process cannot influence the explanatory variables. The behaviour of the exogenous variable is assumed to be known. A VAR-X model can allow for past behaviour of the exogenous variable to influence the VAR process at time t. Additionally, while we have assumed that the processes in this section have mean zero, we can allow for non-zero means, and potential changes in the mean, with the VAR-X setting by having the intercept as one of the exogenous variables.

If we have m exogenous variables, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, then these can be incorporated. A VAR-X (p, s) model does this by assuming that the values of the exogenous variables at lags 1 to s can affect the current observation. If we have a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from a d-dimensional VAR-X (p, s) process, with zero mean, then we can form the following equations with our data:

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \sum_{j=0}^{s} \boldsymbol{\pi}_j \boldsymbol{x}_{t-j} + \boldsymbol{\epsilon}_t,$$

where again $\boldsymbol{y}_t$ is a $d \times 1$ vector of observations at time t, $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{d,t})^T$ and $\boldsymbol{x}_t$ is a $m \times 1$ vector of the exogenous variables observed at time t and $\boldsymbol{e}_t$ is a d-dimensional vector of errors. Again, the $\boldsymbol{\phi}_i$, $i = 1, \ldots, p$ are $d \times d$ matrices of coefficients representing the relationship between the $\boldsymbol{y}_t$ at time t and at lags $1, \ldots, p$. The $\boldsymbol{\pi}_j$ are $d \times m$ matrices of coefficients representing the relationship between the exogenous variables $\boldsymbol{x}$ at time

t to $t-s$ and $\boldsymbol{y}_t$.

This can be written in the same matrix form as (2.2.2):

$$\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{\beta}^T + \boldsymbol{E},$$

where

$$\boldsymbol{Z}^T = \begin{pmatrix} y_{0,1} & \cdots & y_{0,d} & \cdots & y_{1-p} & \cdots & y_{1-p,d} & x_{1,1} & \cdots & x_{1,m} & \cdots & x_{1-s,1} & \cdots & x_{1-s,m} \\ \vdots & & & & & & & & & & & & & \vdots \\ y_{n-1,1} & \cdots & y_{n-1,d} & \cdots & y_{n-p,1} & \cdots & y_{n-p,d} & x_{n,1} & \cdots & x_{n,m} & \cdots & x_{n-s,1} & \cdots & x_{n-s,m} \end{pmatrix},$$

an $(n-p)\times(dp+ms)$ matrix and $\boldsymbol{\beta}^T = \begin{pmatrix} \boldsymbol{\phi}_1 & \cdots & \boldsymbol{\pi}_p & \boldsymbol{\pi}_1 & \cdots & \boldsymbol{\pi}_s \end{pmatrix}^T$, a $(dp+ms)\times d$ matrix of the VAR-X coefficients. The coefficients of this model can be estimated by least squares in the same way as above.

2.2.4 Stability and stationarity conditions

The work presented in Chapter 3 focuses on VAR processes that are stable and weakly stationary between changepoints. A non-stationary process has a mean, variance or covariance that changes over time. A process is said to be weakly stationary if its mean, variance and covariance processes are not dependent on time.

A process is defined as stable if the roots of the characteristic polynomial,

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p = 0,$$

lie strictly outside the unit circle (Box and Jenkins, 1970, Section 3.2). To extend this

to the multivariate setting, Lütkepohl (2013) describe a d-dimensional VAR(p) process

$$\boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t$$

defined through the matrix

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \dots & \boldsymbol{\phi}_{p-1} & \boldsymbol{\phi}_p \\ \boldsymbol{I} & \boldsymbol{0} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \dots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{I} & \boldsymbol{0} \end{pmatrix},$$

and where $\boldsymbol{\epsilon}_t$ is multivariate white noise, give the result that a VAR(p) process is stable if the roots of the reverse characteristic polynomial are greater than 1 in modulus:

$$\det(I_d - \boldsymbol{\phi}_1 z - \ldots - \boldsymbol{\phi}_p z^p) \neq 0 \quad \text{for} \quad |z| \leq 1$$

Lütkepohl (2013) gives the result that a stable VAR process, $\boldsymbol{y}_t$, defined for all $t \in \mathbb{Z}$ is also stationary.

2.2.5 Changepoint detection in VAR processes

Finding changes in VAR models represents an extension from the methods in the multivariate setting mentioned earlier in this chapter where we assume independence over time and between series. Examples of earlier work that consider changes in VAR models include Kirch et al. (2015), Wang et al. (2019), Safikhani and Shojaie (2022) and Cho et al. (2024). Additionally, Cho and Fryzlewicz (2015) introduce a wavelet-based method for detecting changes in the second order structure of a multivariate time series with dependence on time and between series. The complexity of VAR models

means that changepoint detection methods aiming to find parameter changes quickly become very computationally complex. Many changepoint detection methods for VAR processes incorporate measures to reduce the computational complexity. For example, Wang et al. (2019) incorporate LASSO regression into the change detection statistic, and Cho et al. (2024) apply a factor model to account for some of the autocorrelation and cross-correlation in the series, allowing for the rest to be modelled as a VAR process.

2.3 Missing Data

In this section we introduce some key concepts in the missing data literature that are relevant to the problem and method introduced in Chapter 4. We direct the reader to Little and Rubin (2019) and Van Buuren (2018) for in-depth treatments of missing data.

2.3.1 Missingness Mechanisms

Missing data occur in a wide variety of applications, yet there are few papers that explore the impact of missing data on changepoint problems, or that introduce methods that handle missing data within the changepoint problem. An important aspect of missing data is the nature of the probabilistic process that induces the missingness. This is known as the missingness mechanism (Rubin, 1976). The missingness mechanism influences which methods are appropriate for handling the missing data. If an inappropriate method is used to handle the missing data given its missingness mechanism, estimates from the data set can be severely biased.

We follow Little and Rubin (2019) in our discussion of missingness mechanisms, and refer the reader to this text for more detail. We first consider $\boldsymbol{X}$, an $n \times p$ data set of observations. This matrix has no missing observations. We also introduce a response

indicator matrix $\boldsymbol{R}$, where the i, jth entry is 1 if the i, jth entry of $\boldsymbol{X}$ is observed, and 0 if it is missing. We refer to the ith row of $\boldsymbol{X}$, $\boldsymbol{R}$ as a case, the jth column as a variable, and the i, jth entry as an observation.

We introduce a process $\boldsymbol{\psi}$, the missingness mechanism: this is the distribution of missing data given the 'true' underlying data, $\boldsymbol{X}$. This is what creates the missingness in the data, and what information is available to us about it. Missingness mechanisms fall broadly into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). If the data are Missing Completely at Random then the probability distribution of $\boldsymbol{R}$ is not dependent on $\boldsymbol{X}$, and the probability of $x_{i,j}$ being missing is evenly distributed across all of the entries of $\boldsymbol{X}$:

$$f_{\boldsymbol{R}|\boldsymbol{Y}}(r|x, \boldsymbol{\psi}) = f_{\boldsymbol{R}|\boldsymbol{Y}}(r|\boldsymbol{\psi}).$$

As an example, consider data collected from a thermometer measuring air temperature where some of the data is missing. If the thermometer fails to record some observations in a random fashion — that is, the probability of it failing to record an observation is the same for all observations and does not depend on any other variable, then we can say the data is MCAR.

In order to describe MAR and MNAR data, we introduce notation to denote the elements of $\boldsymbol{X}$ that are observed, and those that are missing: $\boldsymbol{X} = \{\boldsymbol{X}^{\text{obs}}, \boldsymbol{X}^{\text{mis}}\}$. For data that is MAR, the missingness mechanism depends on $\boldsymbol{X}$, but it is on values of $\boldsymbol{X}$ that are not missing. For our thermometer example, if the data are more likely to be missing the higher the value of another variable is — for example if the thermometer works less well in humid conditions — and we have records of humidity, then the data

is MAR. A wide range of methods perform well on MAR data.

$$f_{\boldsymbol{R}|\boldsymbol{Y}}(r|x^{\text{obs}}, x^{\text{mis}}, \boldsymbol{\psi}) = f_{\boldsymbol{R}|\boldsymbol{Y}}(r|x^{\text{obs}}, \boldsymbol{\psi}),$$

The most difficult missingness mechanism to handle is MNAR. This is when the missingness mechanism itself depends on some data that is not observed, (2.3.1) does not hold. As an example, if our thermometer data is more likely to be missing when the air temperature is higher, then the missingness mechanism would be MNAR. We cannot infer much about the missing data as the data we would need in order to do this is, itself, missing.

2.3.2 Missingness Patterns

Another important aspect to missing data is the pattern in which the data appears. This is often influenced by the way the data has been collected. Three common patterns are:

1. **Univariate missingness**: where only one of p variables contains missing data. For example,

$$\boldsymbol{R} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.3.1}$$

2. **Monotone missingness**: where the columns of $\boldsymbol{R}_{i,j}$ can be ordered in such a way that for the ith case, if $x_{i,j}$ is observed, then $\boldsymbol{x}_{i,j+1}$ is also observed for all $j = 1, \ldots, k-1$, for some k. Then an example of $\boldsymbol{R}$ with a monotone missingness

pattern is:

$$\boldsymbol{R} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

3. **General missingness**: where the missingness pattern is random across cases and variables. For example,

$$\boldsymbol{R} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Other missingness patterns exist, but fall outside the scope of this thesis. We direct the reader to Little and Rubin (2019) for a more detailed discussion.

2.3.3 Common approaches to dealing with missing data

The missingness mechanism and the missingness pattern influence what methods are appropriate for dealing with missing data. An inappropriate choice of method can severely bias an analysis. In this section we give an overview of common methods of dealing with missing data. Little and Rubin (2019) define four broad groupings of methods: analysis only of the cases where no data is missing; weighting to re-balance the

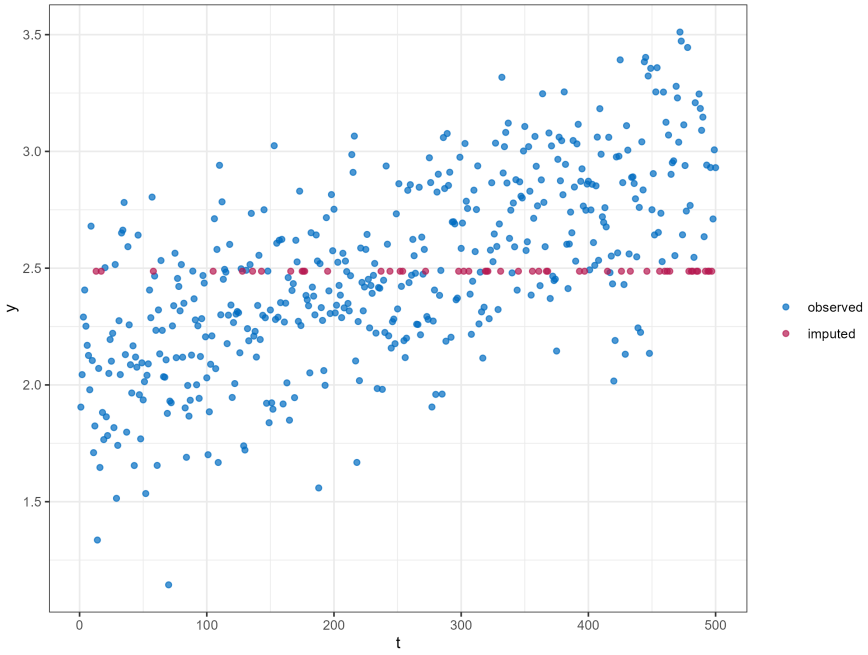


Figure 2.3.1: Linear regression with 10% of independent variable missing in a Missing at Random pattern and imputed using unconditional mean imputation. Plot created using `ggmice` package (Oberman et al., 2023).

analysis of the data given missing observations; replacing missing values with estimates (known as imputation); using models that can handle the presence of missing data. More detailed descriptions of common techniques, their advantages and drawbacks are available in Little and Rubin (2019) and Van Buuren (2018).

In this thesis we are primarily concerned with considering missing data among the independent variables of linear regression models, and direct the reader to Little (1992) for a detailed discussion of this. There are broadly three approaches to dealing with missing data in the linear regression models covered by Little (1992). These are: selectively deleting data (Complete Case Analysis, Available Case Analysis); single imputation based on least squares or maximum likelihood methods; iterative methods (EM algorithm, multiple imputation).

The simplest way of handling missing data is Complete Case Analysis (CCA), where any case — or row —of $\boldsymbol{X}$ that contains any missing observations is deleted. Available Case Analysis uses all of the available observations to calculate estimates from the data.

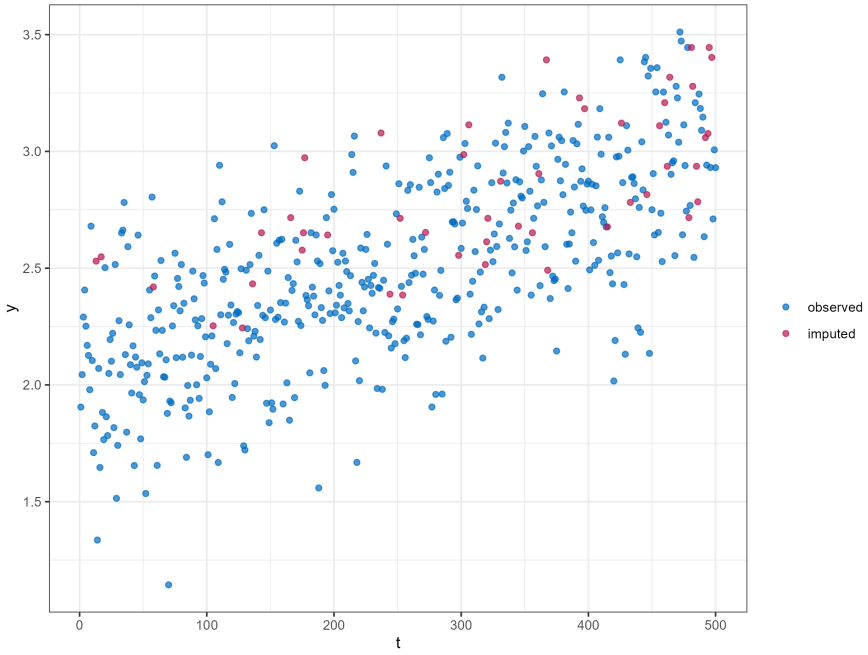


Figure 2.3.2: Linear regression with 10% of independent variable missing in a Missing at Random pattern and imputed using unconditional mean imputation. Plot created using `ggmice` package (Oberman et al., 2023).

In regression this would mean using all observed cases of variable X_j to calculate its mean and variance, and using all the cases where X_j and X_k are observed to calculate the covariance between the two. Both methods are computationally and conceptually straightforward. But they have drawbacks. Both are only suitable for data that is MCAR — otherwise they can yield estimates that are severely biased. Complete Case Analysis in particular is inefficient: it involves potentially throwing away a lot of information. Additionally, in linear regression models, available case analysis can yield a non-invertible variance-covariance matrix and can perform much worse than CCA when data is highly correlated (Little, 1992).

Imputation methods involve replacing missing observations with estimates. Some of the simplest methods involve replacing missing observations with the mean, or with the predicted value based on regression, with or without a random error term. We refer the reader to Van Buuren (2018) and Little and Rubin (2019) for a more comprehensive treatment.

A class of model-based methods is maximum likelihood-based approaches. These do not explitly impute, but look to maximise a likelihood that allows for missing data. Little (1992) finds that ML methods outperform those that impute the data and then estimate parameters using least squares, although a good choice of weights can improve performance of the latter.

A class of maximum likelihood methods suited to univariate or monotone missingness patterns, is factoring the likelihood. It is introduced by Anderson (1957) and applied to regression models by Gourieroux and Monfort (1981). This breaks up the likelihood into completely observed parts: one of the complete cases, then of the fully observed cases for the variable/s that is subject to missingness, conditional on the complete variables for this set of cases. If the data is monotone in missingness, then the likelihood can be split further. This method has the advantage of maximum likelihood based methods, and it is much simpler computationally than iterative methods for dealing with missing data, such as the Expectation-Maximisation algorithnm or multiple imputation (described below). Factoring the likelihood is the basis for the method developed in Chapter 4, where we introduce a method for detecting changepoints in linear regression models that are subject to univariate or monotone missingness in the covariates.

When we do not have data in such a missingness pattern, then likelihood-based methods are iterative, cycling between estimating the missing values then using these estimates to maximise the likelihood and estimate parameters. An example is the Expectation-Maximisation algorithm (Dempster et al., 1977). Iterative methods are more computationally complex than single imputation or factoring the likelihood.

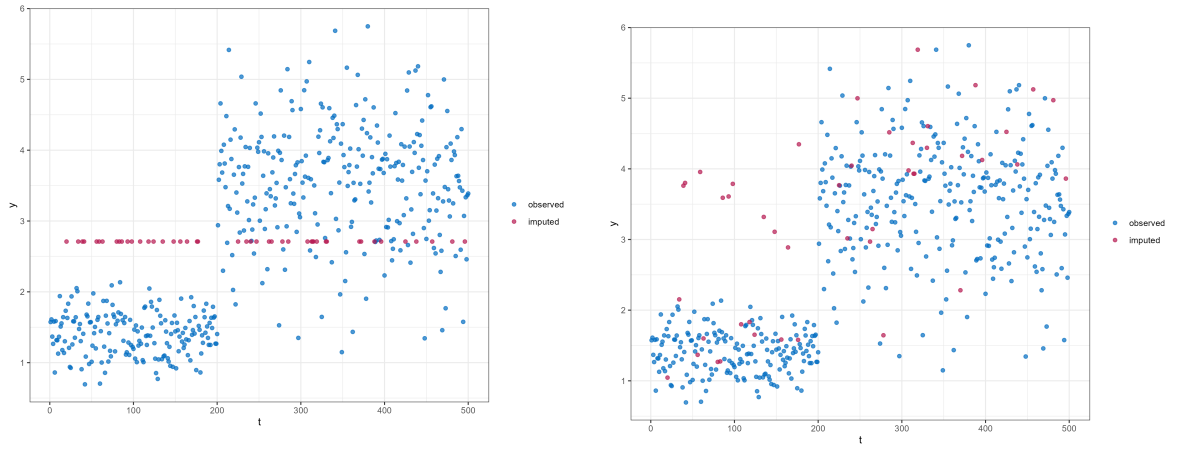
The drawback of single imputation methods is that it is difficult to account for the uncertainty that we introduce to our sample that is down to imputing. We have a set of replacement values, but we treat our imputed data set as if it is fully observed. Multiple imputation (Rubin, 1978, 2004) directly allows estimation of the uncertainty

introduced by imputing the missing data. These construct a predictive distribution for the missing data conditional on the observed data and estimated parameters. Multiple imputation draws from this distribution in order to impute missing observations. This is performed several times, and the data is analysed as if it were a complete data set after each time. The estimates derived from these analyses are then pooled. This allows calculation of the between imputation variance, as well as the average variance found in the analyses, giving understanding of how much variance in the estimates is down to the imputation process.

Through this thesis we will use single and multiple imputation (Chapters 3 and 4) using the `mice` package in `R` (Van Buuren and Groothuis-Oudshoorn, 2011), either to handle missing data as part of our analysis (Chapter 3) or to compare it with a method that we have introduced to handle missing data in regression (Chapter 4). When we do impute, we use the default in the `mice` package, which is known as predictive mean matching. This derives a predicted value for each missing value, by constructing a predictive distribution from the conditional distribution of the missing value given the observed data and the latest round of imputation estimates for other missing data. It then selects an observed value at random from a set of 'donor' observed values that are close to the predicted value.

2.3.4 Changepoints with missing data

Figure 2.3.4 illustrates some of the pitfalls of dealing with missing data when there is a changepoint an an unknown location. The modelling of the data used to predict imputed values can be complicated by the presence of a changepoint. In Figure 2.3.4, we see that predictive mean matching is clearly superior to unconditional mean imputation for imputing missing data in the presence of a changepoint. Predictive mean matching, however, still imputes some missing observations before the changepoint using the data model after the changepoint. This is because predictive mean matching predicts a



(a) Imputed using unconditional mean imputation

(b) Imputed using predictive mean matching

Figure 2.3.3: Linear regression model with a change in the parameters at $t = 200$. Independent variable plotted against time. Plots produced using `ggmice` (Oberman et al., 2023).

value for the missing data, then selects the imputed value from a set of observed data that is nearest this predicted value. In this example it appears that some of the donor observations for imputations before the changepoint are observations from after the changepoint. Work that deals with missing data explicitly within changepoint analyses either impute the data, then perform changepoint analysis, or incorporate the treatment of missing data into the changepoint estimation problem. For example, in the first instance, Bayesian methods, like that described in Corradin et al. (2022), incorporate an imputation step into an iterative process of modelling the location of changepoints, given the estimated model parameters between changepoints. In the second instance, an explicit imputation step is not needed, and the changepoint detection statistic is adapted to handle missing observations. An example of this is Follain et al. (2022), but this method is not suited to regression models. In Chapter 4 we introduce a method specific to detecting changes in a linear regression model with missing data in the covariates.

Chapter 3

Detecting a change in the structure of a multivariate time series exhibiting dependence on time and across variates

3.1 Introduction

Changepoint detection has been used with a wide variety of applications since it was introduced in the latter half of the 20th century (Page, 1954, 1955; Scott and Knott, 1974). In the last 20 years, increasing attention has been paid to the question of detecting multiple changepoints in a univariate series (see, for example, Killick et al. (2012) and Fryzlewicz (2014)); and to detecting multiple changepoints in multivariate datasets. Recent examples include Cho and Fryzlewicz (2015), Bardwell et al. (2019) and Tickle et al. (2021). Detecting changes is important across many application areas including finance (Aue et al., 2009; Cho and Fryzlewicz, 2015), health and medicine (Kirch et al., 2015; Avanesov and Buzun, 2018; Dehning et al., 2020) and communications data

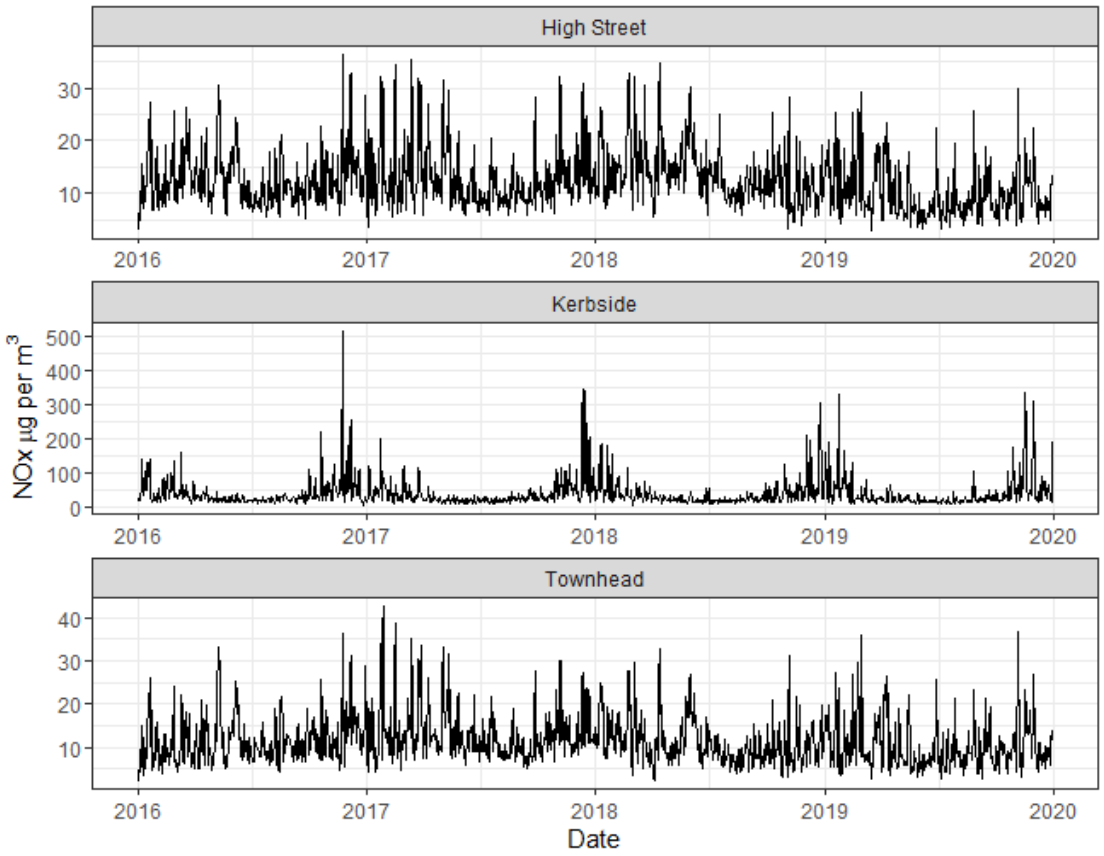


Figure 3.1.1: Nitrous Oxides as Nitrogen Dioxide levels observed across selected sites 2016 – 2019

(Verbesselt et al., 2010; Bardwell et al., 2019), among many others.

In this chapter we consider detecting changes in multivariate data with dependencies across series and time. As a motivating example data from air quality sensors that measure the level of a certain pollutant — Nitrous Oxides as Nitrogen Dioxide (NOx) — in the atmosphere at sites across the Glasgow area (Air Quality in Scotland, 2024). This data forms a multivariate time series, which is depicted in Figure 3.1.1. At the end of 2018, Glasgow introduced stricter emissions standards for buses operating in the city. We explore whether the introduction of this policy coincided with a shift in air quality.

The air quality sensor data exhibits spatial dependence as well as dependence on time. The presence of cross-correlation and/or auto-correlation can lead to excessive

noise in the residuals of a model that does not account for it. Many existing changepoint detection techniques for multivariate time series assume that there is no correlation either between series or across time and are vulnerable to identifying spurious changepoints when applied to data with these features. They are also unable to identify changes in cross-correlation, which can be of interest in itself. Additionally, there are multiple seasonal patterns and exogenous variables — predominantly local meteorology but also pollution from outside Glasgow plus extraneous events (Department for Environment Food and Rural Affairs, 2017, 2018, 2019, 2020). The series of pollution measures are also subject to missing data.

Our approach to detecting if the policy has had an effect is to look to detect changes in the second order structure of the data between January 1, 2018 and December 31, 2019. As suggested by, for example, Friede et al. (2006), if the intervention has had an impact then we would expect there to be a change around the intervention date. An alternative method, when the time or date of the intervention is known, is to use Interrupted Time Series analysis (see, for example, McDowall et al. (2019) for more detail), to assess the impact of the intervention. However, Friede and Henderson (2003) and Friede et al. (2006) find that this method is inferior to changepoint detection since it is vulnerable to false positives from changes in a time series that are not due to the intervention under investigation. Additionally for this application, while we know the date of the introduction of the ULEZ, detecting the impact of this policy intervention is complicated further as it does not happen in isolation, and was telegraphed to road users well in advance of implementation.

We cannot, however, simply look for a change in mean levels of air pollution in Glasgow. The causes of air pollution in Glasgow are myriad, complex and interlinked. Road traffic pollution is just one cause. For example, Glasgow's LEZ was announced in October 2017, but not introduced until the end of 2018. Some bus fleets were upgraded in advance of the change (see, for example, Intelligent Transport (2018)). It

is evident in Figure 3.1.1, that pollution levels appear to be slowly declining over the period of interest. In terms of modelling air pollution levels in order to detect a change, we can treat causes and infuences on air pollution as latent factors — many of them affecting pollution in all areas of Glasgow — that cause a certain dependence structure in the data. We expect the introduction of the LEZ, if it has the intended impact of changing road traffic patterns, to alter the relationship between sites inside the LEZ and these extraneous factors. In Section 3.5 we look to model air quality data in the LEZ together with selected extraneous factors as a multivariate time series, and look to detect a change in the parameters of this model.

Established changepoint detection methods have focused on changes in mean in uncorrelated univariate data — for example, Killick et al. (2012), Jackson et al. (2005), Fryzlewicz (2014). In the past decade there has been an interest in extending methods to multivariate data, other types of changes, and to allow for correlation. For example, Tickle et al. (2021) and Wang and Samworth (2018) focus on searching for a change in mean in multivariate data. Further considerations come with evaluating changes in multivariate series. One challenge is developing methods that can detect both dense changes, where a change occurs in all or nearly all of the variables in a multivariate time series; and sparse changes, where a change occurs in just a few of the variables. See, for example, Enikeeva and Harchaoui (2019).

When dealing with multivariate series, it is possible to simply add the evidence for a change across all series and use this sum as a test statistic. However, the threshold for such a test will need to be large due to the variance of the test statistic being the sum of the variances from the variate. The signal from a change that only affects one or a small number of variates may be small relative to the threshold for the test. Sparsified Binary Segmentation (SBS), introduced by Cho and Fryzlewicz (2015), and SUBSET, detailed in Tickle et al. (2021) are two methods proposed to overcome the issue of detecting sparse and dense changes. Most multivariate changepoint methods

assume independence across series and time, though see Tveten et al. (2022) and Cho and Fryzlewicz (2015) for methods for detecting a change in mean in multivariate time series that allows for cross-correlation; and Romano et al. (2022) and Cho and Fryzlewicz (2024) for work that allows for auto-correlation when detecting changes in mean in univariate data.

One promising line of work for detecting changes in multivariate series which have both temporal dependence and cross-correlation, is by modelling the series as a vector autoregressive (VAR) model, and detecting changes in the parameters of such a model. VAR models, a multivariate extension of the autoregressive time series model, allow for auto-correlation as well as cross-correlation by modelling each variate of a multivariate time series at time t as some linear combination of previous observations — across all series — plus noise. Examples of work on changepoints in VAR models include Cho et al. (2024), Kirch et al. (2015), Safikhani and Shojaie (2022) and Wang et al. (2019). These methods can all search for multiple changepoints, but are encumbered by computational complexity: they require repeated estimation of the parameters of VAR models for each segment under consideration. Several also require the selection of a large number of hyperparameters, making them hard to tune. Another issue is that none currently can include exogenous variables. Exogenous variables can be accounted for in a VAR process with the VAR-X model. This models the current value of a VAR-X process as a linear combination of previous observations, plus a linear combination of exogenous variables and a white noise error term (see, for example, Lütkepohl (2013, p. 387).

In this chapter we propose a novel approach for exploring the issue of detecting a single changepoint in a VAR process with exogenous variables. We seek to combine the VAR-X approach with the SUBSET approach of Tickle et al. (2021). This is a computationally simple approach that allows us to detect a sparse or dense change in a multivariate time series while accounting for the complex cross-correlation and/or

auto-correlation structures that can be modelled in a VAR process. We will compare the performance of this method, which we call SUBSET VAR, to Sparsified Binary Segmentation from Cho and Fryzlewicz (2015) and the VAR based method described by Wang et al. (2019).

In the rest of this chapter we give an overview of the SUBSET-VAR method, before discussing key components of it: VAR and VAR-X models and the SUBSET changepoint detection method (Section 3.2). In Section 3.3 we give the results of a simulation study that compares the performance of our method with existing multivariate changepoint detection techniques. In Section 3.4 we briefly discuss missing data, since our air quality and meteorological data have missing observations. We describe a method of imputing the data and show through simulation how missing data affects the performance of method. We then, in Section 3.5 demonstrate our method on the air quality sensor readings in Glasgow to determine whether the impact of the introduction of the Low Emissions Zone in at the end of 2018 can be detected. We end the chapter in Section 3.6 with conclusions and a description of avenues for further work.

3.2 Background

In this section we briefly describe the SUBSET VAR method, before introducing the VAR, VAR-X and changepoint detection models that underpin it.

3.2.1 SUBSET VAR: detecting a single change in a VAR-X process

Our approach to detecting changes in the second order structure of a multivariate time series is to first transform the data so that, in the absence of a change, we have approximately normally distributed data time series with little to no serial correlation. We then use an existing method, SUBSET, to detect a change in the marginal variance

for one or more time series. We give an overview of the steps of the process below.

1. Divide dataset into a training and a test set.
2. Fit a VAR-X model to the training set.
3. Use the parameters of this model to obtain predicted values for the test set.
4. Obtain residuals of predicted values versus observations in test set.
5. Apply SUBSET to detect a single change — either sparse or dense — in the behaviour of the residuals.

If we have a model that fits the data well, then the residuals should behave as white noise — with no obvious dependence on time (Tsay, 2013) and showing no structural changes. We apply this idea with our method: if we fit a model to observed data, then the residuals up to a changepoint should accord with the above assumption. If there is a changepoint, the model should stop fitting and we will be able to detect a change by inspecting the behaviour of the residuals. Tsay (1988) uses examination of the residuals to detect changepoints and outliers in univariate time series. If our assumptions are not met then our method may be able to pick up the signal of a change in spite of this. This is explored through simulation in Section 3.3.

Since our method depends on detecting when the model stops fitting, care should be taken with the fit of the model to the training data. Checking goodness of fit involves verifying that the residuals are white noise. Additionally, since with SUBSET-VAR we search for a change in the behaviour of normally distributed residuals, we include the common assumption that the error terms of our VAR or VAR-X models are Gaussian and include checks that the residuals are normally distributed. These checks can include examining plots of the auto-correlation function of the residuals to check for auto-correlation, seasonality and non-stationarity; plots of the residuals to examine them for normality. Additionally tests for hetero-skedasticity can be done as

the residuals should not have changing variance: see, for example, Tsay (2013, pp. 66-80) and Lütkepohl (2013, pp. 157-189).

3.2.2 The VAR and VAR-X models

VAR models were developed in the first half of the 20th century (see, for example, Qin (2011) for further details on the development of this class of models). The VAR-X model, where a VAR process is influenced by one or more time series of exogenous variables, is of interest for analysis of the data set under consideration in this chapter, because meteorological variables can be considered to influence pollution levels but not *vice versa*.

We return to the VAR and VAR-X models described in Chapter 2 Section 2.2.1. For a VAR (p) process, we consider a multivariate series, $\boldsymbol{y}$ with dimension d, where the d-dimensional vector of observations at time t is:

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t.$$

As described previously, $\boldsymbol{y}_t$ is a $d \times 1$ vector of observations at time t, each $\boldsymbol{\phi}_i$ matrix is a $d \times d$ coefficient matrix describing the relationship between the process at time and $t - i$, $i = 1, \ldots, p$, and t. The $\boldsymbol{e}_t$ are d-dimensional vectors of error terms. And for the VAR-X (p, s) process, with a m dimensional multivariate time series of exogenous variables:

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \sum_{j=0}^{s} \boldsymbol{\pi}_j \boldsymbol{x}_{t-j} + \boldsymbol{\epsilon}_t.$$

The $\boldsymbol{\pi}_j$ are $d \times m$ matrices of coefficients representing the influence of the exogenous variables $\boldsymbol{x}$ on $\boldsymbol{y}_t$ at lag j, $j = 0, \ldots, s$.

As described in Chapter 2 Section 2.2.1, the coefficients of these models can be estimated using least squares (Tsay, 2013). Throughout this chapter, we use the `R`

package MTS (Tsay and Wood, 2018), to estimate the parameters of VAR and VAR-X processes through least squares.

Obtaining the residuals from a test set

To run our method, we require a measure of the predicted values of the test set versus the observations. Defining a test set of length N, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ where the observations at time t are $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{d,t})^T$, and with corresponding observed m-dimensional series of exogenous variables $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{m,t})^T$. We obtain the residuals of observations versus predicted values, denoted $\boldsymbol{r}$ using the parameters estimated from the training set, $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_p, \hat{\boldsymbol{\pi}}_1, \ldots, \hat{\boldsymbol{\pi}}_s$, by the methods described above in Section 3.2 by the following:

$$\boldsymbol{r}_t = \boldsymbol{y}_t - \sum_{i=1}^{p} \hat{\boldsymbol{\phi}}_i \boldsymbol{y}_{t-i} - \sum_{j=0}^{s} \hat{\boldsymbol{\pi}}_j \boldsymbol{x}_{t-j}$$

We expect a change in the second order structure of the VAR process to manifest as a change in the variance of the residuals. In the next section we will explain how we use the SUBSET algorithm to search these residuals for a change in variance.

3.2.3 Detecting a sparse or dense change in variance

We begin by describing how to find a change in variance in a univariate sequence, before moving on to the multivariate context. Following the notation in Killick et al. (2012), let there be a series of sequential data of length n, denoted $r_{1:n}$, with a single changepoint, τ. If we have a sequence of realisations of Gaussian data that are independently and identically distributed, and at some point τ in the sequence there is a change in variance (but the mean remains zero) then:

$$r_t \sim \begin{cases} N(0, \sigma_1^2), & \text{if } 1 \leq t \leq \tau, \\ N(0, \sigma_2^2) & \text{if } \tau + 1 \leq t \leq n. \end{cases}$$

To find a single changepoint it is common to use a likelihood-ratio test. To define the loglikelihood ratio test statistic we need to define the log-likelihood for a segment of data. The log-likelihood for a segment of data $r_{s:t}$, assuming it comes from a single segment with common variance σ^2 is:

$$l(\sigma^2; r_s, \ldots, r_t) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=s}^{t} r_j^2.$$

We denote the maximum of this as:

$$\ell(r_{s:t}) = (t-s+1)\left[\log(2\pi) + \log\frac{\sum_{j=s}^{t} r_j^2}{t-s+1} + 1\right].$$

Then the log likelihood ratio statistic for a change at $t = \tau$ for $t = 1, \ldots, n$ is:

$$T_\tau = \ell(r_{1:\tau}) + \ell(_{\tau+1:n}) - \ell(r_{1:n}) =$$
$$n\left(\log\frac{\sum_{i=1}^{n}(r_{i,j})^2}{n}\right) - t\left(\log\frac{\sum_{i=1}^{t}(r_{i,j})^2}{t}\right) - (n-t)\left(\log\frac{\sum_{i=t+1}^{n}(r_{i,j})^2}{n-t}\right)$$

SUBSET (Tickle et al., 2021) is designed to detect changes in multivariate series. This method is able to test for a change in some of the series (a sparse change), or across all (a dense change) by considering sums of the penalised likelihood ratio test statistics $T_{i,\tau}$ calculated for all of the individual series, $i = 1, \ldots, d$, in a multivariate time series. It uses two different penalty regimes one for a dense change, and one for a sparse (the sparse also identifying which series undergo a change). Then the penalised test statistic that has the most evidence for a change determines whether or not a sparse change or a dense change is deemed to have occurred. For a sparse change, we define: $T'_{i,\tau} = \max(T_{i,\tau} - \alpha, 0)$, and the test statistic is constructed as $\sum_{i=1}^{d} T'_{i,\tau} - \beta$. For the dense change, it is $\sum_{i=1}^{d} T_{i,\tau} - K$. α, β and K are chosen constants. If at least one of these statistics is greater than zero at time τ, a change in declared — it will be deemed either a sparse change or a dense change depending on which of the test statistics is

greater. We give the test statistic, S_τ for SUBSET below:

$$S_\tau = \max\left\{\sum_{i=1}^{d} T'_{i,\tau} - \beta, \sum_{i=1}^{d} T_{i,\tau} - K\right\}.$$

3.3 Simulation Study

In this section we test our method on five simulation scenarios. As set out in Section 3.1, we compare SUBSET VAR to SBS and a method implementing the LASSO-VAR estimator described by Wang et al. (2019). We give a brief overview of these methods below.

The first is the Sparsified Binary Segmentation (SBS) method of Cho and Fryzlewicz (2015) implemented in the R package `hdbinseg` (Cho and Fryzlewicz, 2018). This is a general method for detecting changes in the second order of a multivariate series. For each series, SBS first transforms it to wavelet periodograms of a Locally Stationary Wavelet model, so that a change in the second order structure becomes a change in mean of the wavelet parameters. It uses a CUSUM test for a change in mean for each series and then uses SBS to combine information across series. The last step sums up the evidence for a change at a given location, τ across series but only for those where the CUSUM test statistic for a change at τ is greater than a pre-specified threshold. The other method, of Wang et al. (2019), detects changes in the structure of a Vector Autoregressive process, based on a likelihood test statistic that uses LASSO regression to estimate the fit of a VAR model on the data. We denote this method LASSO-VAR It applies dynamic programming to the problem of searching for multiple changepoints. This is implemented in the `changepoints` R package (Padilla et al., 2022). We adapt it to search for a single changepoint for this application. Hyperparameters for all methods are determined by simulation (see Appendix A.1.2 for further details).

As both of these competitor methods only work for VAR, rather than VAR-X, we

do not model any scenarios including exogenous covariates. The first three scenarios assume the VAR model for the data is correct, and just differ in the type of change and the number of series that are affected. The last two scenarios test the robustness of methods to situations where the VAR assumptions do not hold. As described in Section 3.2.1, we require a training period where we fit a VAR model to the data, before a test period where we obtain residuals from the fitted values predicted by the model fitted in the training period. We discuss selection of training and test data periods in Section 3.5. In our simulations we keep the training and test periods equal or roughly equal to reflect the ratio in the application that is necessitated by the nature of the data and its availability. Since SBS and LASSO-VAR do not require fitting to training data, we run them on the test data only. τ denotes the location of the true changepoint in the test data. We introduce a VAR(1) process, $\boldsymbol{y}_t$:

$$\boldsymbol{y}_t = \boldsymbol{\phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{y}_t$ is a d-dimensional vector of observations at time t, $\boldsymbol{\phi}_1$ is a $d \times d$ matrix of coefficients. The $\boldsymbol{\epsilon}_t$ normally distributed with covariance matrix I_d. For the first three simulations we vary d, $d = 5, 10, 50$. For the final two simulations $d = 5$. The eigenvalues of $\boldsymbol{\phi}_1$ are denoted λ_i, $i = 1, \ldots, d$:

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}.$$

We describe the five simulation scenarios below:

1. Simulation 1: $d = 5, 10, 50$. $\alpha = -0.9, -0.5, -0.1, 0.1, 0.5, 0.9$. Coefficients to change: first 20% of entries of Λ, i.e., $\lambda_1, \ldots, \lambda_{0.2d}$. $n_{train} = 1000$, $n_{test} = 1000$, $\tau = 500$. In this simulation, we control the size of change using a constant α,

$\alpha = -0.9, -0.5, -0.1, 0.1, 0.5, 0.9$. This is the constant by which we multiply the entries of λ that are to undergo a change. The changes get smaller as α gets closer to 1.

2. Simulation 2: $\boldsymbol{y}_t = \boldsymbol{\phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t$ where $t \leq \tau$. For for $t > \tau$ $\boldsymbol{y}_t = \boldsymbol{\phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t + \boldsymbol{x}_t$, where $\boldsymbol{x}_t = \boldsymbol{\psi}_1 \boldsymbol{x}_{t-1} + \boldsymbol{u}_t$ is a stable VAR process of dimension $f = 0.4d$ (sparse change). ψ_1 is an $f \times f$ matrix of coefficients, $\boldsymbol{e}_t$ and $\boldsymbol{u}_t$ are respectively d and f-dimensional vectors of error terms. $n_{train} = 1000$, $n_{test} = 1000$, $\tau = 500$. It is readily verified by (Lütkepohl, 1984) that adding two stable VAR processes together results in a stationary process. For this simulation and the following simulation, we vary the size of change using a constant, α, $\alpha = -0.9, -0.5, -0.1, 0.1, 0.5, 0.9$. Here α represents the constant that the nuisance series, $\boldsymbol{x}_t$ is multiplied by. This means that the closer α is to 0, the smaller the change.

3. Simulation 3: $\boldsymbol{y}_t = \boldsymbol{\phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t$ where $t \leq \tau$. For $t > \tau$ $\boldsymbol{y}_t = \boldsymbol{\phi}_1 \boldsymbol{y}_{t-1} + \boldsymbol{e}_t + \boldsymbol{x}_t$, where $\boldsymbol{x}_t = \boldsymbol{\psi} \boldsymbol{x}_{t-1} + \boldsymbol{u}_t$ is a stable VAR process of dimension d (dense change). $\boldsymbol{\psi}_1$ is a $d \times d$ matrix of coefficients, $\boldsymbol{e}_t$ and $\boldsymbol{u}_t$ are both d-dimensional vectors of error terms, with covariance matrices I_d, $n_{train} = 1000$, $n_{test} = 1000$, $\tau = 500$.

4. Simulation 4: following the notation of Taylor et al. (2019),

$$\boldsymbol{y}_t = \sum_{j=1}^{J} \sum_{l=1}^{T} \boldsymbol{V}_j(l/T) \xi_{j,l}(t) \boldsymbol{z}_{j,l}.$$

Here the $d \times d$ lower triangular matrix, $\boldsymbol{V}_j(u)$, denotes the transfer function, $\xi_{j,l}$ are a set of Daubechies extremal phase wavelets for location l and level j, and the $\boldsymbol{z}_{j,l}$ are error vectors that are uncorrelated and have mean zero. The spectral representation of the process is given by:

$$\boldsymbol{S}_j(u) = \boldsymbol{V}_j^T(u) \boldsymbol{V}_j(u)$$

for $u := \frac{t}{T} \in (0, 1)$. In our simulation $T = 1024$, $d = 5$, there are 10 levels ($J = 10$). We vary α, the level at which we introduce a change. We describe the predetermined structure of the data as follows:

$$\boldsymbol{S}_\alpha(u) = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

for $t < \tau$, and

$$\boldsymbol{S}_\alpha(u) = \begin{pmatrix} 5 & 5 & 0 & 0 & 0 \\ 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We also define

$$\boldsymbol{S}_4(u) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5, \end{pmatrix}$$

for all t. We vary the level in which a change is made, simulating $\alpha = 2, 4, 6, 8$. $T_{train} = 500$, $T_{test} = 524$, $\tau = 300$. $d = 5$ for all repetitions.

5. Simulation 5: gradual change taking place between τ_1 and τ_2.

Our d-dimensional process, $\boldsymbol{y}_t$, is of the form:

$$\boldsymbol{y}_t = a\boldsymbol{x}_t + b\boldsymbol{z}_t$$

where $\boldsymbol{x}_t, \boldsymbol{z}_t$ are stable, stationary VAR processes: $\boldsymbol{x}_t = \boldsymbol{\psi}_1\boldsymbol{x}_{t-1} + \boldsymbol{u}_t$, and $\boldsymbol{z}_t = \boldsymbol{\zeta}_1\boldsymbol{z}_{t-1} + \boldsymbol{v}_t$. $\boldsymbol{\psi}_1$ and $\boldsymbol{\zeta}_1$ are $d \times d$ matrices of VAR coefficients and $\boldsymbol{u}_t, \boldsymbol{v}_t$ are vectors of normally distributed error terms with mean zero. In this scenario $d = 5$ for all repetitions.

Then for $t \leq \tau_1 : a = 0.95, b = 0.05$:

For $\tau_1 < t \leq \tau_2$, $a = 0.95\frac{\tau_2 - t}{\tau_2 - \tau_1}$ and $b = 0.05\frac{t - \tau_1}{\tau_2 - \tau_1}$

For $t > \tau_2, a = 0.05, b = 0.95$

$n_{train} = 1000$, $n_{test} = 1000$, $\tau_1 = 300, \tau_2 = 600$.

Each of our methods have hyperparameters, such as thresholds or penalties, which determine the amount of evidence required to infer a change. We tune these to return a false positive rate of 1% for all algorithms used in our simulation study. We determine these by simulating data without a change and running all three methods on these datasets (we run them on the same datasets so that variation in the datasets does not lead to one method being given a poorer or better choice of hyperparameter than the others). See Appendix A.1.2 for more details on the hyperparameters for each method.

The result plots show the average distance from the true change by method. We only include results from repetitions where both methods identified a change. This avoids the case where a method may be penalised for being able to pick up a more difficult change — if it picks up difficult changes, where competitor methods fail, but when detecting difficult changes it is less accurate in its detection of the time point at which the change occurred than in those cases where an easy change is detected, then it may appear to be less accurate than those methods that can only pick up easy changes. We also report the percentage of runs where a method identifies a changepoint

— versus declaring a false negative — in each scenario.

3.3.1 Results

Table 3.3.1: Percentage of true positives flagged by each method in Simulation Scenario 1, according to size of change, α, and number of dimensions

Dimensions	α	LASSO-VAR	SBS	SUBSET VAR
5	-0.9	100.0	100.0	100.0
5	-0.5	100.0	100.0	100.0
5	-0.1	100.0	100.0	100.0
5	0.1	100.0	99.9	100.0
5	0.5	100.0	99.6	47.0
5	0.9	24.2	26.9	1.1
10	-0.9	100.0	100.0	100.0
10	-0.5	100.0	100.0	100.0
10	-0.1	100.0	100.0	100.0
10	0.1	100.0	100.0	100.0
10	0.5	100.0	100.0	95.8
10	0.9	51.3	70.3	2.4
50	-0.9	100.0	100.0	100.0
50	-0.5	100.0	100.0	100.0
50	-0.1	100.0	100.0	100.0
50	0.1	100.0	100.0	100.0
50	0.5	100.0	100.0	100.0
50	0.9	100.0	100.0	0.6

In Scenario 1 SUBSET VAR and LASSO-VAR just outperform SBS when the change is large, but are considerably less accurate when the changes are small. Both methods show some drop off in accuracy as the change in the eigenvalues gets closer to one, but SUBSET VAR shows a much sharper decrease in accuracy, averaging just below 400 time points away from the true change when looking for the smallest change simulated in a five dimensional series. It also loses accuracy when looking for the second smallest change, but this is less obvious in the higher dimensional simulations. SBS's accuracy improves for the smaller changes when the number of dimensions in the simulation is higher. Because the proportion of variables changing is constant across the simulations, those in higher dimensions result in a larger number of series undergoing a change. It

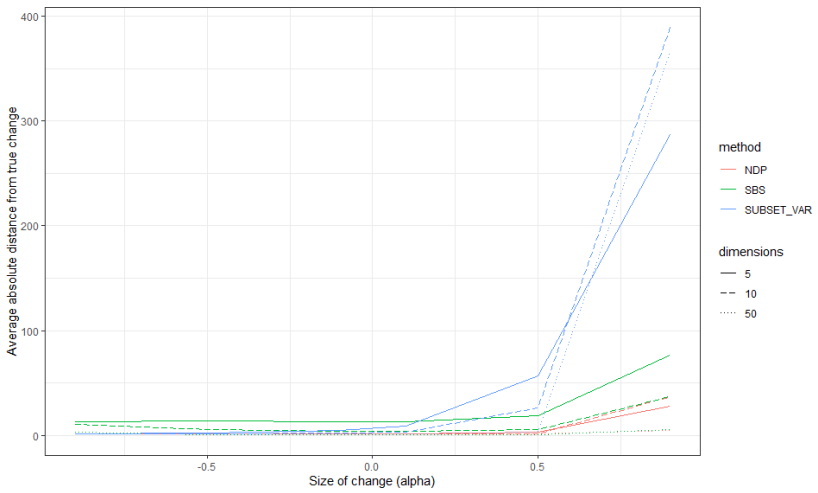


Figure 3.3.1: Accuracy reported Simulation Scenario 1. For every repetition where all three methods find a changepoint, we calculate the absolute distance from the true changepoint of each method. plotted are the average absolute distances against the size of the change, α.

is likely, therefore, that these simulations exhibit more evidence for a change to be accepted.

Table 3.3.2: Percentage of true positives flagged by each method in Simulation Scenario 2, according to size of change, α, and number of dimensions

Dimensions	α	LASSO-VAR	SBS	SUBSET VAR
5	-0.9	99.7	70.7	100.0
5	-0.5	71.2	26.3	93.7
5	-0.1	1.4	1.0	1.0
5	0.1	1.1	1.3	1.0
5	0.5	70.5	24.6	94.2
5	0.9	99.9	73.4	100.0
10	-0.9	100.0	83.5	100.0
10	-0.5	97.9	17.2	100.0
10	-0.1	1.5	1.0	2.0
10	0.1	1.4	0.8	2.1
10	0.5	97.2	18.8	100.0
10	0.9	100.0	83.4	100.0
50	-0.9	100.0	99.4	100.0
50	-0.5	100.0	16.6	100.0
50	-0.1	11.6	2.4	2.4
50	0.1	11.4	3.8	2.2
50	0.5	100.0	16.8	100.0
50	0.9	100.0	99.4	100.0

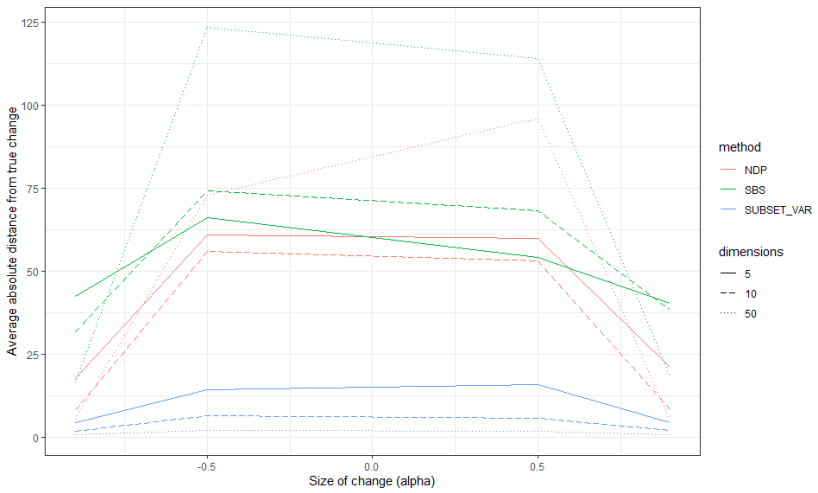


Figure 3.3.2: Results of Simulation Scenario 2. For every repetition where all three methods find a changepoint, we calculate the absolute distance from the true changepoint of each method. plotted are the average absolute distances against the size of the change, α.

In Scenario 2, we see that SBS and LASSO-VAR are less accurate than SUBSET VAR in five and 10 dimensions, although the difference becomes less acute in the higher dimensional simulations. This is potentially because the number of series that undergo a change is higher, producing more evidence for a change. We note, however, that, similarly to Scenario 1, SUBSET struggles to pick up a change when the change is small. In Scenario 3 we see a similar pattern to Simulation 2 with SBS and SUBSET VAR, except that the difference in accuracy between the two methods is less marked. This is likely to be because the number of series that undergo a change is higher than in Simulation 2. We note that again SUBSET struggles to pick up a change when the change is small. None of the methods are very accurate when the change is small. As might be expected, SBS, which was developed with wavelet processes in mind, exhibits the most accuracy in Scenario 4, particularly on the higher levels. SUBSET VAR is less accurate than SBS, particularly in the higher levels. LASSO-VAR is consistently inaccurate. SUBSET VAR and LASSO-VAR fail to pick up many changes at level 8. in Figure 3.3.5 we plot the difference each method is from the mid-point of the gradual change simulate in Scenario 5, so we find anything between 0 and 150 to be accurate.

As with previous plots, we only report a change where all three methods had found identified a change. We see that all methods perform well, with most of the changes identified within 150 time points from the middle of the transition period.

Table 3.3.3: Percentage of true positives flagged by each method in Simulation Scenario 3, according to size of change, α, and number of dimensions

Dimensions	α	LASSO-VAR	SBS	SUBSET VAR
5	-0.9	100.0	99.2	100.0
5	-0.5	100.0	74.0	100.0
5	-0.1	2.5	0.9	1.0
5	0.1	2.5	0.6	0.9
5	0.5	100.0	77.0	100.0
5	0.9	100.0	99.4	100.0
10	-0.9	100.0	100.0	100.0
10	-0.5	100.0	80.0	100.0
10	-0.1	4.3	0.7	5.6
10	0.1	4.3	0.9	5.1
10	0.5	100.0	80.1	100.0
10	0.9	100.0	100.0	100.0
50	-0.9	100.0	100.0	100.0
50	-0.5	100.0	86.6	100.0
50	-0.1	96.0	2.6	98.4
50	0.1	94.6	2.2	97.2
50	0.5	100.0	86.8	100.0
50	0.9	100.0	100.0	100.0

Table 3.3.4: Percentage of true positives flagged by each method in Simulation Scenario 4, according to the level, α, in the LSW process where the change is made.

α	LASSO-VAR	SBS	SUBSET VAR
2	100.0	100.0	29.5
4	100.0	100.0	28.3
6	15.0	99.2	18.2
8	2.4	77.1	10.5

Table 3.3.5: Percentage of true positives flagged by each method in Simulation Scenario 5.

Dimensions	LASSO-VAR	SBS	SUBSET VAR
5	100	99.4	99.9

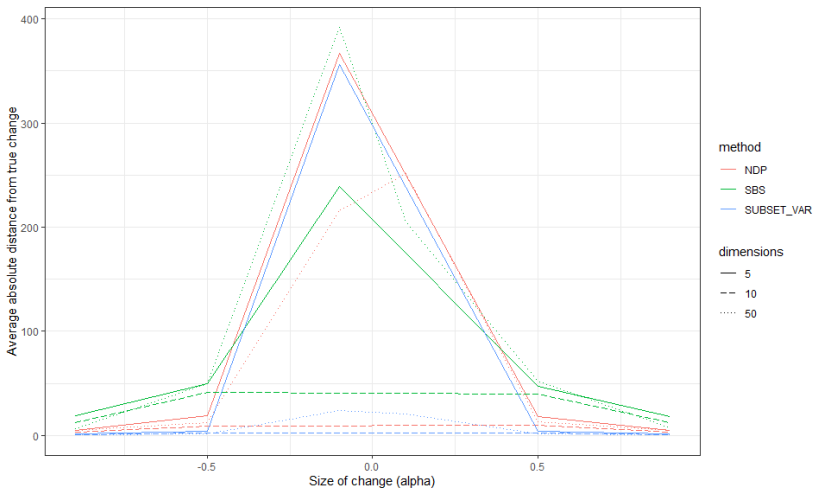


Figure 3.3.3: Results of Simulation Scenario 3. For every repetition where all three methods find a changepoint, we calculate the absolute distance from the true changepoint of each method. plotted are the average absolute distances against the size of the change, α .

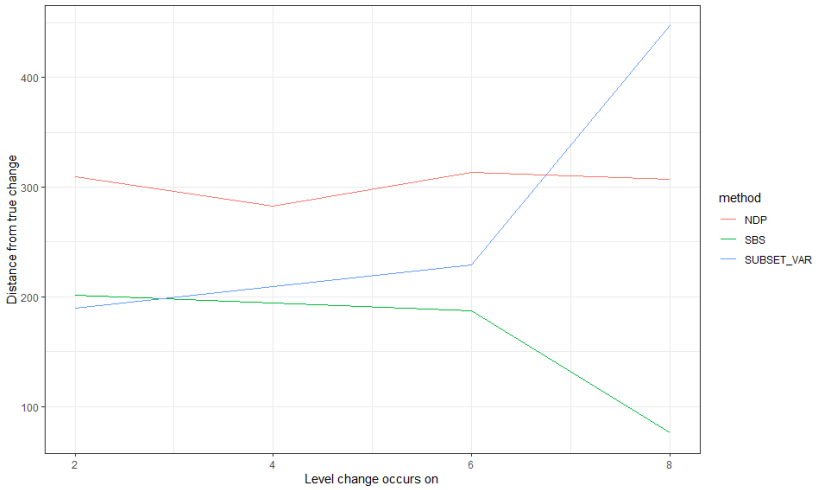


Figure 3.3.4: Results of Simulation Scenario 4. For every repetition where all three methods find a changepoint, we calculate the absolute distance from the true changepoint of each method. plotted are the average absolute distances against the level in which the change is made, α.

3.4 Missing data

As described in Section 3.1, this chapter is motivated by search for a change in a multivariate time series of air quality data in Glasgow, with the aim of determining whether a change in public transport emissions standards coincides with a change in

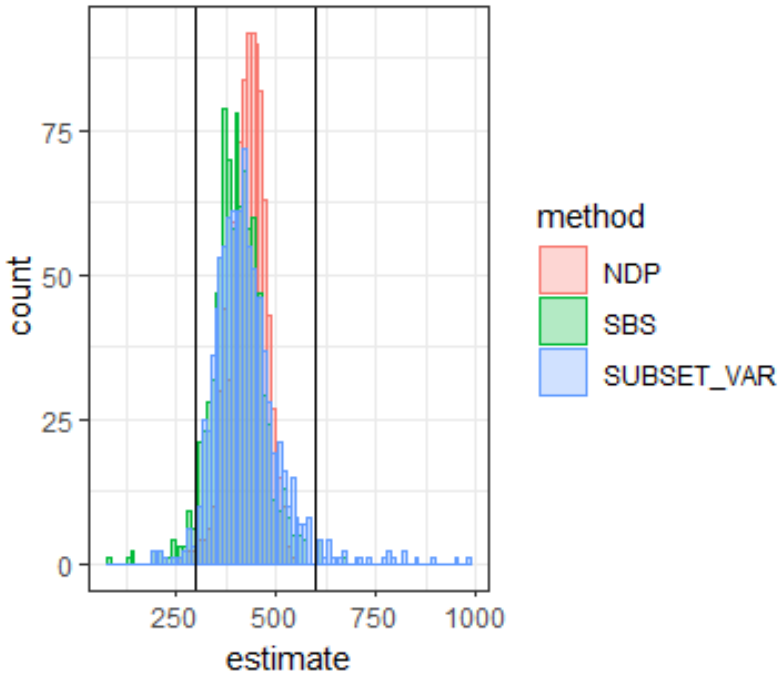


Figure 3.3.5: Reporting the accuracy of methods in Simulation Scenario 5. We plot here a histogram of identified change locations, for every repetition (of 1000 runs) where all three methods identify a change. In this scenario we simulate a gradual change, which begins at $t = 300$ and ends at $t = 600$ (represented on the plot with vertical black lines), so any changepoints flagged outside these lines is deemed inaccurate.

air quality in the city. In doing so we encounter the problem of missing data: we use a time series of air quality data and one of meteorological data for this analysis, and both contain missing values. In our analysis, we remove series that have a run of missing data over a certain length, but if we removed all the series with any missing data we would have few left to analyse. In the air quality data, after we remove variables we deem unsatisfactory due to too long a run of missing data, we retain a series where around 2.7% of cells are missing. In the weather data, about 0.8% of cells are missing.

A straightforward method of dealing with missing data is to delete it. One such method is Complete Case Analysis (see, for example, Van Buuren (2018) and Little and Rubin (2019), where any case (or in a time series, for any t) where there are any missing cells, the whole row is deleted. This can cause biased estimates if the probability that a cell is missing is not a random chance uniformly distributed across all of the data (this

is a missingess pattern known as Missing Completely at Random (MCAR)).

An alternative to deleting missing entries is to replace the missing data with an estimated value — known as imputation. In our application we impute our data using multiple imputation with chained equations. Multiple imputation involves the creation of several imputed data sets, which are analysed separately and the results combined. The advantage of creating several imputed data sets and then analysing them is that it quantifies the uncertainty introduced into the analysis by the presence of missing data (Little, 1988).

Multiple imputation using chained equations, which we will implement in the `R` package `MICE`, introduced by Van Buuren and Groothuis-Oudshoorn (2011), is an algorithm that imputes missing values for each variable in a dataset one at a time. It uses Markov Chain Monte Carlo (MCMC) to draw from the conditional distribution of the missing entries for that variable, given the observed entries of the variable, and the observed and imputed values of the other variables in the dataset. This method of dealing with missing data can be implemented using a variety of imputation methods. In our application we use predictive mean matching (Little, 1988), which is the default method in the `mice` package. In a time series setting, if variable j is missing at time t then predictive mean matching imputes the value of a missing cell of data with the value of that variable's observed data with the closest value to the missing variable's predicted mean. In the `MICE` package the predicted mean for variable j at time t is based on a parameter drawn from the conditional distribution of variable j's missing data, given the observed data and the imputed and observed data of the other variables present. The previous imputed value of the data is then replaced with a entry that is observed for variable j, drawn at random from a group of observed values that are closest to the predicted mean value of variable j at time t. For the first iteration, the imputations for variable j are values of observed data for variable j, selected at random.

Multiple imputation with predictive mean matching is not developed for time series

Table 3.4.1: Percentage of changepoints flagged by SUBSET VAR, presented by size of change and imputation

Imputation	α = -0.9	α = -0.5	α = -0.1	α = 0.1	α = 0.5	α = 0.9
0	100	100	100	100.0	97.5	1.9
1	100	100	100	99.5	34.3	42.5
2	100	100	100	99.7	35.4	43.1
3	100	100	100	99.6	32.6	43.3

data but it does take into account the relationship between variables. We provide below the results of a simulation study where we simulate a VAR (1) process, then remove 3.5% of the data cells in a Missing Completely at Random (MCAR) pattern using the `ampute` function in the `mice` package and impute three times using the method described above. We apply our methods on the original data set and on the three imputations and we compare their performance.

We see that for large changes SUBSET VAR detects a changepoint consistently and accurately, and the variation between imputations is low. For the smaller changes, accuracy declines sharply and the variation of estimates is high. We conclude that we can have reasonable confidence in SUBSET VAR's estimates on imputed datasets if the variance between imputations is low.

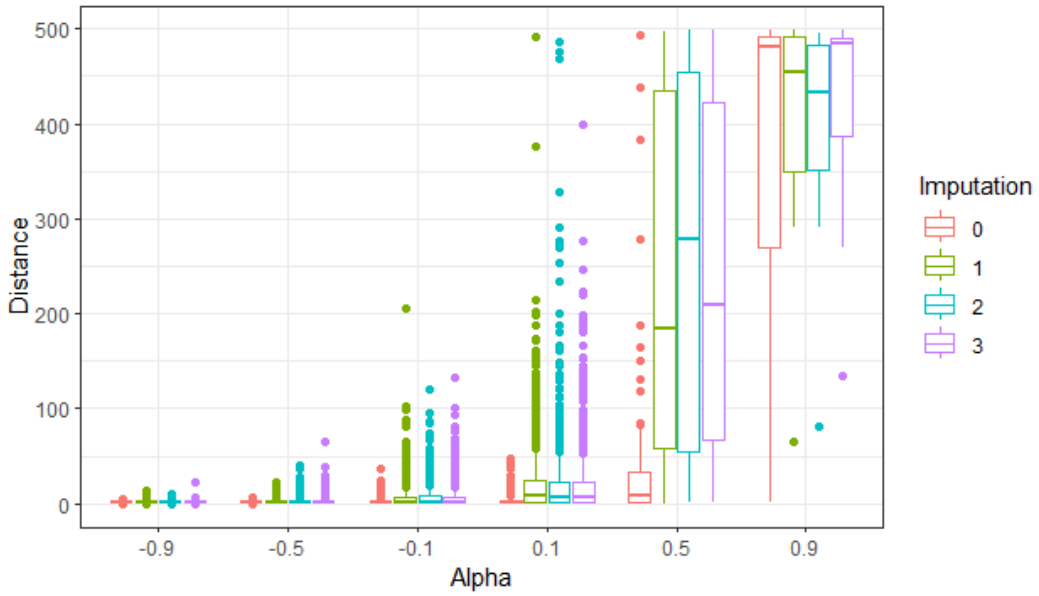


Figure 3.4.1: Accuracy of SUBSET VAR on dataset with approximately 2.7% data removed in a Missing Completely at Random pattern and then multiply imputed (denoted imputations 1-3 in the plot), versus accuracy of SUBSET VAR on dataset prior to removal (denoted imputation 0). We report results only for those repetitions where SUBSET VAR detects a changepoint in all three imputations as well as in the dataset prior to removing data.

3.5 Application: Air Quality in Glasgow

Returning to our application, we aim to determine whether we can identify a change in the covariance structure of a time series of air quality data across Glasgow, near the time a policy geared towards reducing vehicle-based emissions was introduced. The policy intervention, the Low Emission Zone (LEZ), was intended to improve air quality by introducing phased emissions standards for buses at the end of 2018. We model observations of air quality measures in Glasgow and surrounding areas as a multivariate time series. If we detect a changepoint near the time of the intervention — the end of 2018 — we conclude that the impact of this intervention is detectable via a change in the behaviour of the air quality measures, and it therefore had an impact on air pollution in Glasgow. If we detect a change, we end our analysis there, rather than, for example, looking to re-test for a changepoint in the period following the intervention.

Further work could include Interrupted Time Series analysis to examine the extent of the impact on pollution in Glasgow.

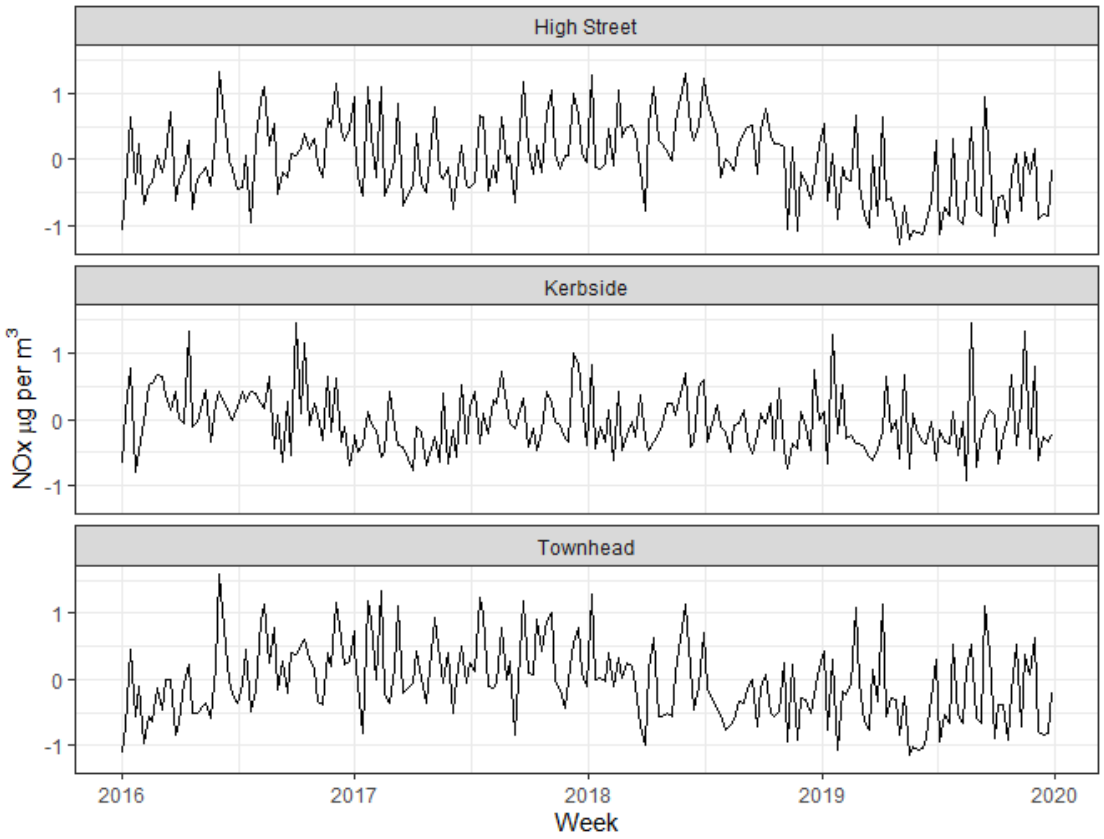


Figure 3.5.1: Daily Nitrous Oxides as Nitrogen Dioxide measurements for central Glasgow sites 2016 – 2019 observed as weekly averages, after seasonal adjusting and outlier removal

We have daily means of Nitrogen Oxides as nitrogen dioxide (NOx) and particulate matter (PM10) levels between 2016 and 2019, inclusive, from air quality monitoring stations in Glasgow (Air Quality in Scotland, 2024). For several of the DEFRA sensors, we have runs of missing data. We exclude any variables at a particular location that have more than 14 days of consecutive missing data between January 1, 2016 and December 31, 2019. We also remove data on days where DEFRA reports an air pollution episode that brings levels in Scotland into the Moderate category or above. See Appendix A.2 for more details.

We perform variable selection on the available data — after excluding those with

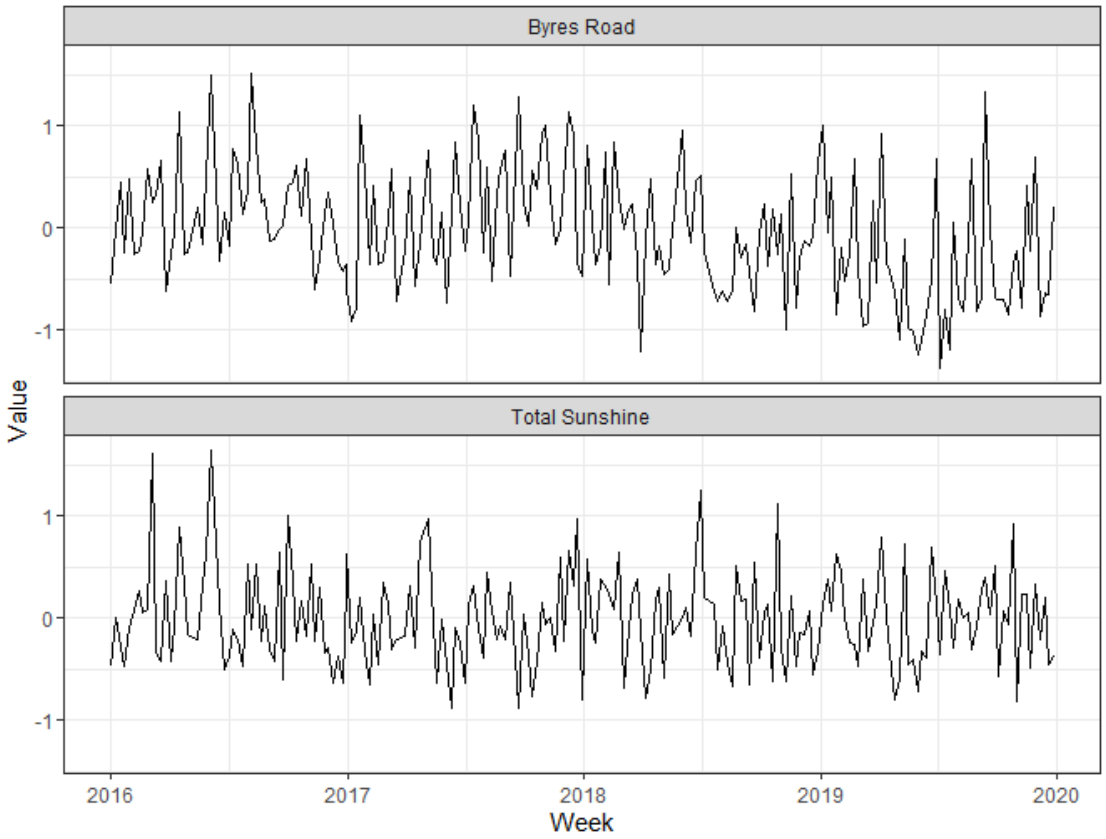


Figure 3.5.2: Daily sunshine data and PM10 levels at Glasgow Byres Road observed as weekly averages across selected sites 2016 – 2019, after seasonal adjusting and outlier removal Met Office (2024)

runs of missing data over 14 days, by testing for Granger causality (Granger, 1969) using the `R` package `lmtest` (Hothorn et al., 2015) at the first lag with series at air quality measuring sites outside and inside the LEZ, and with series of meteorological data. We exclude series inside the zone for which no other variables have Granger causality significant at at least the 10% level, and we exclude meteorological series and air quality series which do not appear to be a predictive factor for series inside the LEZ. We model the multivariate time series as a VAR-X model, with meteorological series and air quality series outside the LEZ as exogenous variables, and the selected air quality series inside the LEZ as the endogenous variables. If the introduction of the LEZ had some effect, then we might expect to see the observed behaviour of pollutants within the affected zone decouple from those in the outer Glasgow areas and the meteorological

series.

After variable selection we then use predictive mean matching from the Multiple Imputation using Chained Equations (MICE) package in R to impute missing data. We take daily meteorological measurements provided by the UK's Met Office (Met Office, 2024): mean temperature, minimum grass temperature, total rainfall, total sunshine and mean windspeed. We also impute any missing data using the same method as used for the pollution data. For the meteorological data, we do this before excluding non-business days from the time series.

The SUBSET VAR method requires splitting our data into a training set and a test set. There are several studies looking at best practice when splitting data into a training and test data set (see, for example, Medar et al. (2017) and Bichri et al. (2024). In practice, however, the choice of split is heavily impacted by data availability and the properties of the data being analysed. In our case, we take 2016 and 2017 as training data and use 2018 and 2019 as test data. We are unable to use data from 2020 onwards due to the onset of the Covid 19 pandemic disrupting the population's activity through stay at home orders and therefore vehicle usage (Scottish Covid-19 Enquiry, 2025). Because the data clearly exhibits annual seasonality we wish to have at least two full years in the training set to enable us to model this. Additionally, as the suspected changepoint is around the end of 2018 or the beginning of 2019 — if we are to see an impact from the Low Emissions Zone — then we include the full years 2018 and 2019 as a test set in which to search for a change. We expect a change could be gradual, as companies and individuals prepare for the new LEZ, but including a large test period allows us to see if we have a spurious change — for example in January 2018 or late 2019. If we include 2015, then we would not be able to use data from Glasgow High Street as it has too long a run of missing data in 2015.

As is clear from Figure 3.1.1, the series of pollution have marked annual seasonality. They also have weekly seasonality as some causes of pollution, like traffic volumes, will

change through a seven day week. To adjust for seasonality, we take the weekly average from our daily data, to handle the weekly seasonality. Then to deal with the annual seasonal pattern we seasonally standardise it by subtracting a smoothed average for that week of the year (the mean of the relevant week as well as the preceding week and the following week) over the four year period, and dividing by the smoothed standard deviation for that week of the year. The adjusted series, after outlier removal, are plotted in Figure 3.5.1. These series suggest that there is potentially a level shift in late 2018 in at least two of the series (Townhead and High Street). We also seasonally adjust the data that are the exogenous variables in our model. These are plotted in Figure 3.5.2

The seasonally adjusted data, depicted in Figure 3.5.1, appear to suggest a change in the form of a level shift, and potentially a variance shift, in NOx levels at the High Street and Townhead sites in the latter half of 2018. We fit a VAR-X(2,3) model, following the fit recommended by the Akaike Information Criterion Akaike (2003) in all five imputed sets of training data and the Hannan and Quinn Information Criterion (Hannan and Quinn, 1979) for four of the five. Diagnostic tests provided by the `MTS` package (Tsay and Wood, 2018), show satisfactory model fit. In three of five imputed data sets multivariate ARCH tests for heteroskedasticity do not reject the null hypothesis of homoskedasticity across four tests at the 5% significance level. In two imputations the hypothesis is rejected for one of four tests — Q(m) of squared series Lagrange multiplier test — one at the 5% level of significance but not at 1% and one (the fifth) at the 1% level of significance. The series of residuals satisfies tests for multivariate normality in all imputations and the Ljung-Box statistics indicate that there is no cause to suspect serial correlation for the first 24 lags. The autocorrelation function of the residuals and the normal quantile-quantile plots are shown in Figures 3.5.3 and 3.5.4, respectively.

We present the results of SUBSET VAR in detecting a changepoint in the data set. Since the change is potentially subtle or slowly evolving, using the higher threshold for

SUBSET VAR (as per the simulation studies) does not detect a change. Results are given in Table 3.5.1 and plotted in Figure 3.5.5 for the case where we force SUBSET VAR to detect a single change. It consistently detects a change in December 2018, within a few weeks of the introduction of the LEZ.

Table 3.5.1: Week of changepoints flagged by SUBSET VAR, over five imputations

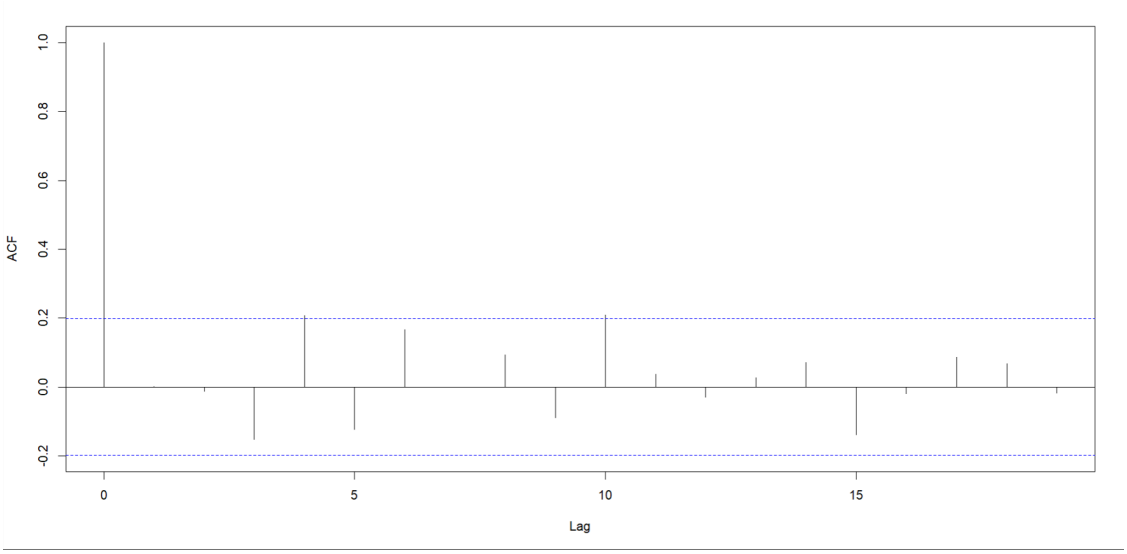
Imputation	Week
1	December 21, 2018
2	December 21, 2018
3	December 14, 2018
4	December 21, 2018
5	December 21, 2018

3.6 Discussion

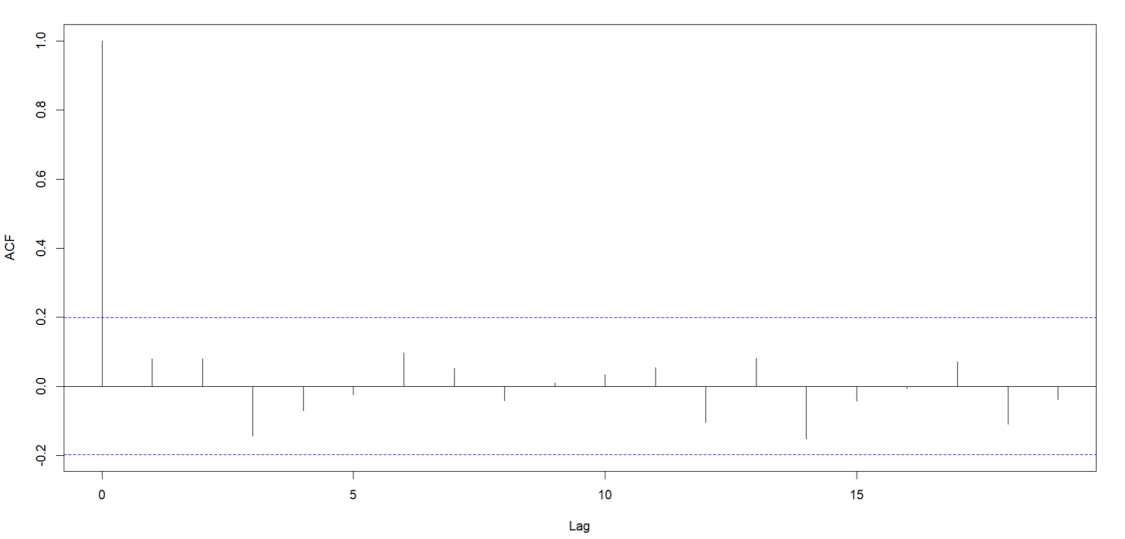
We have introduced a method that brings some of the benefits of VAR-based changepoint detection models without the ponderous computation that can hamper these methods in higher dimensions. We performed a simulation study to compare the performance of this method against two others in variety of scenarios. We then use the method introduced in this chapter to detect the impact of a policy change on pollution patterns in Glasgow. We include a small study on the impact of running changepoint methods on simulated data sets that contain missing data that has then been imputed. Possible extensions include adapting this method to run in an online setting, and adjusting it to be able to search for multiple changepoints. Another potentially fruitful area of investigation is further investigation into the performance of changepoint methods in the presence of missing data.

3.7 Acknowledgements

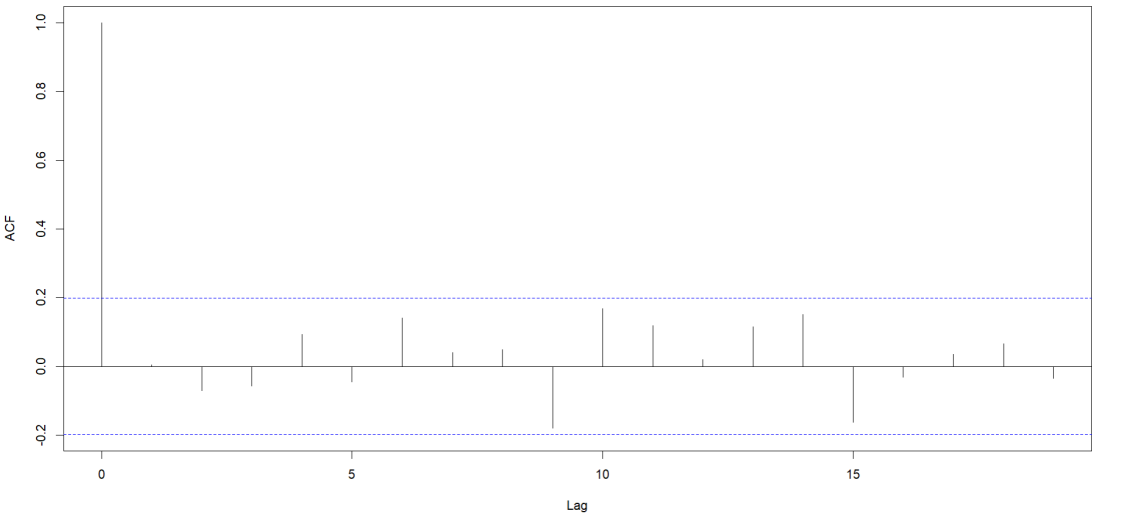
The air quality data used in this chapter was provided by Scottish Air ©Crown 2016 copyright Scottish Government via www.scottishairquality.co.uk, licenced under the Open Government Licence (OGL) (https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/) and the meteorological data was also provided under the OGL and is ©Crown Copyright 2022. Information provided by the National Meteorological Library and Archive — Met Office, UK.



(a) High Street NOx

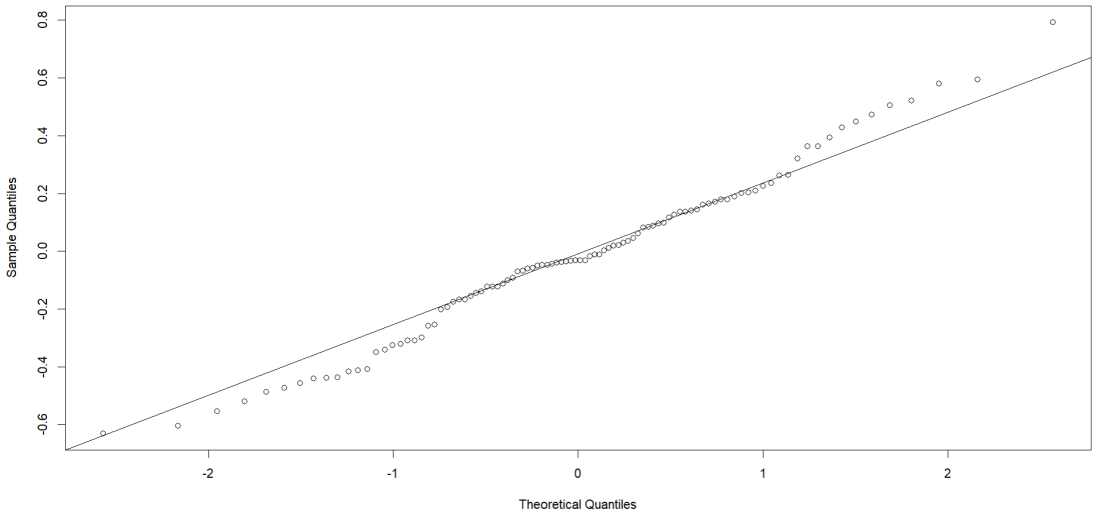


(b) Kerbside NOx

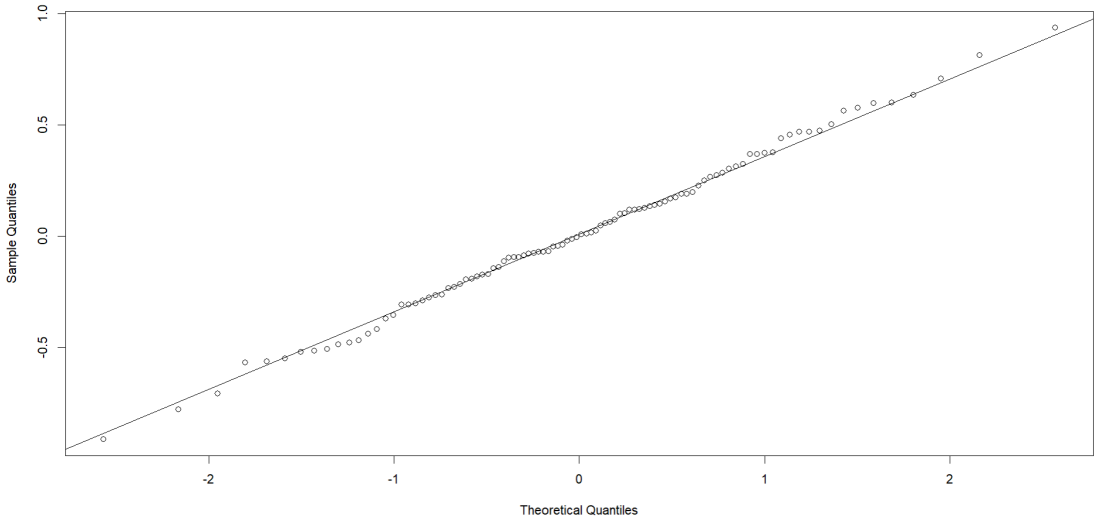


(c) Townhead NOx

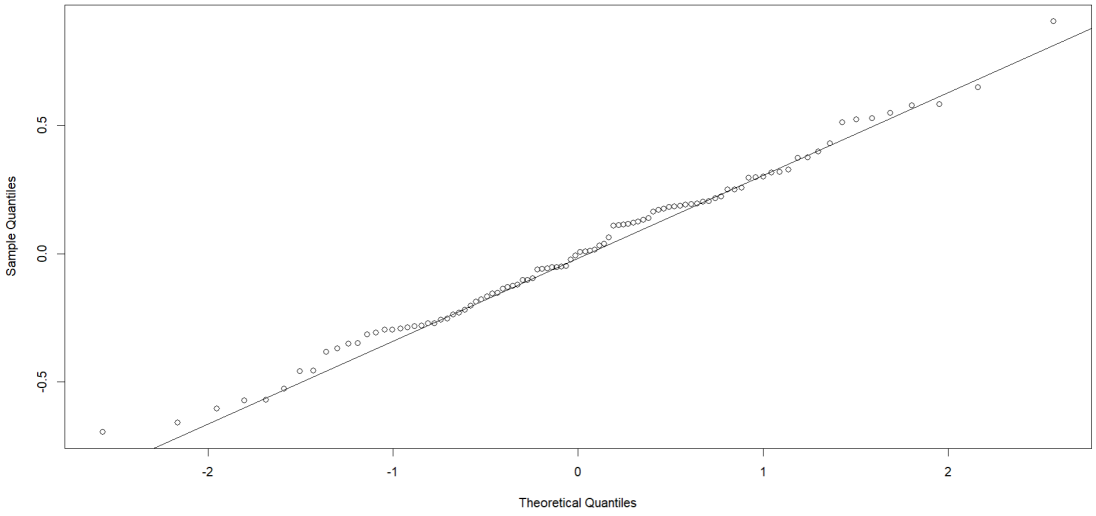
Figure 3.5.3: Plots of the autocorrelation function of the residuals of the VAR-X model (imputation 5).



(a) High Street NOx



(b) Kerbside NOx



(c) Townhead NOx

Figure 3.5.4: Normal quantile-quantile plots of the residuals of the VAR-X model (imputation 5).

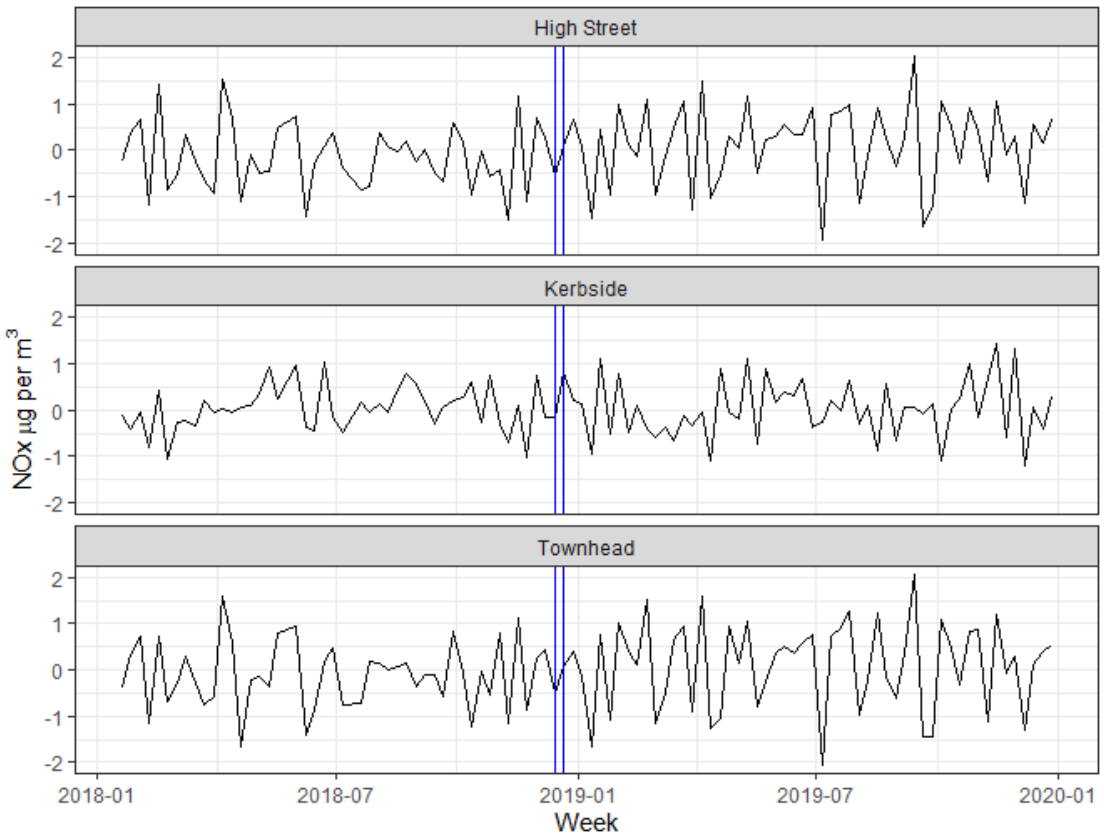


Figure 3.5.5: Plot of the residuals after fitting a VAR-X model to the pollution levels and weather variables after adjusting for seasonality and outliers. The blue lines represent the date on which SUBSET-VAR identifies a change in mean. The plot corresponds to imputation 1 in Table 3.5.1, but identified changepoints from all imputations are shown.

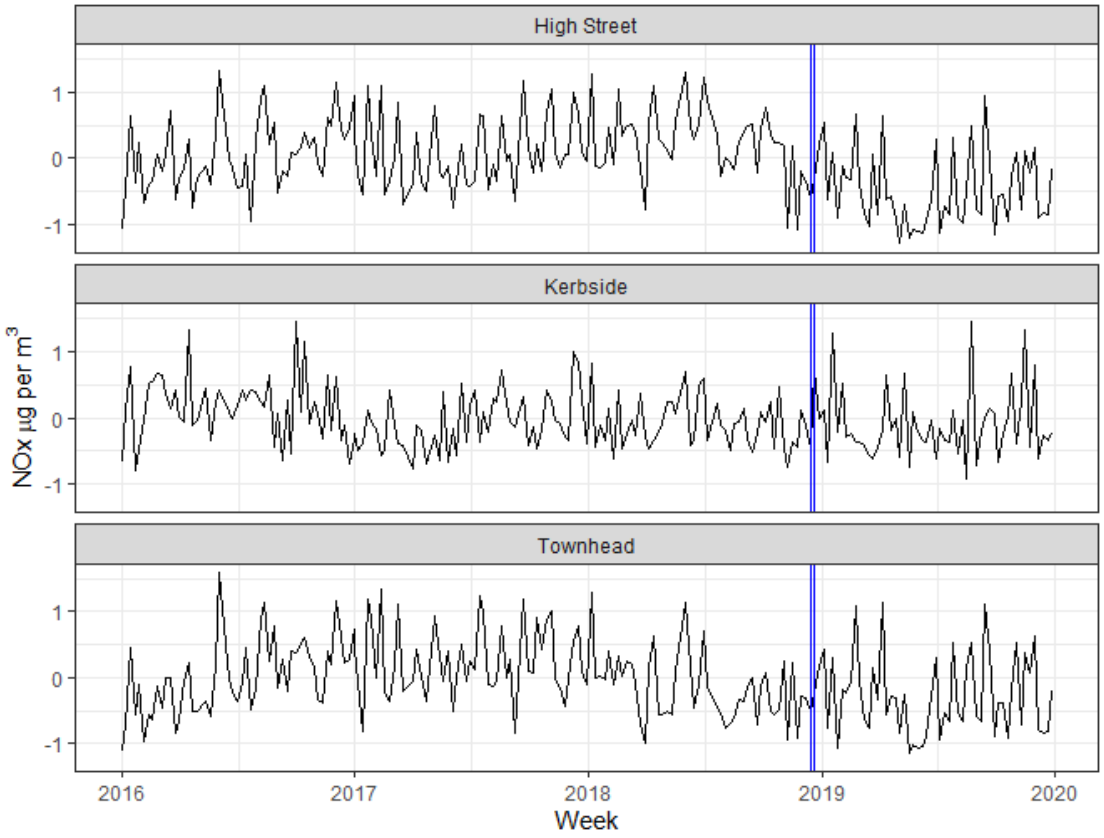


Figure 3.5.6: Plot of the seasonally adjusted series with changepoints identified by SUBSET-VAR shown as a blue vertical line.

Chapter 4

Detecting changepoints in regression models with missing covariate data using a factored likelihood approach

4.1 Introduction

We consider the problem of finding a change in a linear regression model in the presence of missing data. Detecting changes in a linear regression model is a well studied problem that has been applied in a range of areas, including: financial and economic data (Bai and Perron, 1998; Datta et al., 2019); crime rates (Ratcliffe, 2012; Leonardi and Bühlmann, 2016); climate data (Robbins et al., 2016; Friedrich et al., 2020); and industrial applications (Bae et al., 2015). To date these methods have assumed that the data has been completely observed, despite the presence of missing data being a common challenge faced by practitioners. In this chapter we present an approach to detecting changes in linear regression models in the presence of certain types of patterns

of missing data.

A simple way to handle missing data is to remove any cases subject to missingness and perform the statistical analysis using only complete cases (Little and Rubin, 2019). This is known as Complete Case Analysis (CCA), and has been applied (though recommended only in limited circumstances) by Little (1992) to regression models, but without any changepoints being present. One option in our setting would be to extend this work to changepoint detection, but there are several issues with using CCA. The first is that estimators can be biased unless the data is Missing Completely at Random. The second is that it is inefficient. In a regression context this may mean that large amounts of data are removed when only a few covariates contain missing observations. Finally, from a changepoint detection perspective, removing data near to a change could mean that the detection method misses it completely.

An alternative approach is to impute the data. This involves replacing missing values with estimates or other observed values, and then proceeding with the analysis as if the data had been completely observed. Imputation has been applied to problems concerning changepoint detection with missing data by various authors, for example Corradin et al. (2022), Lu (2023), Murph and Storlie (2022) and Zhao et al. (2022). None of these works consider when the data are drawn from a linear regression model. Seidou et al. (2007) address the problem of regression with missing data using a Bayesian imputation method. The use of imputation is not without drawbacks: it can be computationally expensive to impute every missing value; the presence of a changepoint can mean that the model being used to impute missing data is incorrect for the post-change observations.

A commonality between CCA and imputation is that they separate the process of removing the missing data and performing the statistical analysis. An alternative approach would be to incorporate the handling of the missing data explicitly into the changepoint detection method. In our setting, this would mean that the missing data is

taken into account directly in the changepoint detection step of a method, rather than treating them as separate problems entirely or iterating between an imputation step and a changepoint detection step in the same algorithm. In the past five years several authors have tackled the problem from this perspective, for example Cao et al. (2019); Faber et al. (2021); Enikeeva and Klopp (2025); Follain et al. (2022); Londschien et al. (2021). None of these works are designed for use in a linear regression setting.

In this chapter we present a method for changepoint detection in regression models that explicitly models missing data. To achieve this we extend the method of factoring the likelihood, introduced by Anderson (1957) and applied to regression models by Gourieroux and Monfort (1981), to changepoint detection. Factoring the likelihood allows us to break down the full likelihood, which includes the missing data, into a series of conditional and marginal densities. These conditional densities contain the information about the missing data, and each factor of the likelihood can be maximised separately using only the completely observed data. This allows us to easily maximise the full data likelihood, which we then use in a likelihood ratio-based changepoint detection test. Our method has several advantages over CCA and imputation. Compared to CCA, we are maximising the full data likelihood rather than the likelihood for only complete observations. It is also more computationally efficient than explicitly imputing the data, and avoids the risk of introducing errors when imputing post-change data using an incorrect model. On top of this, the factored likelihood approach is based on modelling the distribution of the responses given the covariates and also the distribution of the subset of covariates with missingness given the fully observed covariates. Our change-point approach has power to detect changes in either or both of these.

The rest of this chapter is organised as follows. In Section 4.2 we describe the relevant missing data background for our setting, introduce our model, and outline our proposed approach for changepoint detection. After this in Section 4.3 we demonstrate the effectiveness of our method on simulated data, before showcasing our approach on

financial data and wind farm data in Section 4.4 and then closing with a discussion in Section 4.5.

4.2 Methodology

4.2.1 Background and Model

In this section we will introduce the relevant missing data background and notation, before describing the linear regression model for our data.

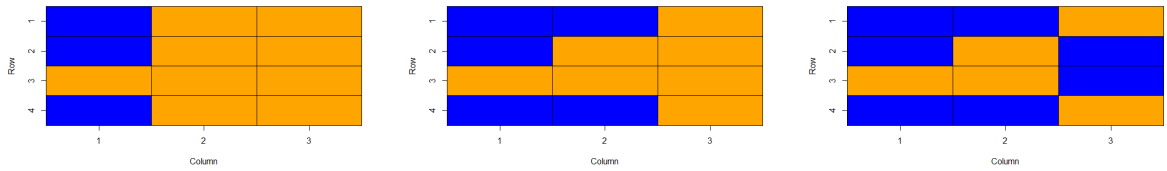
We begin with our general notation for the missing data. We assume, to begin with, that we observe $\boldsymbol{X}$ an $n \times p$ matrix of data with each entry of $\boldsymbol{X}$, $x_{i,j} \in \mathbb{R}$. Later we will introduce additional data $\boldsymbol{y}$ and $\boldsymbol{Z}$, but these will be fully observed. To describe the missingness patterns in $\boldsymbol{X}$, we introduce an $n \times p$ matrix $\boldsymbol{R}$ that indicates which components of $\boldsymbol{X}$ are observed. So, $r_{i,j} = 1$ if we observe $x_{i,j}$ and $r_{i,j} = 0$ otherwise. Furthermore we define $\boldsymbol{X}^{\text{obs}}$ to be the observed data — the entries of $\boldsymbol{X}$ for which $r_{i,j} = 1$; and $\boldsymbol{X}^{\text{mis}}$ the missing data; so $\boldsymbol{X} = \{\boldsymbol{X}^{\text{obs}}, \boldsymbol{X}^{\text{mis}}\}$. Following Little and Rubin (2019), in this chapter we refer to the ith row of $\boldsymbol{X}$, $\boldsymbol{x}_i$, as a case, and denote the i, jth entry, $x_{i,j}$ as an observation. The missingness mechanism describes the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{X}$: $p(\boldsymbol{R}|\boldsymbol{X}, \psi)$ where ψ is an unknown parameter that can be thought of as the relationship between the missingness and the data.

Rubin (1976) introduce classifications for different missingness mechanisms, that is the form of the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{X}$. There are three main classes. Firstly, Missing Completely at Random (MCAR), where the probability of an observation being missing does not depend on $\boldsymbol{X}$ so $p(\boldsymbol{R}|\boldsymbol{X}, \psi) = p(\boldsymbol{R}|\psi)$. Secondly, Missing at Random (MAR), where the probability of x_{ij} being missing depends on some other, observed, variables $\boldsymbol{x}_k, k \neq j$. Under MAR data: $p(\boldsymbol{R}|\boldsymbol{X}, \psi) = p(\boldsymbol{R}|\boldsymbol{X}_{\text{obs}}, \psi)$ for all missing values of $\boldsymbol{X}$. Missing Not at Random describes the case where the probability of $x_{i,j}$ being missing cannot simplify to the MCAR or MAR case. In this chapter we

consider MCAR and MAR data.

The missingness pattern describes which variables are affected by missingness and how the missingness is distributed over those variables. Little and Rubin (2019), for example, give an overview of missingness patterns. In this chapter we consider the univariate missingness pattern, where only one variable is subject to missingness, and monotone missingness patterns (Figure 4.2.1). Data that is missing according to a monotone missingness pattern, described by Little (1992), is where we can order the columns of $\boldsymbol{X}$ such that for each case i for all $j = 1, \ldots, k-1$, if $\boldsymbol{x}_{i,j}$ is observed then $\boldsymbol{x}_{j+1}$ is also observed.

Finally, we introduce indicator functions to denote our fully observed, or missing, datasets. We define I_c as the set of complete cases and I_m the set of missing cases. Then $\mathbf{1}_{i \in I_c}$ and $\mathbf{1}_{i \in I_m}$ are indicator functions for when cases are either observed or missing.



(a) Univariate (b) Monotone (c) General

Figure 4.2.1: Missingness patterns: blue denotes cells that are missing, orange observed. The rows denote cases and the columns, covariates. Univariate missingness: only column one contains missing cells. Monotone: columns 1 and 2 contain missing cells, but for each entry where column 1's cell is observed, column 2's is also. General: neither univariate nor monotone. This chapter deals with univariate and monotone missingness patterns only

4.2.2 Model Inference for Univariate X, Z

Now that we have described missingness mechanisms, missingness patterns and introduced notation, we can introduce our regression model, with a dependent variable, Y, which is fully observed, an independent variable, X, that is subject to missingness, and a fully observed independent variable, Z. Initially we follow Gourieroux and Monfort

(1981) and consider X and Z to be one-dimensional, before showing how the method can be extended to consider higher dimensions of X and Z in Section 4.2.4. The linear regression model is given by:

$$\boldsymbol{y} = \boldsymbol{x}\beta_1 + \boldsymbol{z}\beta_2 + \boldsymbol{\epsilon}, \tag{4.2.1}$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ are $n\times 1$ vectors. The dependent variable Y has mean $\beta_1 x+\beta_2 z$, with β_1, β_2 scalars. The independent variables, X and Z, have normally distributed error term ϵ, with variance σ, though there are no assumptions on the marginal distribution of Z. Following Gourieroux and Monfort (1981) we model that the Xs are related to the Zs as:

$$\boldsymbol{x} = \boldsymbol{z}\gamma + \boldsymbol{\xi}, \tag{4.2.2}$$

where each $\xi_i \sim N(0, \eta)$. Accordingly our regression model has five unknown parameters, $\boldsymbol{\phi} = (\beta_1, \beta_2, \sigma, \gamma, \eta)$.

We now describe our approach for estimating the regression model in (4.2.1) and (4.2.2). This is based on the factored likelihood method introduced by Anderson (1957), and applied to regression models without a change by Gourieroux and Monfort (1981).

The likelihood $L(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r}, \boldsymbol{\phi}|\boldsymbol{z})$ can be expressed as a factored likelihood:

$$L(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r}, \boldsymbol{\phi}|\boldsymbol{z}) = p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\phi})p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\phi})p(\boldsymbol{r}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{y}, \boldsymbol{z}, \psi) \tag{4.2.3}$$

For this likelihood model the vector $\boldsymbol{\phi} = (\beta_1, \beta_2, \sigma, \gamma, \eta)$ represents the five parameters of the regressions models in equations (4.2.1) and (4.2.2), and ψ is the parameter of the missingness mechanism. In this work we are interested in detecting a change in $\boldsymbol{\phi}$. We assume that $\boldsymbol{\phi}$ does not depend on ψ and so using the fact that the data are either MAR or MCAR we can drop the $p(\boldsymbol{r}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{y}, |, \psi)$ term from the likelihood.

Following Gourieroux and Monfort (1981), we notice the remaining two terms are linear models that can be derived from (4.2.1) and (4.2.2). For the first term:

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{x}\beta_1 + \boldsymbol{z}\beta_2 + \boldsymbol{\epsilon} = (\gamma\boldsymbol{z} + \boldsymbol{\eta})\beta_1 + \boldsymbol{z}\beta_2 + \boldsymbol{\epsilon} = \boldsymbol{z}(\beta_1\boldsymbol{\gamma} + \beta_2) + (\beta_1\boldsymbol{\eta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{z}b + \boldsymbol{\nu},\end{aligned}$$

where $\nu_i \sim N(0, a)$, with $b = \gamma\beta_1 + \beta_2$ and $a = \beta_1^2\eta + \sigma$. This is a linear model between Y and Z.

Standard results for the conditional distributions of multivariate normal distributions give that the distribution of $\boldsymbol{x}^{\text{obs}}|\boldsymbol{y}, \boldsymbol{z}$ is also a linear model. Let $\boldsymbol{y}^{\text{obs}}$ and $\boldsymbol{z}^{\text{obs}}$ be the entries that correspond to the observed cases, then by independence across cases $\boldsymbol{x}^{\text{obs}}$ only depends on $\boldsymbol{y}^{\text{obs}}$ and $\boldsymbol{z}^{\text{obs}}$, and this can be written:

$$\boldsymbol{x}^{\text{obs}} = \boldsymbol{y}^{\text{obs}}d + \boldsymbol{z}^{\text{obs}}e + \boldsymbol{\kappa},$$

where $\kappa_i \sim N(0, c)$. See Gourieroux and Monfort (1981) (and Appendix B.1.1) for how the parameters of this model relate to the parameters of the original model (4.2.1) and (4.2.2).

As Gourieroux and Monfort (1981) observed, maximising the likelihood of our original model, (4.2.3), is equivalent to maximising

$$p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\phi})p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\phi}) = p(\boldsymbol{y}|\boldsymbol{z}, a, b)p(\boldsymbol{x}^{\text{obs}}|\boldsymbol{y}, \boldsymbol{z}, c, d, e). \tag{4.2.4}$$

This is because there is a one-to-one and invertible mapping between the two sets of parameters: a to e and those contained in $\boldsymbol{\phi}$. To maximise (4.2.4) we can maximise the terms on the right hand side separately. Compared to the full model, the advantage of this is that there are no missing data in the covariates — we can fit $\boldsymbol{y}|\boldsymbol{z}$ for all cases as $\boldsymbol{y}$ and $\boldsymbol{z}$ are fully observed — and $\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{z}$ for the observed cases, as for these $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$

are fully observed — and maximising the likelihood for each linear model can be done analytically.

Having described this model without a change, we are in a position to consider the model with a single change in parameters and to introduce the concept of factoring the likelihood to changepoint detection, which we will do in the next section.

4.2.3 Change Detection with Univariate X, Z

We now consider the regression model with a change in the parameters. That is we assume the data for each case is observed over time, and we denote the data of the tth observed case by y_t, x_t, z_t. We then consider a model with a potential changepoint as:

$$y_t = x_t\beta_1 + z_t\beta_2 + \epsilon_t,$$

$t = 1, \ldots, n$. For a change in any or all of β_1, β_2 and γ at τ, where τ is the location of the changepoint:

$$y_t = \begin{cases} x_t\beta_1 + z_t\beta_2 + \epsilon_t & \text{if } \ t \leq \tau \ , \\ x_t\beta_1^* + z_t\beta_2^* + \epsilon_t^* & \text{if } \ t > \tau \ . \end{cases}$$

$$z_t = \begin{cases} z_t\gamma + \xi_t & \text{if } \ t \leq \tau \ , \\ z_t\gamma^* + \xi_t^* & \text{if } \ t > \tau \ . \end{cases}$$

No change corresponds to $\tau = n$. If there is a change, $\tau < n$, then this corresponds to a change in the model of $\boldsymbol{y}$ given $\boldsymbol{x}, \boldsymbol{z}$, or in the model of $\boldsymbol{x}$ given $\boldsymbol{z}$, or both. Our model allows for there to be a change in $\beta_1, \beta_2, \epsilon$ without a change in γ, ξ, and *vice versa*. For example for a change in β_1 only, $\beta_1 \neq \beta_1^*$. but $\beta_2 = \beta_2^*, \epsilon = \epsilon^*$ and $\gamma = \gamma^*, \xi = \xi^*$.

Letting $\boldsymbol{\phi} = (\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \tau)$, $\boldsymbol{\phi}^{(1)} = (\beta_1, \beta_2, \sigma, \gamma, \eta)$ and $\boldsymbol{\phi}^{(2)} = (\beta_1^*, \beta_2^*, \sigma^*, \gamma^*, \eta^*)$, we

work with the likelihood introduced in (4.2.3). As before, we can ignore the $P(\boldsymbol{r}|\ldots)$ term since ψ does not depend on $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}$ or τ, and we can factorise the likelihood as in (4.2.3).

We define as $\mathcal{L}_\tau$ the maximum log likelihood for a changepoint at τ.

$$\begin{aligned}\mathcal{L}_\tau = &\sum_{t=1}^{\tau} \log\Big(P(y_t|z_t, \boldsymbol{\phi}^{(1)})\Big) + \sum_{t=1}^{\tau} \log\Big(P(x_t^{\text{obs}}|y_t, z_t, \boldsymbol{\phi}^{(1)})\Big) \\ &+ \sum_{t=\tau+1}^{n} \log\Big(P(y_t|z_t, \boldsymbol{\phi}^{(2)})\Big) + \sum_{t=\tau+1}^{n} \log\Big(P(x_t^{\text{obs}}|y_t, z_t, \boldsymbol{\phi}^{(2)})\Big)\end{aligned}$$

We will use $\mathcal{L}_n(\phi^{(1)}, \phi^{(2)})$ to denote the model for no change, with the convention that sums from $\tau = n+1$ to n are 0, and so this likelihood does not depend on $\phi^{(2)}$. Then the likelihood ratio test statistic (Hinkley, 1970), T_τ, is given by

$$T_\tau = 2\big[\max_{\substack{\tau=1,\ldots,n-1,\\ \phi^{(1)},\phi^{(2)}}} \mathcal{L}_\tau(\phi^{(1)}, \phi^{(2)}) - \max_{\phi^{(1)}} \mathcal{L}_n(\phi^{(1)}, \phi^{(2)})\big].$$

Then $\hat{\tau} = \text{argmax}_\tau\, T_\tau$; and we detect a change if $T_{\hat{\tau}} > \Gamma$, where Γ is the chosen threshold. In this chapter Γ is determined by simulation.

4.2.4 Change detection with multivariate X, Z

The ideas described in Sections 4.2.2 and 4.2.3 can easily be extended to multivariate data with monotone missingness patterns. The approach to detecting the changes will be the same, and the only difference is how to maximise the likelihoods for each segment and calculate the segment cost.

We consider the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{Z}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{4.2.5}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector, $\boldsymbol{X}$ is a $n \times p$ matrix subject to missing observations in a monotone pattern, and $\boldsymbol{Z}$ is a $n \times k$ matrix which is always observed. $\boldsymbol{X}$ is related to $\boldsymbol{Z}$ as below:

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\xi}, \tag{4.2.6}$$

Our data parameters are $\phi = \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma, \boldsymbol{\gamma}, \eta$. $\boldsymbol{\beta}_1$ is a $p \times 1$ vector of coefficients, $\boldsymbol{\beta}_2$ is a $k \times 1$ vector of coefficients and $\boldsymbol{\gamma}$ is a $k \times p$ matrix of coefficients. $\boldsymbol{\epsilon}, \boldsymbol{\xi}$ are $n \times 1$ vectors of normally distributed error terms: σ, η.

As in previous sections we factorise the likelihood as:

$$L(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\phi}|\boldsymbol{Z}) = p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\phi})p(\boldsymbol{x}_p|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\phi})p(\boldsymbol{x}_{p-1}|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{x}_p, \boldsymbol{\phi}) \dots p(\boldsymbol{x}_1|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{X}_{2:p}, \phi), \tag{4.2.7}$$

where $\boldsymbol{X}_{i:j} = (\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_j)$.

As before, we have dropped the $p(\boldsymbol{R}|\dots)$ term, since the missingness mechanism does not depend on $\boldsymbol{\phi}$. Furthermore, as in the univariate case, results on the conditional probabilities of a Gaussian allow us to relate each term in the likelihood to a probability of a linear model. For example, $p(\boldsymbol{x}_i|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{X}_{i:p}, \boldsymbol{\phi}_i)$ can be written as

$$\boldsymbol{x}_i = \boldsymbol{y}d_i + \boldsymbol{z}_1 e_{i,1} + \dots + \boldsymbol{z}_K e_{i,K} + \boldsymbol{x}_{i+1} f_{i,i+1} + \boldsymbol{x}_{i+2} f_{i,i+2} + \dots + \boldsymbol{x}_p f_{i,p} + \boldsymbol{\kappa}_i, \tag{4.2.8}$$

where $\boldsymbol{\phi}_i = d_i, e_{i,1:k}, f_{i,i+1:p}, \kappa_i$ and $\kappa_{t,i} \sim N(0, c_i^2)$. We relate the parameters of (4.2.8) to those of (4.2.5) and (4.2.6) in Appendix B.1.2.

Maximising the likelihood in (4.2.7) can be done by maximising each term separately, with the likelihood trivial to maximise as it is that of a normal linear model, and we only consider the entries corresponding to observed values of $\boldsymbol{x}_i$. Importantly, as we

have monotone missingness, for such an $\boldsymbol{x}_i$ we observed all entries of $\boldsymbol{y}, \boldsymbol{Z}$ and $\boldsymbol{X}_{i+1:p}$, so this likelihood corresponds to a fully observed linear model.

Identical to the univariate case, maximising the likelihood in this way is equivalent to maximising the likelihood for the full model. We can therefore use a likelihood-based approach to detecting a changepoint similar to in Section 4.2.3. Defining the likelihood for a change at τ, with pre-change parameters $\phi^{(1)}$ and post change, $\phi^{(2)}$:

$$\begin{aligned}\mathcal{L}_\tau(\phi^{(1)}, \phi^{(2)}) &= \sum_{t=1}^{\tau} \log\left(P(y_t|z_t, \boldsymbol{\phi}^{(1)}\right) + \sum_{t=1}^{\tau} \log\left(P(x_{t,p}|y_t, \boldsymbol{Z}_t, \phi^{(1)})\right) \\ &+ \sum_{t=1}^{\tau}\sum_{L=1}^{p-1} \log\left(P(x_{t,L}|y_t, \boldsymbol{z}_t, x_{t,L+1:p}, \boldsymbol{\phi}^{(1)})\right) + \sum_{t=\tau+1}^{n} \log\left(P(y_t|z_t, \boldsymbol{\phi}^{(2)}\right) \\ &+ \sum_{t=\tau+1}^{n} \log\left(P(x_{t,p}|y_t, \boldsymbol{Z}_t, \phi^{(2)})\right) + \sum_{t=\tau+1}^{n}\sum_{L=1}^{p-1} \log\left(P(x_{t,L}|y_t, \boldsymbol{z}_t, x_{t,L+1:p}, \boldsymbol{\phi}^{(2)})\right)\end{aligned}$$

As in Section 4.2.3, we denote the model for no change as $\mathcal{L}_n(\phi^{(1)}, \phi^{(2)})$.

Then the likelihood ratio test statistic, T_τ, is given by:

$$2\big[\max_{\substack{\tau=1,\ldots,n-1,\\ \phi^{(1)},\phi^{(2)}}} \mathcal{L}_\tau(\phi^{(1)}, \phi^{(2)}) - \max_{\phi^{(1)}} \mathcal{L}_n(\phi^{(1)}, \phi^{(2)})\big]$$

and, as before $\hat{\tau} = \text{argmax}_\tau\, T_\tau$; and we accept a change if $T_{\hat{\tau}} > \Gamma$, which is determined by simulation.

4.3 Simulation Studies

We simulate scenarios from the models described in Sections 4.2.3 and 4.2.4. For each simulation scenario the model is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta_1} + \boldsymbol{Z}\boldsymbol{\beta_2} + \boldsymbol{\epsilon},$$

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\xi},$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{Z}$ is $n \times 5$, $\boldsymbol{X}$ is $n \times p$. $\boldsymbol{\beta}_1$ is $p \times 1$, $\boldsymbol{\beta_2}$ is 5×2 and $\boldsymbol{\gamma}$ is $5 \times p$. In each scenario the $\boldsymbol{z}_i$ are simulated as independent $U(0,1)$ processes of length $n = 1000$. Each entry of the coefficient matrices $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$ and $\boldsymbol{\gamma}$ pre-change are simulated at random as $U(0,1)$. We define a change factor, α, which is a constant that we multiply the regression coefficients by to effect a change. If the scenario specifies a change in these coefficients they are multiplied by a change factor at $t \geq \tau$. We denote a small change when the change factor is 1.5, a medium change when it is 1.75 and a large change when it is 2. An overview of the five scenarios is given below. We also include a sixth, which simulates data with no change.

1. $p = 1, \epsilon_t \sim N(0,1), \xi_t \sim N(0,\eta), \eta = 1$. Coefficients to change: $\beta_1, \boldsymbol{\beta}_2$.
2. $p = 2, \epsilon_t \sim N(0,1), \boldsymbol{\xi}_t \sim MVN(0,\boldsymbol{\eta}), \eta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Coefficients to change: $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$.
3. $p = 2, \epsilon_t \sim N(0,1), \boldsymbol{\xi}_t \sim MVN(0,\boldsymbol{\eta}), \boldsymbol{\eta} = \begin{pmatrix} \eta^2 & \omega \\ \omega & \eta^2 \end{pmatrix}$. Coefficients to change: $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$.
4. $p = 1, \epsilon_t = 0.5\epsilon_{t-1} + \upsilon_t$, $\upsilon_t \sim N(0,1), \xi \sim N(0,\eta), \eta = 1$. Coefficients to change: β_1, β_2.
5. $p = 1, \epsilon_t \sim N(0,1), \xi_t \sim N(0,\eta), \eta = 1$. Coefficients to change: $\boldsymbol{\gamma}$.
6. $p = 1$, $k = 1$. $Z_t = q \times t + \delta_t$, where $q \sim U(0, 0.01)$ and $\upsilon_t \sim N(0,1)$. $\epsilon_t \sim N(0,1), \xi_t \sim N(0,1)$. No coefficients change. Data removed in MCAR and MAR patterns.

Scenario 1 implements the model described in Section 4.2.3. Scenarios 2 and 3 cover the extensions to this described in Section 4.2.4. In the former the X variables are

independent, and in the latter they are correlated with covariance matrix $\boldsymbol{\eta}$, a randomly generated positive definite covariance matrix. Scenario 4 is intended to test our methods when the assumptions of Section 4.2 are not met. In Scenario 4 the regression errors are not independently and identically distributed, but have an autoregressive structure. In Scenario 5 we consider the situation where the conditional distribution of the covariates x on Z changes but the conditional distribution of Y on X and Z does not. Such a situation constitutes a change under the assumptions of our test, but not under the null hypothesis of the test used by CCA or Imputation. In Scenario 6 we simulate data with no change but linear structure on Z, and remove data according to two missingness mechanisms. This is to investigate the impact that the missingness mechanism has on Type I errors.

Unless specified otherwise, we remove data via a Missing at Random mechanism — the higher the value of each $\boldsymbol{z}_{t,j}$ the more likely x_t is to be missing — and, where $p > 1$, in a monotone missingness pattern, using the `ampute` function in the `mice` package in `R` (Van Buuren and Groothuis-Oudshoorn, 2011). We vary the size of change and the proportion of missing data, and compare our method against two others: Complete Case Analysis — where we delete any cases where $X_{t,i}, i = 1, \ldots, p$ is missing; and imputation — where we impute the whole series once using predictive mean matching from the `mice` package, before running a likelihood ratio test based test for a single change.

Predictive mean matching creates a prediction of the value of the missing observation using a Bayesian process — deriving the posterior distribution of the parameters of the missing data given observed data and previous iterations of imputations. It then draws an imputed value at random from the closest observed values to the predicted value.

For the factored likelihood method, the cost per segment is calculated as in Sections 4.2.3-4.2.4. For Single Imputation and Complete Case Analysis we calculate the

likelihood ratio test statistic as (notation given for univariate Y, X, Z):

$$T_\tau = 2\big[\sum_{t=1}^{n} \log\Big(P(y_t|x_t, z_t, \boldsymbol{\phi}^{(1)})\Big) - \max_\tau \mathcal{L}_\tau\big],$$

where

$$\mathcal{L}_\tau = \sum_{t=1}^{\tau} \max_{\phi^{(1)}} \log\Big(P(y_t|x_t, z_t, \boldsymbol{\phi}^{(1)})\Big) + \sum_{t=\tau+1}^{n} \max_{\phi^{(2)}} \log\Big(P(y_t|x_t, z_t, \boldsymbol{\phi}^{(2)})\Big),$$

and we detect a change if $\hat{\tau} = \text{argmax}_\tau T_\tau$ is greater than the chosen threshold, Γ. We maximise the test for $h \leq \tau \leq n - h$ to avoid estimation of regression parameters on a very small number of observations.

4.3.1 Results

We present a comparison of the performance of the methods in finding changes in a simulated linear model of length n. For Scenarios $1 - 4$ (Tables 4.3.1 and 4.3.2) we present a measure of accuracy, and give results on detection power in Appendix B.2. Similar relative results are seen if we look at accuracy or probability of detection. When there is missing data, Complete Case Analysis is consistently the least accurate of the three methods, and its performance relative to the other methods deteriorates as the amount of missing data increases. Factored Likelihood slightly underperforms Single Imputation and Complete Case Analysis when there is no missing data, but as the proportion of missingness increases then it is more accurate than Single Imputation and much more accurate than Complete Case Analysis. Its detection power is slightly lower than Single Imputation, but the difference is negligible. This is because Factored Likelihood is, unlike the other two methods, searching for a change in two sets of parameters: the regression parameters $Y \sim X, Z$ and additionally the parameters of the regression $X \sim Z$.

We see in Table 4.3.3, that Factored Likelihood detects a change in the relationship

		Scenario 1			Scenario 2		
Missing	Method	S	M	L	S	M	L
0.0	CC	**89.1**	**97.2**	**99.2**	**96.6**	**99.4**	**100.0**
0.0	FL	87.9	97.0	99.1	95.6	**99.4**	**100.0**
0.0	Imputed	**89.1**	**97.2**	**99.2**	**96.6**	**99.4**	**100.0**
0.2	CC	84.6	95.7	97.8	92.1	98.5	99.6
0.2	FL	**88.6**	**96.9**	**98.2**	**94.0**	98.6	**99.9**
0.2	Imputed	87.6	96.5	**98.2**	93.4	**98.7**	**99.9**
0.4	CC	75.3	89.7	95.9	86.1	96.5	98.5
0.4	FL	**85.2**	**96.4**	**98.7**	**94.8**	**99.1**	**99.6**
0.4	Imputed	82.8	95.9	98.6	93.7	99.0	99.5
0.6	CC	61.0	80.1	90.4	71.3	88.1	94.3
0.6	FL	**87.6**	**96.9**	98.8	**92.5**	**98.4**	**99.9**
0.6	Imputed	84.7	95.1	**99.0**	90.3	97.3	99.7
0.8	CC	35.5	56.2	69.7	50.0	69.0	77.4
0.8	FL	**83.7**	**95.0**	**98.9**	**92.9**	**98.4**	**99.6**
0.8	Imputed	81.4	92.2	97.7	89.0	97.5	99.5

Table 4.3.1: Results for Scenarios 1 and 2. Table shows the true positive rate — the percentage of repetitions where a method identifies a changepoint within $+/-10$ of the true change. The threshold for each method was chosen to give a false positive probability of 0.05 based on 1000 simulated data sets with no change. The proportion of data missing is rounded to the nearest 20%. We display results by size of change — the constant by which we have multiplied the slope parameters of the regression ($S = 1.5, M = 1.75, L = 2$). The best results for each scenario by level of missingness and size of change are in bold.

between X and Z. CCA correctly does not as it is not looking to detect such a change. Imputation is also not looking to detect a change but we see an increase in its Type I error because it is imputing from the wrong model due to the presence of a change. Examination of the parameters of the models X given Z and Y given X, Z on either side of the identified changepoint give some indication as to which model the Factored Likelihood method has detected a change in. In Table 4.3.4 we present the percentage of Type I errors by method and by missingness mechanism where Z has a linear relationship with t. The threshold for each method is chosen based on a false positive rate of 5% for the 1000 repetitions of the scenario with 0% and 20% missing. The false positive rate of Imputation increases as the proportion of missingness goes up, and this effect is more marked when the data is MAR. This is because the imputation method,

		Scenario 3			Scenario 4		
Missing	Method	S	M	L	S	M	L
0.0	CC	**88.7**	96.6	**99.5**	**54.0**	**73.9**	**87.8**
0.0	FL	87.6	**96.8**	99.2	51.7	72.4	86.4
0.0	Imputed	**88.7**	96.6	**99.5**	**54.0**	**73.9**	**87.8**
0.2	CC	85.0	94.4	97.9	46.8	68.1	82.7
0.2	FL	**88.0**	**95.3**	**98.5**	49.7	**76.7**	**85.9**
0.2	Imputed	83.8	94.1	98.1	**50.9**	74.6	84.5
0.4	CC	77.3	90.0	95.1	37.7	56.9	75.0
0.4	FL	**83.0**	**94.6**	**97.1**	**50.9**	**72.1**	**85.1**
0.4	Imputed	79.2	92.9	95.6	48.6	70.8	84.3
0.6	CC	67.8	83.3	89.9	22.4	47.9	66.1
0.6	FL	**78.1**	**91.8**	**97.8**	**46.1**	**70.8**	**85.9**
0.6	Imputed	71.4	86.6	93.9	43.7	69.5	84.0
0.8	CC	44.5	60.6	74.0	11.2	29.6	45.3
0.8	FL	**74.4**	**89.9**	**95.5**	**45.4**	**69.4**	**84.5**
0.8	Imputed	62.5	79.5	91.0	41.5	66.0	81.6

Table 4.3.2: Table showing the true positive rate in detecting a change for Scenarios 3 and 4. The threshold for each method was chosen to give a false positive probability of 0.05 based on 1000 simulated data sets with no change. The proportion of data missing is rounded to the nearest 20%. Results are presented as in Table 4.3.1

		Probability of detection			True positive rate		
Missing	Method	S	M	L	S	M	L
0.0	CC	5.20	4.50	5.60	0.1	0.0	0.0
0.0	FL	77.90	97.20	99.40	31.2	58.1	76.6
0.0	Imputed	5.20	4.50	5.60	0.1	0.0	0.0
0.2	CC	5.50	5.40	4.10	0.0	0.0	0.1
0.2	FL	66.50	93.90	99.00	25.9	52.3	72.8
0.2	Imputed	5.30	6.20	6.30	0.0	0.0	0.1
0.4	CC	4.20	6.00	5.70	0.0	0.1	0.0
0.4	FL	56.60	89.20	97.70	17.4	43.5	61.4
0.4	Imputed	6.50	7.00	10.50	0.0	0.4	1.1
0.6	CC	4.60	4.90	5.00	0.0	0.1	0.1
0.6	FL	46.60	85.20	95.20	12.1	35.3	57.6
0.6	Imputed	7.80	14.80	21.90	0.3	1.5	4.0
0.8	CC	3.50	3.50	5.20	0.0	0.2	0.3
0.8	FL	34.10	71.90	89.20	7.5	24.4	44.4
0.8	Imputed	11.00	23.00	40.90	0.9	3.0	10.2

Table 4.3.3: Results for Scenario 5. Table shows the probability of detection and true positive rate of methods in detecting a change in the relationship of X given Z. as in Table 4.3.1

	MAR			MCAR		
Missing	CC	FL	Imputed	CC	FL	Imputed
0.0	6.2	5.1	5.5	5.2	4.2	5.0
0.2	4.5	5.0	5.1	4.7	5.3	4.7
0.4	4.2	4.8	6.1	3.9	4.8	5.9
0.6	3.8	2.5	12.4	3.6	2.6	7.1
0.8	3.6	4.1	22.9	4.0	3.0	14.7

Table 4.3.4: Results for Scenario 6. Table showing the false positive rate of methods when Z has a positive linear relationship with t. The threshold for each method is chosen to give a false positive probability of 0.05 based on the 1000 simulated data sets with no change and either no data missing or 20% of the data removed. Under the Missing at Random mechanism, the probability of x_t being missing increases with the size of z_t.

predictive mean matching, replaces missing values with a random choice from among a set of observations closest to the predicted value of the missing observation. Given the structure of Z, where there are high levels of missingness, the method will often have a poor set of existing observations to pick from. This will be more acute under MAR, where x_t is more likely to be missing the higher the value of z_t, so missingness will be concentrated among large t. As the proportion of data missing increases, Factored Likelihood and CCA show slightly less than 5% false positive rate. This is because, with fewer cases included in their likelihoods, the threshold should be a little lower. The missingness mechanism does not affect the amount of Type I error for either method.

4.4 Data Applications

4.4.1 Application to Financial Data

In this first application we want to assess the impact of increasing levels of missingness on our methods in a controlled fashion. To do this we run our methods on a complete dataset, then on versions of the data set where we have induced missingness artificially in increasing proportions. This allows us to compare the results of running methods on data sets with missingness induced in a known pattern and missingness mechanism

against a standard obtained by running methods on the fully observed data set.

In this first example we consider monthly average European sovereign bond yields (Organization for Economic Cooperation and Development, 2023a,b,c,d,e) between 1999 and 2010. We regress the yield of the UK's 10 year Gilt on a set of European government securities that had a similar credit rating during the period considered. We choose UK debt as the dependent variable in our model, with France the independent variable that is subject to missingness, and Belgium, Germany and the Netherlands as the other independent variables. The model without missing data is:

$$y_t = x_t\beta_1 + \boldsymbol{z}_t\beta_2 + \epsilon_t, \tag{4.4.1}$$

where y_t represents the average UK 10 year Gilt yield for month t, x_t is France's 10 year government bond yield observed at month t and $\boldsymbol{z}_t$ is a vector of length 3 containing the 10 year government bond yield of Belgium, Germany and the Netherlands for month t, ϵ_t is a normally distributed error term for month t. The data are plotted, with detected changepoints, in Figure 4.4.1. The data clearly exhibit non-stationary behaviour, with yields increasing and decreasing over time, however in our model we treat $\boldsymbol{Z}$ as observed. As our changepoint detection method is simply looking to detect a change in the relationship between observed $\boldsymbol{Z}$ and $\boldsymbol{x}$, and in the relationship between $\boldsymbol{y}$ given $\boldsymbol{x}, \boldsymbol{Z}$, we do not need to adjust the analysis to account for non-stationarity.

We follow Tickle et al. (2021) in using Wild Binary Segmentation (Fryzlewicz, 2014) to adapt a method that searches for a single changepoint to searching for multiple changepoints. We first consider the complete data set. We search for a maximum of two changepoints using Wild Binary Segmentation with 5000 intervals in combination with our factored likelihood test statistic, as well as using a likelihood test statistic based on linear regression on a complete dataset (as given in Sections 4.2 and 4.3). For all methods we remove the first and last h observed entries of the cost function (as described in Section 4.3). In this section, because of the small segment sizes necessitated

by Wild Binary Segmentation, we set $h = 5$. These methods select a change at February 2000 and May 2003. In Figure 4.4.1 we see a clear change in the relationship between the UK's and the other bonds at these points. Up until February 2000 the UK Gilt tracks slightly above the Eurozone bonds, then it performs almost completely in line with them until May 2003, when the yield of the UK Gilt sharply increases.

Figure 4.4.1: Bond yields of selected European government bonds. Detected changepoints shown with dashed line.

To check the fit of the model, we fit a linear regression of the 10 year UK Gilt on the French, Belgian, Dutch and German 10 year government bond yields, as per (4.4.1) for the series between February 2000 and May 2003. Model fitting diagnostics are depicted in Figure 4.4.1. The data appears to be approximately normally distributed with some slight deviation away. The residual order plot and ACF suggest a degree of autocorrelation in the residuals. Although this model is not a perfect fit, it is considered satisfactory since we treat the changepoints identified on the complete dataset as known, and are looking to see to what extent introducing differing levels of missing data affects

our algorithms' ability to pick up the signal of a change.

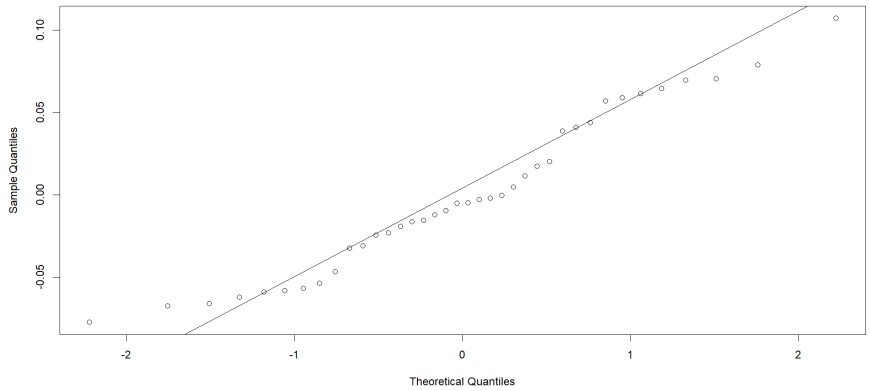


Figure 4.4.2: Normal Quantile-Quantile plot of residuals

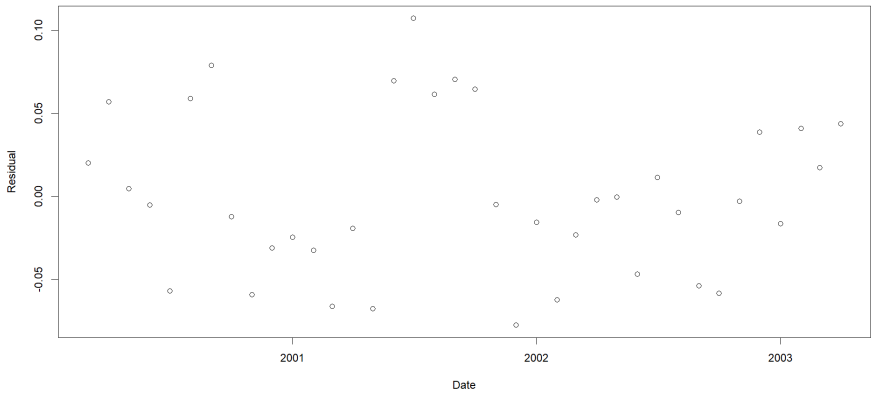


Figure 4.4.3: Residuals plotted against time

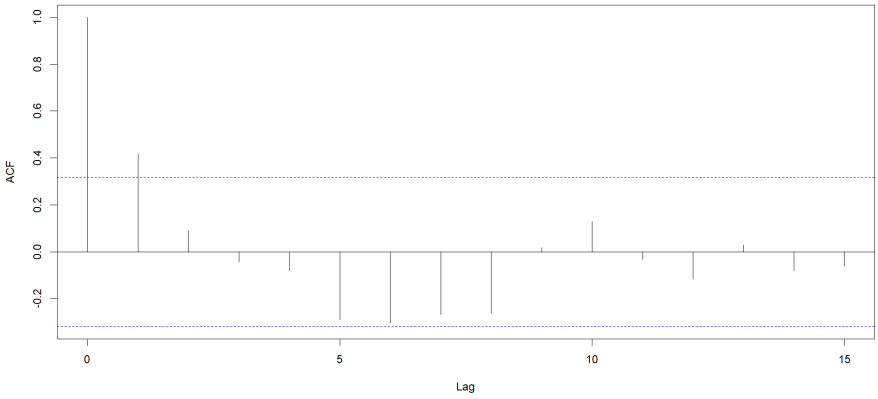


Figure 4.4.4: Auto-Correlation Function of residuals

Figure 4.4.5: Diagnostic plots of residuals of fitted linear regression model of UK Gilts on selected European sovereign bonds between February 2000 and May 2003.

Having identified two changepoints in the complete dataset — in February 2000 and

May 2003 — we then remove differing levels of data — 10%, 20% and 30% — according to a Missing at Random mechanism using the `ampute` function in `R`'s `mice` package, which induces missingness in data sets according to selected patterns and mechanisms. We repeat this 100 times for each missingness scenario. For each repetition we run the three methods on the same set of amputed data, and over the same set of 5000 random intervals for Wild Binary Segmentation. We give the results of this for each method in Table 4.4.1. Factored Likelihood performs more strongly than the others at all three levels of missingness. CCA and Factored Likelihood successfully identify fewer changes as the proportion of missingness increases. CCA performs poorly, with just 62% of the changepoints identified when 10% of the data is missing, and its performance drops sharply as missingness increases, tailing off to just 17% identified when 30% of the data is missing. Imputation is consistently less accurate than Factored Likelihood, though the difference between the two becomes smaller as the proportion of missing data increases, since Imputation's performance does not drop off by as much as the amount of missing data increases. This could be because there is a change in the relationship between France and the other European Union nations' bonds (a change in X given Z — see Scenario 5 in Section 4.3) as well as a change in the relationship between the UK's bonds and the other countries'.

Method	10% missing	20% missing	30% missing
Complete Cases	62%	33%	17%
Factored Likelihood	96%	91%	85%
Single Imputation	84.5%	80.5%	82%

Table 4.4.1: Accuracy of methods on detecting changepoints in February 2000 and May 2003 (previously identified by all methods when no data is missing) in sovereign bond yield data with increasing levels of missingness. There are 100 repetitions. If a method correctly identifies one change but not the other this is counted as 50% accuracy for that repetition.

4.4.2 Application to Wind Farm Data

Having considered changepoint detection in a regression relationship with artificially induced missingness in one of the independent variables, we turn to an example of an incomplete dataset. We consider a linear model between sensor measurements of ambient wind speed, grid inverter temperature and grid power measured at 10 minute intervals in a wind turbine located in an offshore wind farm in the North Sea. Figure 4.4.6 shows ambient wind speed and grid inverter (GI phase 1) temperature plotted against date for the period in consideration. There is typically a positive relationship between wind speed and grid inverter temperature and grid power — except at wind speeds close to the cut-off wind speeds (low and high), when temperature and power reach maximum and minimum levels. This wind turbine has repeated shutdowns in July 2023 — it exhibits 'cycling' behaviour, where at each start-up a quick rise in temperature triggers an automated shutdown. However, we see excessively high values in grid inverter temperature as early as July 2022.

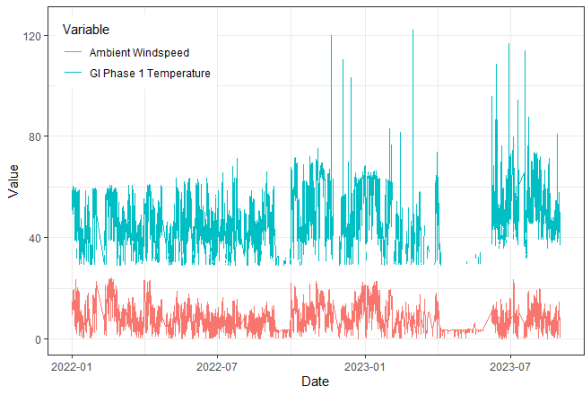


Figure 4.4.6: Changing relationship between ambient wind speed and grid inverter in an offshore wind turbine. There is a positive relationship between Ambient Wind Speed and GI Phase 1 Temperature, but over the period under consideration temperature hits higher and higher levels over time, while wind speed does not increase commensurately.

There is considerable interest in detecting changes or anomalous behaviour in this relationship with the purpose of condition monitoring — detecting a problem before the equipment breaks. Maintenance represents a significant proportion of the total costs of

a wind farm. Due to their offshore location, transit of maintenance staff to the wind turbine is constrained by weather conditions to ensure safety. After filtering out values where the system is not running in steady state (for example, if wind speed is too high or low, or the turbine is shut down for maintenance), grid power and ambient wind speed are completely observed, and grid inverter temperature has some missing values. We regress grid inverter temperature and ambient wind speed on grid power:

$$y_t = x_t\beta_1 + z_t\beta_2 + \epsilon_t,$$

where y_t represents grid power at time t, x_t and z_t represent grid inverter temperature and ambient wind speed, respectively, at time t, and ϵ_t is a normally distributed error term.

Wind speed time series typically exhibit seasonality (Stopa, 2021), but in this model we treat $\boldsymbol{z}$ as observed and are interested in either a change in the linear relationship between $\boldsymbol{y}$, with $\boldsymbol{x}$ and $\boldsymbol{z}$, or between $\boldsymbol{x}$ and $\boldsymbol{z}$, hence we do not need to model the seasonality in $\boldsymbol{z}$. Additionally we are seeking to detect a known change — before the repeated shutdowns in July 2023, so, even if this is a relatively simplistic model for our data, the signal for a change is still apparent.

We run the factored likelihood method to detect a single change in the regression relationship in the series between January 1, 2022 and August 31, 2023. The method detects a change on April 9, 2022 (depicted in Figure 4.4.7).

Plotting wind speed versus grid inverter temperature in Figure 4.4.8, we see a second distinct curve, plotted in red, emerging more and more clearly after the changepoint. Lower wind speeds are creating higher and higher temperatures in the grid inverter.

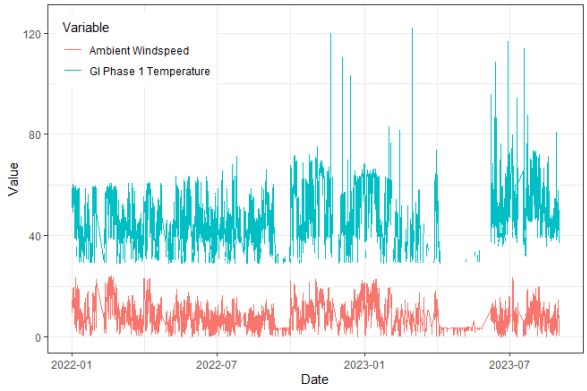
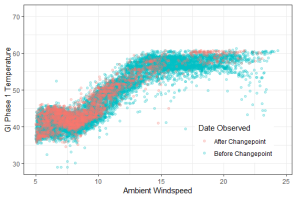
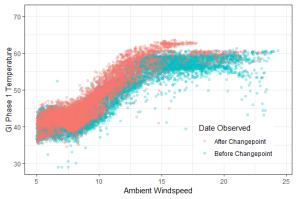


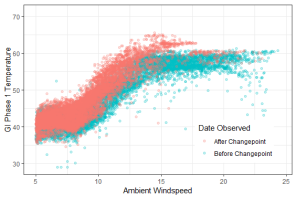
Figure 4.4.7: We show an identified change in the relationship between GI Phase 1 Temperature and Ambient Wind Speed. The changepoint is identified as on April 9, 2022. After this date we see that GI Phase 1 Temperature increases a disproportionate amount given Wind Speed, compared with its behaviour prior to this date.



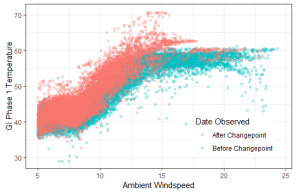
(a) One month after changepoint



(b) Two months after changepoint



(c) Three months after changepoint



(d) Four months after changepoint

Figure 4.4.8: Power curves showing the relationshop between Ambient Windspeed and GI Phase 1 Temperature over time. Plotted in blue are observations occurring before the identified changepoint on April 9, 2022, and in red afterwards. It is clear that a second curve emerges after the identified changepoint — one where for an observed Ambient Windspeed the GI Phase 1 Temperature is higher than it was before the change point — substantiating a change in the relationship between the two variables.

4.5 Discussion

We have presented a method for detecting changepoints in a linear regression model where some of the dependent variables are subject to missingness in certain patterns.

This method is more accurate than whole data methods when the amount of missing data present is sizeable, and it is comparable in accuracy and detection power with no or low amounts of missingness. It is substantially better — in terms of detection power and accuracy — than Complete Case Analysis. By avoiding the need to explicitly impute the missing data, it is less computationally intensive than Imputation. Additionally, our method detects changes that could include a change in X given Z. This is particularly advantageous over Imputation, which, as it is designed to detect changes just in Y given X, Z, can have inflated Type I errors if the distribution of X given Z changes. This problem becomes more acute as the amount of missing data increases. As discussed in Section 4.3, it is possible to compare parameter estimates for the two models before and after a changepoint identified by the Factored Likelihood approach, to indicate in which model the method has detected a change. We have successfully demonstrated our method on two real world data sets: one of bond yields, where we removed increasing amounts of the data in a Missing at Random pattern, and one of sensor measurements from an offshore wind farm where missingness was already present in the data. We have also shown that our method can be extended to detecting more than one changepoint via Wild Binary Segmentation.

Further work includes extending this to more complex linear models, such as Vector Autoregressive models or relaxing the assumption of Gaussian errors. Another fruitful area of work is developing extensions based on specific Missing Not at Random mechanisms or non-monotone missingness patterns.

Chapter 5

Detecting interpolation errors in infant mortality counts in 20th Century England and Wales

5.1 Introduction

In this chapter we consider a set of data collected over a critical period in the history of health in the United Kingdom: infant mortality rates by local government district in England and Wales between 1911 and 1973. This is a key period in the history of infant mortality decline: at the start there were, on average, 130 deaths per thousand births in England. This had dropped to 17 by the end (GB Historical GIS and University of Portsmouth, 2024b,c). For Wales the rate is 135 and 16, respectively (GB Historical GIS and University of Portsmouth, 2024e,f). Annual rates are available for every local government district during the period. But, despite the scale and the rich level of detail of the data available, how and why infant mortality declined so precipitously is not well understood. This is because data, as it exists, is difficult to analyse: a large number of changes were made to local government district boundaries over the years of interest,

meaning that — for many districts — we cannot compare like with like over time. Scholars deal with this problem by aggregating (Lee, 1991; Woods et al., 1988, 1989; Winter, 1982), but these analyses miss the finer scale variations in the pattern of decline in infant mortality, such as differences between rural and urban areas. Alternatively, they have focused on fine detail but have been restricted by geographic area (James et al., 2001) or time (Congdon and Southall, 2004).

Figure 5.1.1 shows the infant mortality rate (the number of deaths per 1000 births) for a random sample of 100 districts in our dataset. It is clear that infant mortality declines over the period in question, but how this decline occurs varies enormously, particularly in the first 30-40 years. For some local government districts, infant mortality rates begin the period very high and drop steeply. For others, infant mortality rates are low at the beginning of the period and the decline is less marked. It is also clear that no single intervention, such as the establishment of a National Health Service in 1948, is responsible for the decline.

Our intention is to be able to understand this data set at a fine scale over the whole of the time period in question. As an example of the kind of analysis that can be done with fine scale data we cluster individual series based on functional Principal Components Analysis (fPCA) — identifying areas where infant mortality evolves in a similar fashion over the period. Studies that have treated mortality curves as functional data include Erbas et al. (2007) and Hyndman and Ullah (2007). Those that treat them as functional data in order to cluster them include Stefanucci and Mazzuco (2022) and Léger and Mazzuco (2021). Clustering historical infant mortality rates according to shape has been performed on 19th century infant mortality rates observed every 10 years, using latent trajectory analysis rather than a functional data approach (Atkinson et al., 2017b,a). it is clear from Figure 5.1.1, which depicts only a fraction of the infant mortality series, that it is difficult to distinguish trends at the fine scale, that being able to group the series according to common behaviour would enhance the understanding

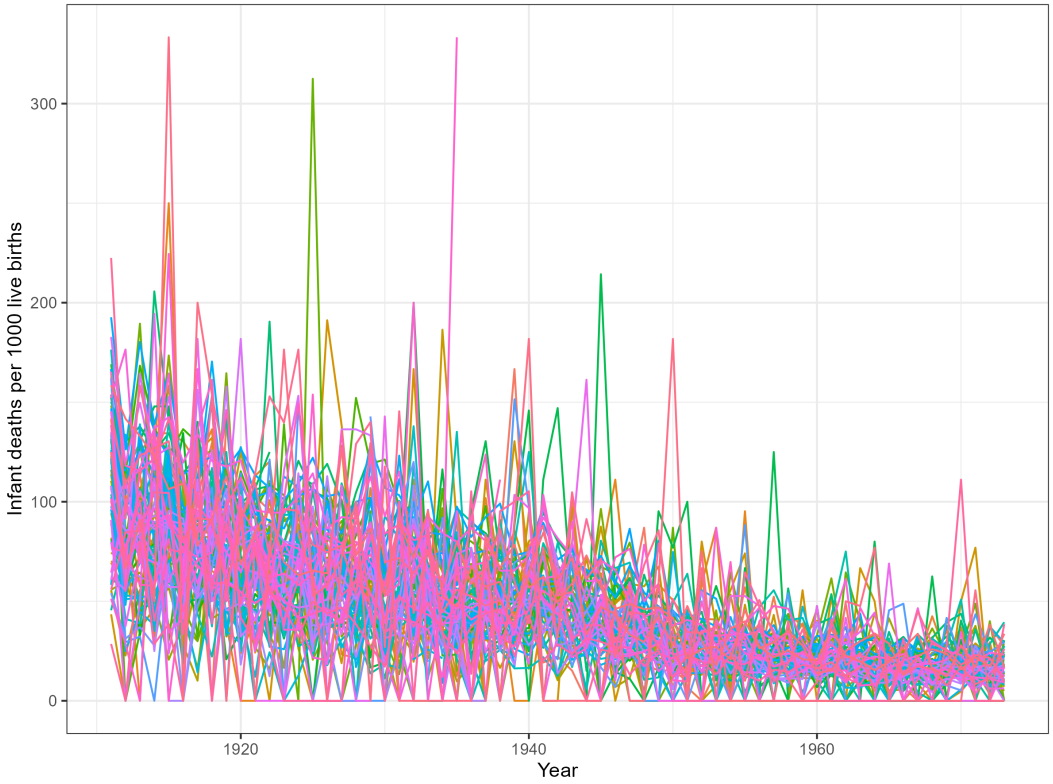


Figure 5.1.1: Infant mortality rate England and Wales, based on the raw data. Random sample of 100 local government districts depicted (Southall and Mooney, 2022).

of this data set.

Before we can apply fPCA to our data to better understand the nuances of the decline in infant mortality over the 20th century, we first need to address the problems caused by the numerous boundary changes afflicting our data set. We will use interpolation to adjust those observations impacted by boundary changes. Interpolation estimates predictions for those areas and years affected, enabling the construction of a time series over a consistent set of boundaries covering the full period 1911-1973. We use an interpolation approach that assumes populations are distributed evenly over area (Goodchild, 1980), and introduce a process to detect situations where the interpolation is poor. We interpolate the raw data: annual counts of infant mortality — instances where a child dies before its first birthday — and annual counts of birth by local government district, rather than the infant mortality rate derived from these.

This is because we want to apply our error detection process to the series of raw and interpolated counts, since it is likely that an interpolation error in the series of deaths will also occur in the series of births, then the error may be harder to detect in the proportion of infant deaths to births than in one of the series of counts individually. For those series in which we detect an error, as a safest approach we will merge areas that have been affected by an interpolation error, rather than trying to improve the interpolation by using a more sophisticated model (although that option is available for future analyses).

Investigations into areal interpolation errors include Liu and Liu (2008), Hawley and Moellering (2005), Gregory (2002a) and Fisher and Langford (1996). However, there is no agreed best method of interpolating. The performance of methods can vary according to the type of interpolation, the data available and the accuracy assessment measure. To the best of our knowledge the only work extant that looks to detect and assess interpolation errors — in a data set similar to ours — by treating them as data observed over time, as well as space, is Gregory and Ell (2006). This work considers decennial population counts for local government districts, which can be approximated as normally distributed data. They detected changes by looking for outliers in the differenced data from one time period to the next. By looking at differences only over consecutive data, this method loses power by ignoring the long-term impact of a change, which will impact all future observations. It is also not suitable for count data that does not fit the criterion for a normal approximation to Poisson data.

In this chapter we introduce a changepoint detection method to find the instances where interpolation has performed poorly on our data. An interpolation error will show as an abrupt change in the series of counts of annual infant deaths by local government district at the same time as a known boundary change in that district. Our data is count data with downwards trend. It includes instances of low or zero counts (Figure 5.1.2) and is therefore not suitable for approximation to normal. We also directly model

the trend within the data, by testing for an abrupt change in the presence of a trend, which makes it more robust than standard change-point detection methods that would ignore any trend. It can can be used directly on the data in our study.

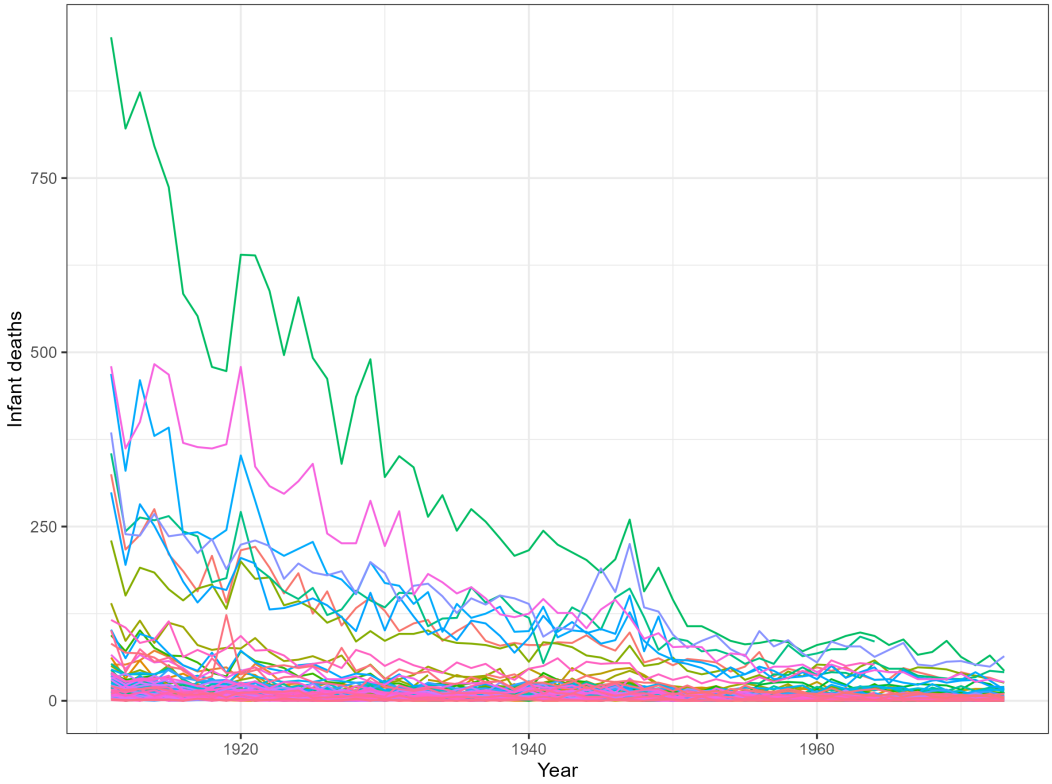


Figure 5.1.2: Counts of infant deaths, England and Wales, based on the raw data. Random sample of 200 local government districts depicted (Southall and Mooney, 2022).

Changepoint detection methods are a group of algorithms that look to identify one or more changes in a specified statistical property of a sequence, or series, of data. See, for example Truong et al. (2020) for an overview of the extensive literature in this area. Our data is count data, which is most naturally modelled by a Poisson distribution. We are interested in detecting interpolation errors which would correspond to an abrupt change in the mean of the data, while allowing for gradual changes in the mean due to other factors — such as public health improvements — which we model as trend. We introduce a method that searches for a single abrupt change in a sequence of Poisson distributed data that is subject to trend. Pein (2021) is among recent work

to explore the detection of an abrupt change in data with a slowly evolving trend, but the method is not suitable for count data. Methods that search for a changepoint in Poisson distributed data, such as Paparas et al. (2023) and Samuel and Pignatjello Jr (1998), either ignore trend or detect changes that would be due to changes in the trend, or the introduction of trend (Perry et al., 2007). Loader (1992) consider the detection of an abrupt change in a Poisson process with trend. This method is not suited to our data since it is developed for modelling arrival times not the number of occurrences in a sequence of time periods of equal length.

In the rest of this chapter we give an overview of the data under consideration and the interpolation method applied to it (Section 5.2) ; then we introduce our changepoint detection method (Section 5.3.1) and give an overview of fPCA (Section 5.3.2). We return to the application in Section 5.4.1 where we use our changepoint detection method to identify errors in the interpolated time series of infant death counts, and, in Section 5.4.2, we show that correcting interpolation errors affects the clustering of the infant mortality curves, and discuss some potential insights gleaned from looking at the data at this fine scale. We conclude in Section 5.5.

5.2 Data and Preliminaries

As mentioned in Section 5.1, the data we consider consists of annual counts of births and infant deaths recorded for Local Government Districts in England and Wales between 1911 and 1973. These were published as the Annual Report of the Registrar General $1911-1920$, and thereafter as the Registrar General's Statistical Review of England and Wales. These have largely been digitised, but in some cases input manually, for local government districts of England and Wales by the Great Britain Historical Database and are held in a vital statistics data set that is made available by the UK Data Service (Southall and Mooney, 2022). From these data we are able to calculate the infant

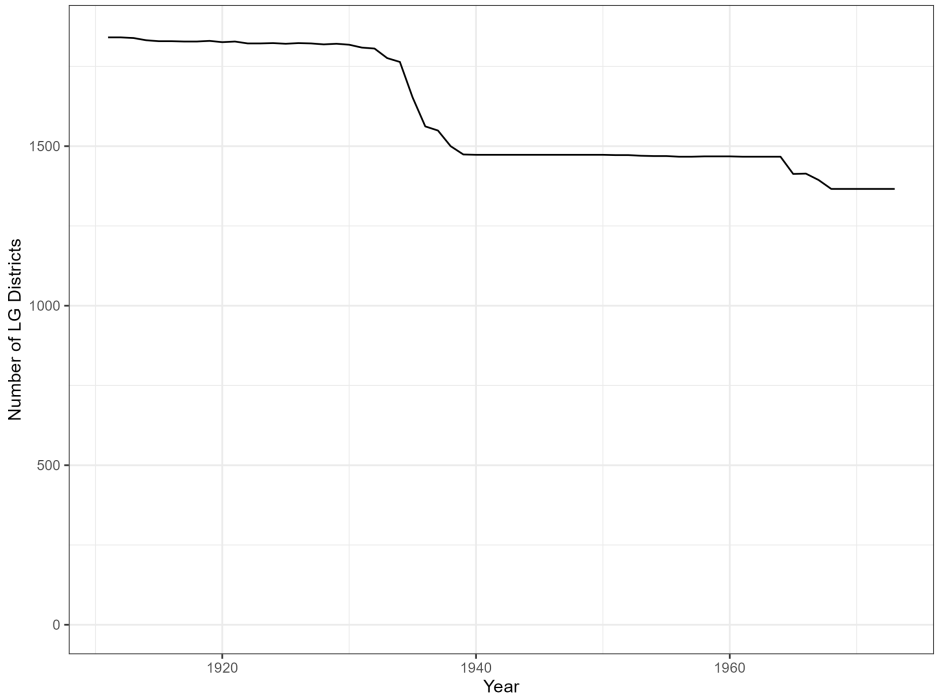


Figure 5.2.1: Counts of LGD names in England and Wales per year (raw data) (Southall and Mooney, 2022).

mortality rate per births in each district. However, this is not straightforward. The boundaries of the approximately $1300 - 1800$ geographical units for which annual data is collected (Figure 5.2.1) undergo considerable change over the period: boundaries are adjusted, districts are created and others abolished. Boundary changes occurring between 1930 and 1973 are held by the GBHDB and published in their vital statistics dataset. We refer the reader to Gregory et al. (2002) and Gregory (2002b) for more detail on the the construction of these data collections.

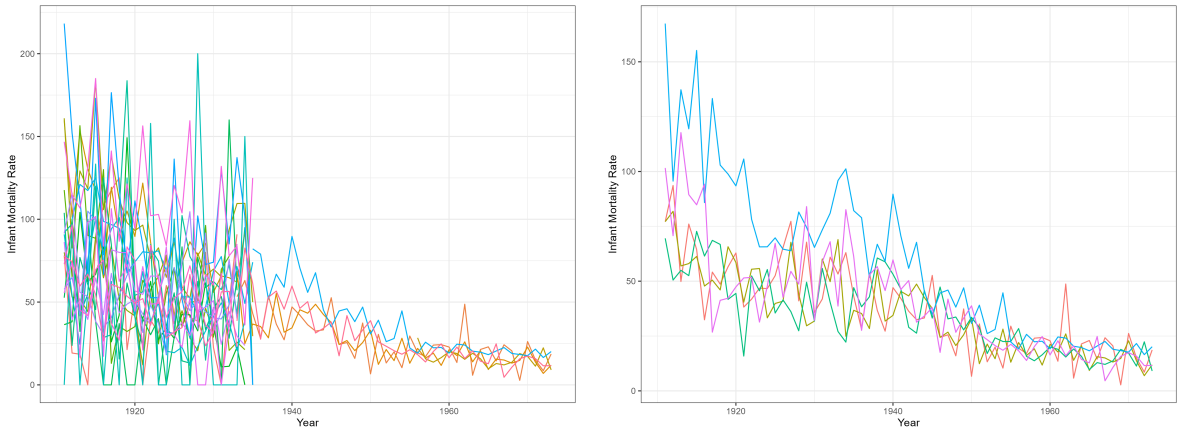
Knowledge of boundary change can help us firstly with interpolation, and secondly to identify a changepoint that is potentially related to a boundary change. The second source of data used in this study are GIS files of local government district boundaries (Southall et al., 2024). These are published for each census year in the period of interest: 1911, 1921, 1931, 1951, 1961 and 1971. Using these files and the database of annual infant mortality statistics by local government districts, we can interpolate the data from boundaries that no longer exist onto the 1971 boundaries.

While the available data covers all of England and Wales, we focus on two trial areas in this chapter because of the large number of districts in the data set. These are the local government districts in the county of Northumberland and in East and West Sussex. We choose these areas because the county boundary of Northumberland and the combined county boundaries of East and West Sussex do not change between 1911 and 1973, allowing us to focus on the changes to the local government district boundaries (GB Historical GIS and University of Portsmouth, 2024a,g,d). We aim to create a consistent time series for all local government districts from $1911-1973$ based on the 1971 boundaries.

The process of interpolating the data involves estimating predictions for districts which have undergone boundary changes, for the years before that boundary change. For example, if a district undergoes single boundary change over the period, say, in 1935, the observations for the years $1911-1934$ have been collected for different geographical areas to the 1971 boundaries, while those from 1935 to 1973 will have been collected on the 1971 boundaries. Therefore the observations representing the 1971 district for the years before the change need to be estimated.

In this work, we employ one of the simplest interpolation methods — areal weighted interpiolation (Goodchild, 1980), which we implement in the `R` package `areal` (Prener and Revord, 2019). More sophisticated interpolation methods exist but these usually involve using extra data to inform the estimates (see, for example, Gregory and Ell (2005) and Hawley and Moellering (2005) for an overview). The size and complexity of our data set, and the paucity of ancillary data available — which is, additionally, often inconsistently available over our time period — mean that these more complex methods are not straightforwardly appropriate for our data. We know that areal interpolation is likely to introduce some errors because of its assumption that populations are evenly spread over area — which is not realistic for England and Wales in the 20th century. This is why it is necessary to search the interpolated series for errors after the

interpolation step. We give a brief overview of the areal weighting interpolation process below. We give more details specific to the application in Appendix Section C.1.



(a) Raw (b) Interpolated

Figure 5.2.2: Raw and interpolated IM rates for local government districts in Sussex or Northumberland that are created or abolished 1911 − 1973 (Southall and Mooney, 2022).

We define those areas and years for which we require estimates as target districts, following the notation of Gregory and Ell (2006) and adding notation for time. The estimate for the jth target district in the tth year is denoted $\hat{x}^T_{t,j}$, $j = 1, \ldots, d'$, where d' is the number of target districts that are known to undergo at least one boundary change and $t = 1911, \ldots, 1973$. For each year we also have a set of observed counts from source districts. These are the areas for which we have observations collected on known boundaries. The observed counts for the ith source area in the tth year is denoted $x^S_{t,i}$, $i = 1, \ldots, s_t$, where s_t is the number of source districts in the tth year under consideration. For each Target Area, one or more Source Districts will overlap with its boundaries, and we will use this subset of observations, together with knowledge of how much of the area of a Source District overlaps with the Target Area, to calculate an estimate for the Target Area. Let A^T_j be the region of the jth target district and $|A^T_j|$ its area. Similarly, let $A^S_{t,i}$ be the region of the ith source district in the tth year and $|A^S_{t,i}|$ its area. Here $i = 1, \ldots, s_{t,j}$ and $s_{t,j}$ is the number of source districts contributing to the jth target area in the tth year. Then if $|A^S_{t,i} \cap A^T_j|$ is the area of intersection

between the ith source and jth target district in year t,

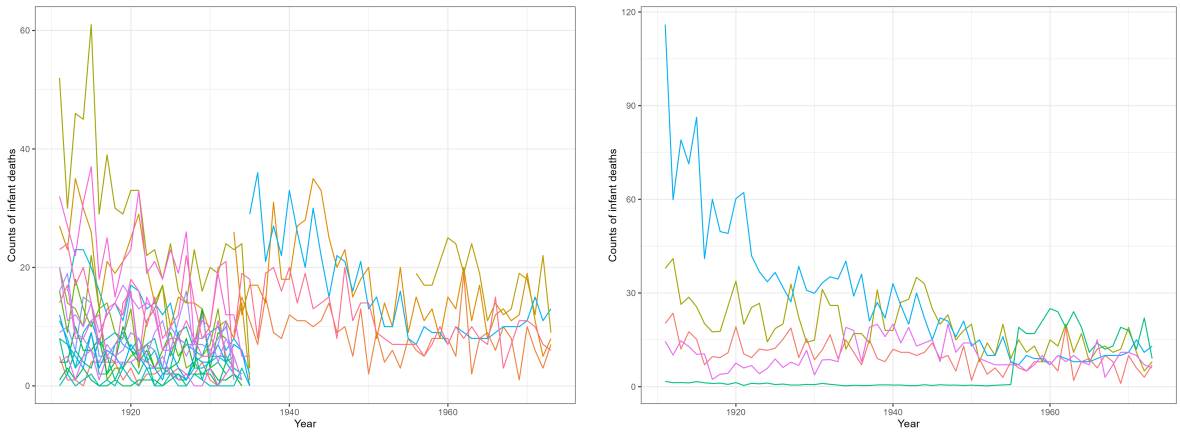
$$\hat{x}^T_{t,j} = \sum_{i=1}^{s_{t,j}} \frac{|A^S_{t,i}| \cap A^T_j|}{|A^S_{t,i}|} x^S_{t,i}.$$

We apply areal weighted interpolation to the raw counts of births per local government district per year, and the counts of infant deaths per local government district per year. From this we calculate the annual infant mortality rate per 1000 live births. In Figure 5.2.2 we plot the interpolated infant mortality rates next to the raw data for areas that were created during the period, as well as those areas affected by their creation.

5.3 Methods

In this chapter we use existing and new methods to perform our analysis. We first introduce a changepoint method developed for this application, before giving an overview of fPCA, which we use to cluster the infant mortality curves.

5.3.1 Changepoint detection



(a) Raw (b) Interpolated

Figure 5.3.1: Raw and interpolated counts of infant deaths in selected areas 1911 – 1973 (Southall and Mooney, 2022).

Although the infant mortality rate is of primary interest for historical analysis, we apply a changepoint detection technique to the interpolated counts of infant deaths. These are plotted for Northumberland and Sussex in Figure 5.3.1 alongside the uninterpolated counts. We consider the count data, rather than the infant mortality rate, because an interpolation error in the counts of infant death is likely to be present also in the counts of births, making it harder to detect an error in the infant mortality rate. We expect an interpolation error to manifest as a sharp jump upwards or downwards in the series of infant death counts at or near the year of a known boundary change.

A Poisson distribution is a common way to model count data. To check that this model is indeed appropriate for our data set we fit a Poisson regression with a time varying parameter to the series in our data set that do not undergo a change, and test to see whether a key property of the Poisson distribution — that of equal mean and variance — holds for most of them. We then test the models for overdispersion and underdispersion (Cameron and Trivedi, 1990) using the `AER` package in `R` (Kleiber et al., 2020). At the 5% level of significance, we reject the null hypothesis of no overdispersion for 6 of 16 series, and for 2 series at the 1% level of significance. For two of the series we reject the null hypothesis of no under-dispersion at the 5% significance level and for one at the 1% level.

We introduce a Poisson distribution with a time varying parameter, β_t, and a time invariant parameter, α, where the probability of observing count x in an interval t is:

$$P(X_t = x_t) = \frac{\lambda_t^x e^{-\lambda_t}}{x_t!},$$

where $\log \lambda_t = \alpha + \beta t$.

We define a single changepoint in this process at $t = \tau$ as:

$$\lambda_t = \begin{cases} e^{(\alpha+\beta t)} & \text{if } t \leq \tau \ , \\ e^{(\alpha^*+\beta t)} & \text{if } t > \tau \ . \end{cases}$$

We then give a likelihood ratio based statistic, T_τ, for detecting a single change at $t = \tau$ as based on the difference between the maximised log likelihood for a change versus that for no change.

Let $\mathcal{L}_n$ be the likelihood for the model with no change (with additive constants dropped):

$$\mathcal{L}_n = -\sum_{t=1}^{n} e^{(\alpha+\beta t)} + \sum_{t=1}^{n} (\alpha + \beta t) x_t,$$

and $\mathcal{L}_\tau$ be the maximised log likelihood for a change at τ (again, dropping additive constants):

$$\begin{aligned} \mathcal{L}_\tau = &-\sum_{t=1}^{\tau} e^{(\alpha+\beta t)} + \sum_{t=1}^{\tau} (\alpha + \beta t) x_t \\ &- \sum_{t=\tau+1}^{n} e^{(\alpha^*+\beta t)} + \sum_{t=\tau+1}^{n} (\alpha^* + \beta t) x_t. \end{aligned}$$

Then T_τ is

$$T_\tau = 2\big[\max_{\alpha,\alpha^*,\beta} \mathcal{L}_\tau - \max_{\alpha,\beta} \mathcal{L}_n \big],$$

This is calculated using the `glm` function in `R`. The proposed candidate for a changepoint, $\hat{\tau} = \arg\max_\tau T_\tau$. Standard practice in the changepoint detection literature is to accept a change at $\hat{\tau}$ if $T_\tau > c$ where c is some pre-determined constant. The thresholds for a given significance level can be calculated using simulation:for instance, by simulating a number of data sets without a change and selecting a threshold that gives

a chosen false positive rate for those data sets. However, for our use, we will require subsequent analysis for each identified change to see whether it is likely due to interpolation error after a boundary change, or is due to other historical effects. Thus we take a different approach of using the test statistic to rank regions from those with the strongest evidence of a change, so that an analyst can identify regions most important to investigate further.

We demonstrate this method in a variety of simulated scenarios. These scenarios and accuracy measures reported over 1000 repetitions of them are described in the Appendix, Section C.2.

5.3.2 Functional Principal Components Analysis

As mentioned in Section 5.1, we want to study our data at fine scale. An example of the kind of analysis that can be done using this data is clustering the data based on its functional principal components.

To do this we treat the interpolated infant mortality curves as functional data. Functional data analysis views the underlying signal as being a continuous function of time that we are observing, with some error, at discrete time points. It uses assumptions about the continuity and smoothness of the function to help estimate it. In our case we conceive of the infant mortality curves as ones that are more or less continuously evolving, but are observed once a year only, due to the data collection method. We then use FDA to estimate the function that represents the evolving infant mortality rate for each district.

Following Ramsay and Silverman (2005) we introduce a real-valued function $u_j(t)$, $j = 1, \ldots, d$, which is a member of L^2 and is defined on the interval $\mathcal{T}$: the years $1911 - 1973$. We can consider our observations of the infant mortality rate of the jth district in year t, $y_{t,j}$ as glimpses of this continuous function (with some measurement

error, denoted $\epsilon_{t,j}$):

$$y_{t,j} = u_j(t) + \epsilon_{t,j}.$$

The idea of functional PCA is that we can write the function $u_j(t)$ in terms of a set of basis functions, and view the variability in the function as due to randomness in the coefficients. Given the variability in the functions across districts, there will be specific choices of basis functions (formally defined as the eigenfunctions, ξ, of the covariance operator of $u(t)$) that have good properties, including the independence of the coefficients for any given district.

$$\hat{u}_j(t) = \mu(t) + \sum_{p=1}^{p} f_{j,p}\xi_p(t), \tag{5.3.1}$$

where ξ_p is the eigenfunction associated with the pth principal component and $f_{j,p}$ is the functional principal component score for the jth series. The set of eigenfunctions are orthonormal.

Functional PCA estimates the covariance operator of $u(t)$ from the data and thus gets estimates of these eigenfunctions. It then estimates the coefficients for each local government district. In our case we proceed as follows. We smooth each series by fitting a smoothed curve to each y_j — the observed infant mortality rate for the jth district — using 12 quadratic B-splines basis functions. These basis functions are denoted ϕ_k, $\phi = 1, \ldots, 12$. By smoothing each series each $u_j(t)$ can be expressed as a known linear combination of these functions:

$$\hat{u}_j(t) = \sum_{1}^{12} c_{j,k}\phi_k(t). \tag{5.3.2}$$

Then the variance-covariance function,

$$\nu(s,t) = \frac{1}{d}\sum_{j=1}^{d} u_j(s)u_j(t),$$

can be estimated from (5.3.2). The estimates of the eigenfunctions and their corresponding principal component scores in (5.3.1) can then be estimated through matrix algebra (see, for example, Ramsay and Silverman (2005, pp 37-58) for more details).

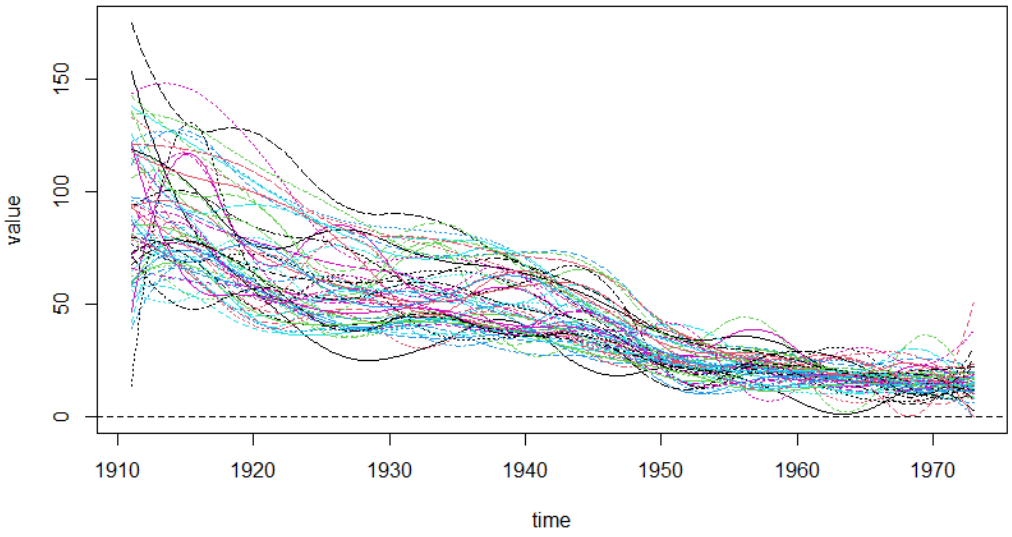


Figure 5.3.2: The smoothed series of infant mortality data. We use 12 quadratic B-splines basis functions to smooth

The fPCA described in this chapter is carried out using the `fda` package (Ramsay et al., 2024) in `R`. We cluster the data based on the principal component scores, $f_{j,p}$ in (5.3.1). Our clustering summarises the data for each region in terms of the coefficients associated with the p most important basis vectors, and clusters the data based on this summary. So data in the same cluster will have similar values for the coefficients for each of these basis functions. We show the first four principal eigenfunctions of the data (after adjusting for interpolation errors) in Figure 5.4.4.

We apply an agglomerative hierarchical clustering method — Unweighted Pair

Group Method with Arithmetic Mean (UPGMA, Sokal and Michener (1958)) using the `hclust` function in base `R`— to the first four Functional principal components scores of the data. These account for 90% of the variation about the mean curves. We determine clustering method and an optimal number of clusters by examining connectivity (Handl et al., 2005), Dunn index (Dunn, 1974) and Silhouette (Rousseeuw, 1987) scores using the `clValid` package (Brock et al., 2008) in `R` across UPGMA and k-means (Hartigan and Wong, 1979). We gain the top scores for UPGMA across the methods. Connectivity and Silhouette recommend two clusters, while the Dunn index has optimal values at six to nine clusters, when we cluster after correcting for data and interpolation errors. Taking this with our knowledge of the data, and that it is likely that dividing our 1500 or so local government districts into a large number of clusters will be most helpful to historians, we select the maximum number with an optimal Dunn index score — nine — as the number of clusters. In Sections and 5.4.2 we discuss the PCA functions and the clusters, including how the clustering is affected by treating areas affected interpolation errors found in 5.4.1.

5.4 Analysis

In this section we apply the changepoint detection method described in Section 5.3.1 to the series of infant death counts, and identify several series affected by errors. We cluster the data based on their FPCA scores, as described in Section 5.4.2 and show that analysing the data at a fine scale reveals interesting aspects which remain hidden when looking at this data at an aggregated scale. We cluster the data before and after addressing the data and interpolation errors. We show that adjusting for errors affects the clustering.

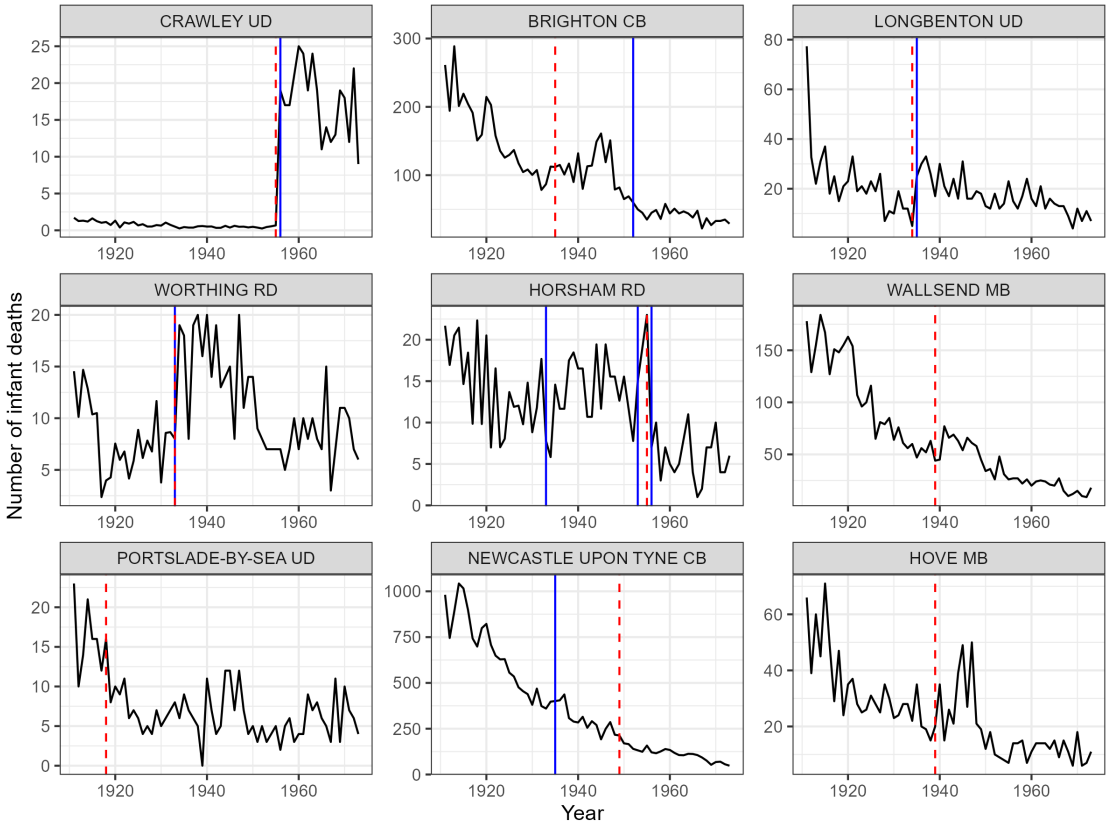


Figure 5.4.1: Results of first application of changepoint detection. Series of infant death counts with the most evidence for a changepoint, in descending order. Blue vertical line depicts the year of a known boundary change. Red dashed line depicts the date of change identified by changepoint detection.

5.4.1 Identifying Errors

We apply our changepoint detection method to search each series of the interpolated counts of infant deaths for Northumberland and Sussex for a single abrupt change, while allowing for trend. As the interpolation introduces non-whole numbers, and our method is designed for count data, we round interpolated observations to the nearest whole number. The years $1940 - 1944$ are excluded from the changepoint detection analysis, given that these represent the complete years of the Second World War, which we expect to cause considerable disruption to the time series. Moreover, we know that no boundary changes were made to the local government districts in Northumberland and Sussex in that period. In Figure 5.4.1 we present line plots of the interpolated series

of infant mortality rates for the nine districts with the most evidence for a change in descending order of evidence. We also show the year of changepoints identified by our method, together with known boundary changes. As discussed in Section 5.1, we consider likely changes to be those where the identified change coincides with a known boundary change. As mentioned in Section 5.1, the identified changepoints are ranked in descending order rather than being accepted or rejected on the basis of a threshold of evidence, because the large size of the dataset and the need for potential changes to be assessed further by hand means that it is most useful for analysts to consider — and correct — the most egregious errors first. In Appendix Figure C.3.1 we show the corresponding time series for the test statistic for a change at τ for different changepoint locations, τ. These can be used to see how much uncertainty there is for the specific location of a change.

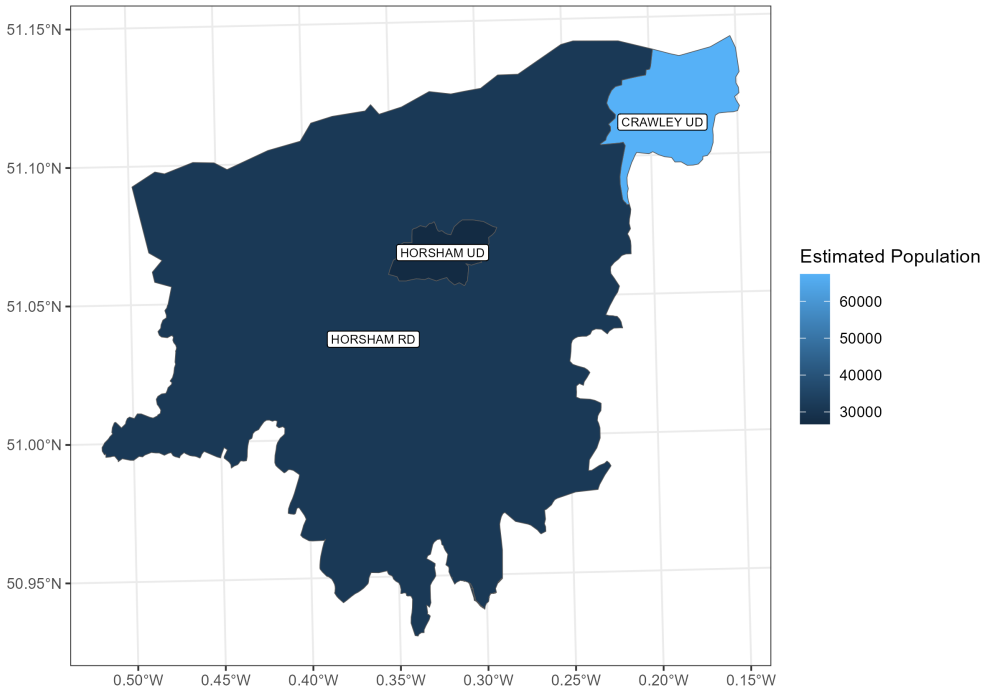


Figure 5.4.2: Horsham Rural District, Urban District and Crawley Urban District. Fill represents total estimated population in 1971 (Southall and Mooney, 2022). Despite occupying a much smaller area than Horsham RD, Crawley UD has a considerably higher population.

We give a summary of the errors detected by our changepoint detection method below.

- **Crawley Urban District (UD)**: appears to be interpolation error. Crawley UD an urban district created out of a rural district, **Horsham Rural District (RD)** (see Figure 5.4.2 to see a map of the districts and their estimated total populations in 1971). Interpolation method assuming that the population of Horsham RD is evenly distributed leads to it drastically underestimating the population occupying the area within Crawley UD's boundary before its creation. Crawley UD and Horsham RD merged. We name the adjusted district Crawley Aggregated District (AD).

- **Brighton County Borough (CB)**: probable historical change.

- **Longbenton RD**: boundary change in 1912 missing from our list of boundary changes. Corrected by adding the change to our interpolation process and re-interpolating. Longbenton RD was created in 1912 out of part of Tynemouth RD following its abolition (see Youngs (1991, p. 336, pp. 347-8, p. 724, pp. 724-5) for details and of other districts affected). Without accounting for the change in 1912 the areal interpolation does not have access to counts from source districts covered by former Tynemouth RD and pertaining to Longbenton UD from 1912 until the known change, and relies for interpolation upon counts from the districts that make up the expanded Longbenton UD from 1935. The very sharp drop in count from 1911 to 1912 is because a count is available for Tynemouth RD in 1911, but not for 1912 on.

- **Worthing RD**: possible interpolation error — district created in 1935. Merged with Chanctonbury RD, which was also created in 1933 and from two of Worthing RD's three source districts. We call the adjusted district Worthing AD.

- **Wallsend Metropolitan Borough (MB), Portslade by Sea UD, Newcastle upon Tyne CB, Hove MB**: probable historical change.

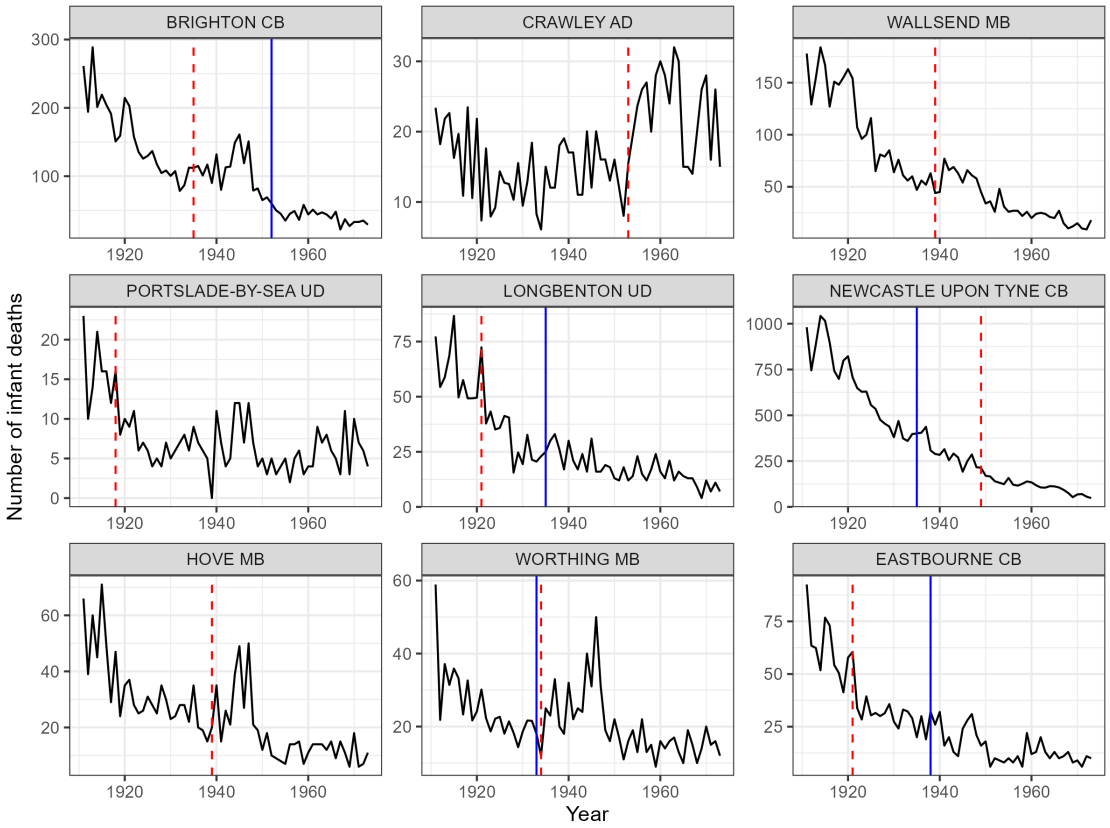


Figure 5.4.3: Results of second round of changepoint detection — after errors have been found (Figure 5.4.1) and corrected. Plot depicts series of infant death counts with the most evidence for a changepoint, in descending order. As before, blue vertical line depicts the date of a known boundary change. Red dashed line depicts the date of an identified boundary change

Having adjusted for those errors identified as interpolation errors by amalgamating affected districts, as described in the list above, we search the series again for a changepoint. In Figure 5.4.3 we see that the only change identified near a known boundary change is Worthing MB. This has a smooth peak to the likelihood function (Appendix Figure C.3.2 and a low value where the function is maximised, so we decide these changes to not need adjusting for. We note also Crawley AD has strong evidence for a change in 1953 though the likelihood test statistic has a smoother peak than in Figure

C.3.1. This changepoint is likely to be a historical change due to strong population growth in the 1950s (GB Historical GIS and University of Portsmouth, 2025).

5.4.2 Clustering the Infant Mortality Curves based on fPCA scores

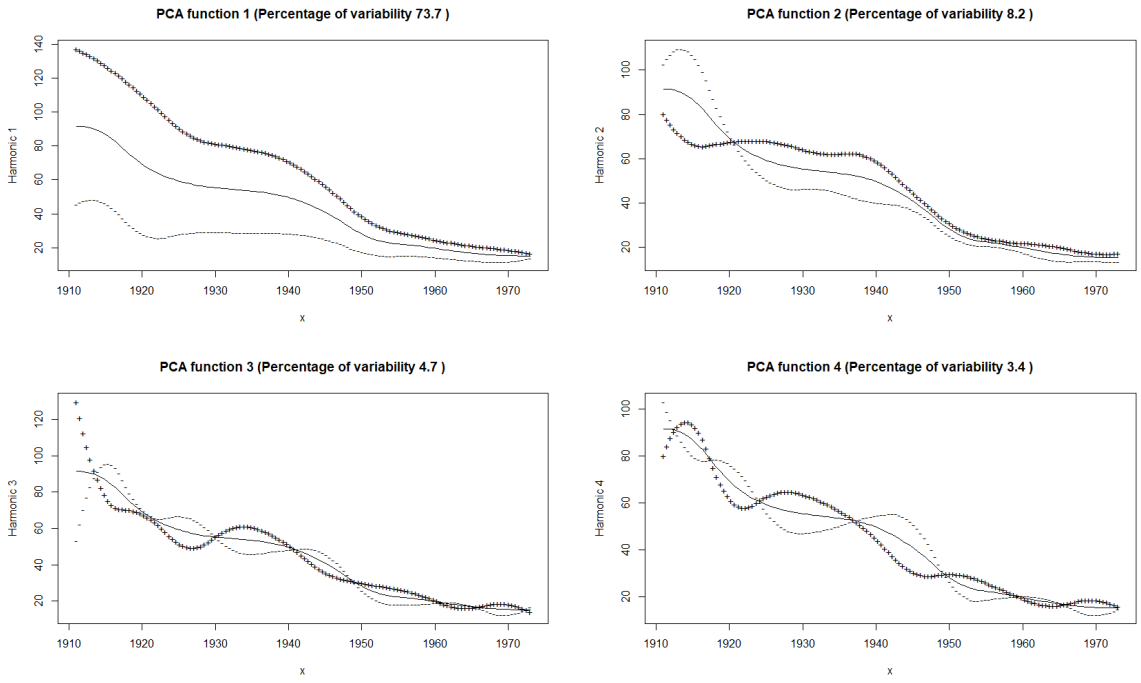


Figure 5.4.4: Plotting the first four functional principal components of the interpolated data (after correcting for error described in Section 5.4.1 but before aggregation). Produced using the `fda` package. Ramsay et al. (2024). The mean of the data is plotted, plus lines showing $+/-$ 2 standard deviations about each principal component function.

In this section we turn to considering the series of infant mortality rates created from interpolated series of infant death counts and counts of live births per local government district, based on the 1971 boundaries. We want to identify districts with commonalities in the behaviour of their infant mortality curves. We do this by treating them as functional data, then clustering them based on their functional PCA scores — the method described in Section 5.3.2. We show that identifying and correcting interpolation errors in series alters the grouping, and is therefore an important step in

aiding historical analysis.

We first cluster the series before correcting for data and interpolation errors, and we compare the clustering to that after adjusting for errors. Following the identification of errors in the interpolation, the two areas affected by the most severe errors are adjusted in a conservative way by amalgamating them with areas that are related to the same boundary change. In this section we first discuss the shape of the principal component functions, which give insight into the variability of the decline in infant mortality rates in Sussex and Northumberland between 1911 and 1973, before looking at the way the clustering is affected by correcting for interpolation errors.

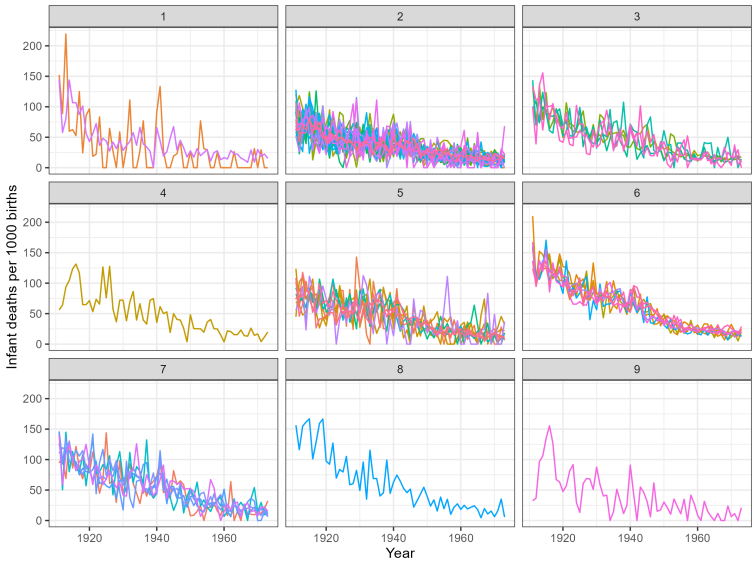
The four PCA functions constructed from the data that has been corrected for data and interpolation errors are depicted in Figure 5.4.4, plotted against the mean curve. The mean curve shows reasonably steady decline that begins to slow in the middle of the 1920s. Decline is steady but slow until the mid-1940s, where there is another acceleration in the decline. By the 1950s there is much less variation and the series decline slowly but steadily. The first PCA function, which accounts for just under three-quarters of the variability in the data, represents the varied starting points for the different districts over the period. The higher bound begins at over 130 deaths per 1000 births in 1911, while the lower bound represents just over 40. The higher bound plotted, which represents two standard deviations above the mean, declines fairly steeply between 1911 and the mid-1920s, it increases slightly then flattens out towards the end of the 1920s, before dropping again towards the end of the 1930s. The pace of decline slows from around 1950 to the end of the years in our dataset. The lower bound shows a slight increase in the first few years of the period, before declining from middle of the 1910s to a low in the early 1920s. It then rises slightly over the 1920s, and doesn't show a decline until the middle-end of the 1940s.

The second PCA function accounts for 8.2% of the variability. The upper curve shows a sharp decline at the beginning of the 1910s, this then increases again to peak

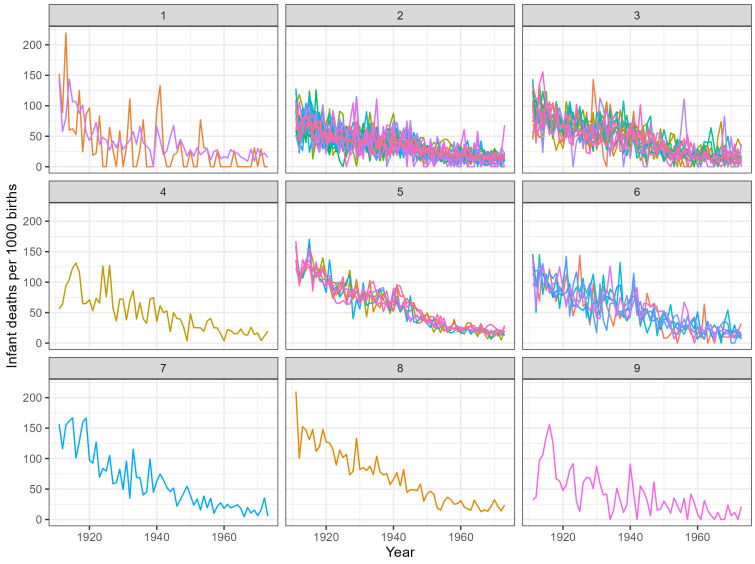
in the middle 1920s — in fact moving above above the lower curve. It then declines slowly, with some undulation, until the late 1930s, where it then declines more sharply. The lower curve increases sharply in the first few years of the 1910s, staying elevated over the years of World War I, before declining steeply by the end of the decade. The rate of decline slows sharply in the 1920s, and the curve declines slowly and undulatingly until the mid 1940s, where the pace of decline increases. This variability around the years of the First World I — which we also see in the third and fourth PCA functions — shows some of the useful insight that can be gained by looking at the data on a fine scale, rather than aggregating to county level. Winter (1982), for example, examines county level data and finds that for the vast majority of counties in Britain, infant mortality declines during the First World War.

The time periods with the most variability around PCA functions 3 and 4 are the first half of the 1910s, the mid 1920s and mid 1930s (PCA function 3); and the mid 1920s to mid 1930s, plus the late 1930s to 1950 (PCA function 4).

We now turn to considering the clustering of the data. We cluster the series before identifying and correcting for errors. We then re-cluster and examine the difference in the clustering. Figure 5.4.2 shows the first two functional principal component scores on a scatter plot, before and after correcting data errors and aggregating for interpolation errors. Before adjusting for errors, clusters 6 and 8 represents the highest scores based on functional principal component 1, with cluster 8, which is made up of Morpeth MB alone, separated from cluster 6 by dint of having a much more negative score on functional principal component 2. Cluster 7 has the next highest score, followed by cluster 4, which consists solely of Berwick upon Tweed UD. Berwick upon Tweed is a clear outlier based on principal components 3 and 4 (Figure C.4 in Appendix Section C.4). In the first clustering, clusters 5, 3, 1 and 9 have roughly similar scores on functional principal component 1, but vary over functional principal component 2. Cluster 9, consisting only of Seaford UD, is a separate cluster from 1 and 3, which it



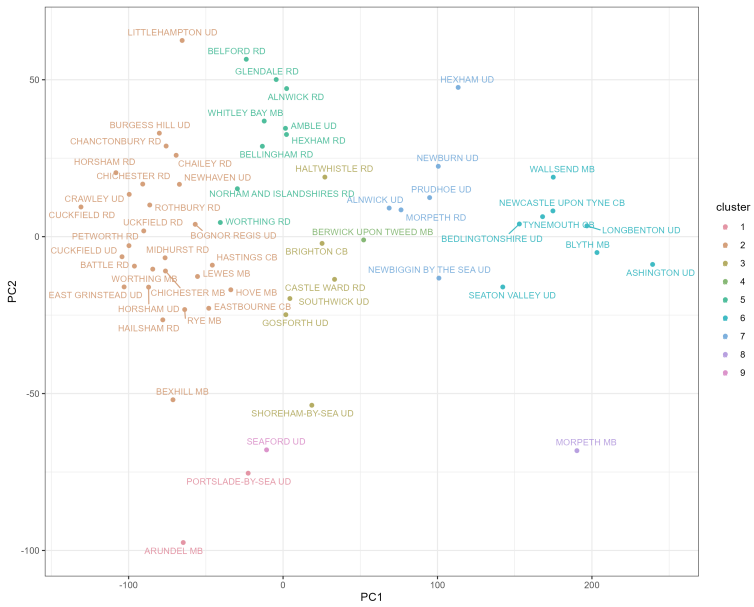
(a) Before adjusting for errors



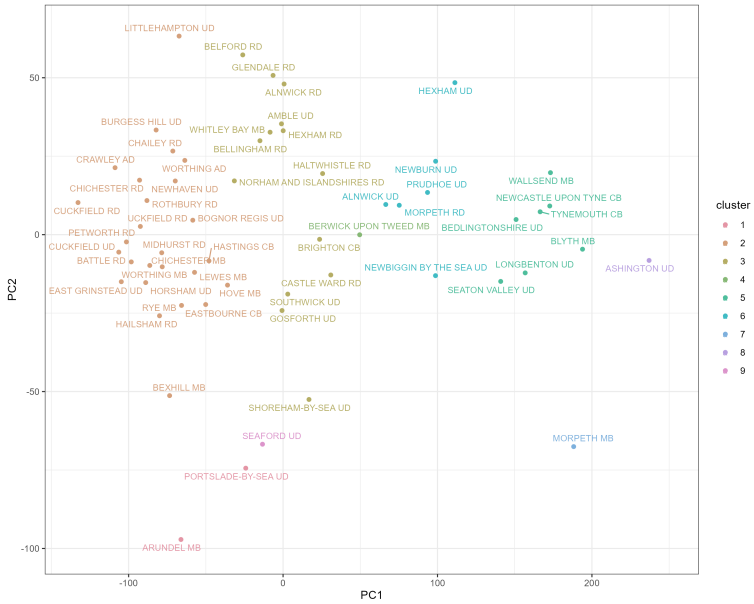
(b) After adjusting for errors

Figure 5.4.5: Local government district infant mortality rates plotted against time. Districts are grouped by cluster.

sits between, because Seaford UD, like Berwick upon Tweed, is an outlier based on principal components 3 and 4. Cluster 2 has the lowest scores on functional principal component 1, and is spread relatively evenly about functional principal component 2. We can see some evidence of a North-South divide in the clustering (Figures 5.4.7 and 5.4.8) — although it is also clear that the picture is more complex than that. Cluster



(a) Before adjusting for errors



(b) After adjusting for errors

Figure 5.4.6: Scatter plot of first two functional principal component scores — local government districts plotted by cluster.

2 is almost entirely local government districts in Sussex, with Rothbury RD the only one from Northumberland in this cluster By contrast, cluster 6 — which scores most highly on principal component 1, is exclusively made up of areas in Northumberland, as are the next highest scoring clusters: 4 and 7 and 8.

As discussed in Section 5.4.1, after detecting data and interpolation errors — and correcting the data errors — we aggregate those areas that we believe are subject to interpolation errors. This is a conservative approach to ensure we do not introduce additional errors due to, for example, modelling assumptions were we to use a more sophisticated interpolation method. We now turn to examining how detecting and adjusting for errors — both in the data itself and in the interpolation — affects the clustering.

Most of the areas that we find are subject to interpolation errors are in cluster 5. As discussed, this is one of the largest clusters and also the consistently healthiest over the period. Crawley AD is in cluster 2, the same as its parent clusters Horsham RD and Crawley UD. This is likely because, despite the aggregation error, the infant mortality rates are low for both districts and so they both sit in the healthiest cluster. Worthing RD is in cluster 5, but sits close to cluster 2.

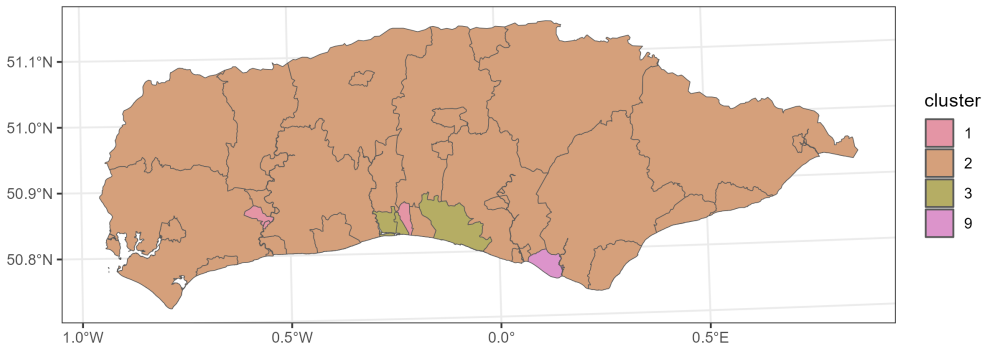


Figure 5.4.7: East and West Sussex: final clusters after changepoint detection and error correction

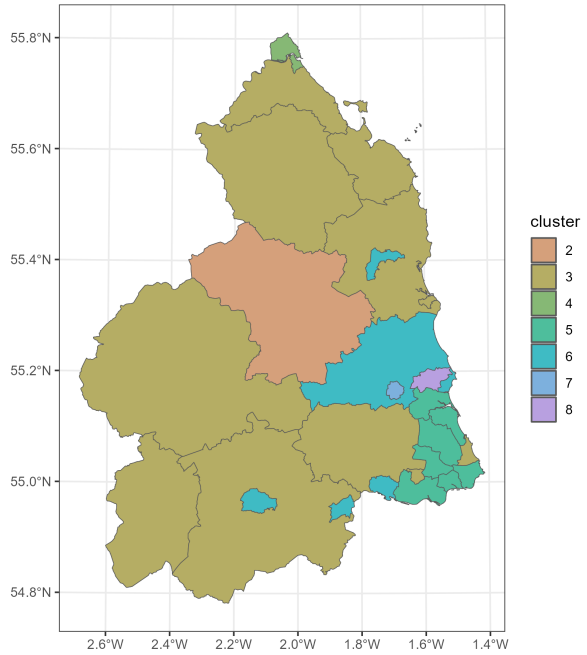


Figure 5.4.8: Northumberland: final clusters after changepoint detection and error correction

The most obvious change in clustering after the adjustments for an unknown boundary change and aggregation of areas where there is a suspicion of an interpolation error, is that clusters 3 and 5, which have similar scores on principal component 1 but are respectively above average and below on principal component 2, are merged into one: cluster 3. The other big difference is that Ashington UD, which has the highest score of any district on principal component 1, has been separated out into its own cluster, while Morpeth MB, remains in its own cluster, separate from the other districts that have the highest scores on principal component 1 (now renamed to cluster 5).

The clusters look more clearly defined in the second clustering — for instance Haltwhistle RD looks awkward in its initial clustering of cluster 3, being above average on PC2 unlike all the other members in its cluster. After adjusting for errors and re-clustering, clusters 2, 3, 5 and 6 the main clusters and the others containing just one or two districts that are outliers.

Ashington UD moves into its own cluster — cluster 8 — in the second round of

clustering. This appears appropriate as it has an exceptionally high infant mortality rate at the beginning of the period — over 20% of births — and this declines in a more or less linear fashion until the late 1940s, as opposed to the accelerated decline over the 1910s and early 1920s that we see in the mean curve. The Medical Officer's report in 1911 (Medical Officer of Health, Ashington U.D.C., 1911) states that the largest cause of infant deaths was diarrhoea and enteritis, at 64, and is attributed to an epidemic over the hot weather in summer months caused by flies to contaminate food, exacerbated by "privy ash-pits" being close to houses. The medical officer notes that improvement work to privies and ash-pits had taken place in the previous years, and recommends domestic hygiene education. Morpeth MB, which is in its own cluster in both rounds of clustering, also has very high rates at the beginning of the period, but sees some spikes in the first decade of our time period that take the IM rate to a higher level than 1911, and it also experiences some elevated levels in the 1930s that interrupt the general pattern of decline. The Medical Officer of Health Report in 1925 (Medical Officer of Health, Morpeth U.D.C., 1925) attributes high death rates and infant mortality rates in the district to overcrowding and the existence of slums, but notes that education through an Infant Welfare Centre has helped to reduce the outbreaks of diarrhoea that were common. Both Morpeth and Ashington are reasonably populous and it is unlikely that spikes in the infant mortality rate are down to volatility that we could expect in areas where the number of births is very low.

Portslade by Sea UD moves from cluster 7 to 6, leaving Arundel MB alone in cluster 7. Arundel MB scores very negatively in terms of fPC 2 and very positively on functional principal component 3. It suffers from spikes in infant mortality in the mid-1910s, the early 1930s and early 1940s. This corresponds to the smoothed series being high at the beginning of the period, corresponding with a negative score on fPCA 2 as well as a positive score on fPCA 3, and the increase in the smoothed curve over end of the 1930s, dropping off in the early 1940s. Portslade by Sea does not exhibit similar spikes. It is

likely that Arundel's spikes are caused by there being a low population. The unusually high figures in the early 1940s are caused by three and four instances of infant deaths (Medical Officer of Health, Arundel Borough, 1940, 1941).

5.5 Discussion

We have demonstrated a method for detecting interpolation errors in a time series of infant death counts, accommodating the fact that the data is observations of counts that exhibit trend over time. We have successfully identified series with interpolation errors in two trial areas of the dataset. We have shown that adjusting for the interpolation errors changes the grouping of the data.

Extensions to the method include: modelling the data as a multivariate time series and accounting for autocorrelation and/or spatial correlation in the series; extending the model to over-dispersed count data; searching for multiple changes in each series; allowing the trend term to slowly evolve over time.

5.6 Acknowledgements

The infant mortality data analysed in this chapter is from the Great Britain Historical Database (GBHDB), held by the UK Data Service, DOI:http://doi.org/10.5255/UKDA-SN-9035-1, ©Southall, H.R., University of Portsmouth, Mooney, G., University of Portsmouth. The GIS files containing digital boundaries of local government districts 1911-1971 are GBHDB DOI:http://doi.org/10.5255/UKDA-SN-9321-1 ©Southall, H.R., University of Portsmouth, Gregory, I., Lancaster University, Aucott, P., University of Portsmouth, Burton, N., University of Portsmouth. These data sets are openly licensed via https://creativecommons.org/licenses/by-sa/4.0/.

Chapter 6

Conclusions

The work in this thesis consists of three methods tackling different aspects of the problem of finding changepoints in the structure of time series. The applications considered in the thesis are diverse: air quality readings, financial time series, data from an offshore wind farm and historical public health data. In this section we describe the contribution of each chapter and adumbrate areas of further research based on this work.

6.1 Key findings

Chapter 3 introduces a method for detecting a single changepoint in VAR and VAR-X processes by fitting the appropriate model on a training set, then using the estimates of the model parameters to fit it to data from a test set, and running a multivariate changepoint detection method on the residuals. Its advantages over competitor methods are that it is much less computationally complex than others without making a commensurate compromise on accuracy. A key issue with detecting changepoints in VAR processes is computational complexity — the estimation of VAR parameters involves inverting a large matrix. It is also flexible — being able to be applied to both VAR and VAR-X models. It is relatively simple to interpret and does not require fine-tuning of a lot of hyper-parameters. A further advantage is that it can find changes in

some or all of the series of residuals — corresponding to a change in a lot, or just a few, of the entries in the variance-covariance matrix.

Chapter 4 introduces a novel method for detecting changepoints in regression models with missing data in the exogenous variables. This is the first method addressing this problem for regression models that explicitly accounts for the missing data directly within the changepoint detection method. That is, it does not require iteration within a maximum likelihood context to account for the missing data; nor imputation; nor deleting cases with any missingness prior to treating the data for changepoint detection. It is also able to detect more than one type of change in the model: both in the relationship between the endogenous and the exogenous variables, but also the model between the exogenous variables themselves. We demonstrate its efficacy via simulation scenarios and real data applications, and compare it to CCA and imputing the data. Usually imputation outperforms CCA but we note one scenario — Scenario 5, described in Section 4.3 — where the imputation mis-models the missing observations.

The work of Chapter 5 introduces a novel method for detecting a single change in count data that is subject to trend. There is also novelty in the application: applying changepoint detection to detect interpolation errors on a historical data set. This is one of only very few instances of changepoint detection being employed with an application in the humanities.

6.2 Further research

The method introduced in Chapter 3, SUBSET VAR, could be developed further by adapting it to detecting online changepoints, or to the detection of multiple changepoints. The issue of computational complexity is pressing both in online changepoint detection and the finding of multiple changepoints in the offline setting, so looking to adapt SUBSET-VAR to either of these settings would most likely require work to fur-

ther reduce the computational complexity of elements of the method. While SUBSET VAR is already a relatively fast process, measures to reduce the complexity of the VAR parameter estimation, such as LASSO (as deployed offline in Wang et al. (2019), could be explored.

The method in Chapter 4 is suitable for MCAR or MAR missingness mechanisms. However, a useful area of further research would be extending it to MNAR patterns, which would require modelling of the missingness mechanism within the likelihood function that is incorporated into the changepoint detection statistic. Gourieroux and Monfort (1981) considers a MNAR case when handling missing data in regression. In Chapter 4 we use Wild Binary Segmentation (Fryzlewicz, 2014) to adapt the method to detecting multiple changepoints, but another fruitful area of research should be adapting this to the detection of multiple changepoints using dynamic programming-based methods such as employed in (Killick et al., 2012). Finally, this method could be extended to more complex models under certain missingness mechanisms and patterns — such as those subject to non-normal errors.

There are two broad areas of further research for the work in this chapter. Firstly, the application: the method in this chapter was developed to detect interpolation errors in a large dataset of infant mortality data in 20th Century England and Wales. The aim is to render this dataset open to analysis at a fine scale over the whole period, 1911-1973 under consideration and the entire area for which data is available. It has been tested on a subset of the data — the counties of East Sussex, West Sussex and Northumberland. A natural step is to expand this analysis to cover all of the local government districts in England and Wales.

There is also scope for methodological innovation. The method models the data as Poisson with trend. Simulations show that the method has a high false positive rate when tested on data which has been simulated from a distribution with fatter tails, such as the Negative Binomial. A clear avenue for further research is extending the method to

deal with data that is distributed according to other count data distributions. Another potentially fruitful area of research would be modelling the series of infant mortality data as a multivariate series that is subject to cross-correlation that can change over time. This is a more realistic model of data — areas that are similar geographically, economically or socially are likely to have common latent factors driving the evolution of infant mortality rates. A further step would be developing a method with the ability to detect a change in the correlation between series, which could indicate an error or historical change. A novel aspect of this work, that could be developed further, is the method's ability to detect one type of changepoint (an interpolation error) while ignoring changes that are due to historical trend. The method could be extended to look for a change in either, reporting which type of change it there is more evidence for.

Appendix A

Chapter 3: SUBSET VAR

A.1 Simulation Studies

A.1.1 Simulating stationary VAR processes

In all of our simulation studies we require stationary VAR processes on either side of a changepoint. This means a process where the mean and variance do not change over time. A non-stationary process can quickly tend to infinity, creating spurious changepoints.

Defining a VAR process as in Section 3.2.2

$$\boldsymbol{y}_t = \sum_{i=1}^{p} \boldsymbol{\phi}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t,$$

Lütkepohl (2013, pp. 15-16) define a stable VAR process as one where the roots of its reverse characteristic polynomial are outside the unit circle.

$$\det(\boldsymbol{I}_d - \boldsymbol{\phi}_1 z, \ldots, -\boldsymbol{\phi}_p z^p) \neq 0 \quad \text{for} |z| \leq 1.$$

For a VAR(1) process it suffices to ensure that the moduli of the eigenvalues of the coefficient matrix, $\boldsymbol{\phi}_1$ are less than 1. To allow our simulations code to simulate

stationary VAR processes where $p > 1$ we use a process introduced by (Ansley and Kohn, 1986) and coded as `random_coefmats2` in the `gmvarkit` package from Virolainen (2021).

Simulation 1

In Simulation 1 we simulate a VAR (1) process with zero mean by calling `random_coefmats2` to generate a $\boldsymbol{\phi}$ matrix at random. To effect a change we multiply the eigenvalues by a predefined constant, $-1 < \alpha < 1$. This allows us to control the size of the change. We also restrict the proportion of variates affected by a change to 20% by multiplying only the first 20% of the eigenvalues by α. We initialise the post-change VAR process using the final observations of the pre-change process, to ensure we do not have an abrupt change in mean at the changepoint.

Simulations 2 and 3

In Simulations 2 and 3 we simulate a zero mean VAR (1) process as before, and we introduce a nuisance process at the changepoint by simulating another zero mean VAR (1) process. Observed values after the change are the sum of the original process at time t and the nuisance process at time t. We multiple the nuisance process by a value, $-1 < \alpha < 1$, to control the size of the change. In Simulation 2 we use a sparsity parameter to make the change a sparse one, with 40% of variates affected, and in Simulation 3 we create a dense change, with all variates affected.

A.1.2 Determining Hyperparameters for Simulation Study

SUBSET has three hyperparameters — α, β and K — used across two penalty regimes.

Our SUBSET algorithm first returns likelihood statistics for a sparse change $D'_{i,\tau} = \max(D_{i,\tau} - \alpha, 0)$ and for a dense change $D_{i,\tau}$.

As described in section 3.2.3 the test statistic for SUBSET is

$$S_\tau = \max\left\{\sum_{i=1}^{d} D'_{i,\tau} - \beta, \sum_{i=1}^{d} D_{i,\tau} - K\right\}.$$

In our code, $K = \beta + \text{Threshold}$.

We set α as the default value, $2\log(d)$. We then ran 1000 repetitions of a VAR(1) process in five, 10 and 50 dimensions with no change.

We fit a VAR(1) model to the process and calculate the residuals, as described in Section 3.2.1. We obtain the $D'_{i,\tau}$ and $D_{i,\tau}$ for every candidate changepoint, τ, and for every repetition of the simulation we return the maximum value of each of these statistics. In each dimensional setting we then calculate the 99.5th percentile of these scores — equivalent to a 0.5% false positive rate for a dense change and for a sparse change. We set these values as the penalties for a sparse and dense change.

In five dimensions these are: $\beta = 16$; threshold $+ \beta = 26.5$.

In 10 dimensions they are: $\beta = 18$; threshold $+ \beta = 35.6$.

In 50 dimensions they are: $\beta = 18.2$; threshold $+ \beta = 102$

For the method of Wang et al. (2019) (LASSO-VAR) there are two hyperparameters to consider. These are γ, a penalty applied to the likelihood statistic to penalise selecting too many changepoints and λ, a tuning parameter for the LASSO. We set λ to the default value of $0.1\sqrt{\log(d)}$ specified in Wang et al. (2019) for the simulations detailed in that paper. For γ we follow a similar process to above for SUBSET —- running the method to return the level of γ at which a change would be accepted on 1000 simulations of a VAR(1) process with no change present. As we are only dealing with a single test statistic, we look to set γ to the level of the 99th percentile of values, which is equivalent to a 1% false positive rate. In five dimensions this is 56.9, in 10 dimensions this is 154, and in 50 dimensions, 2579.

The SBS algorithm has a parameter, q, coded in the `hdbinseg R` package Cho and Fryzlewicz (2018) that runs, by default, 1000 iterations of a bootstrap resampling

for each run of the algorithm to determine an optimal threshold for the false positive rate specified by q. The default value is 0.01, giving a 1% false positive rate. In five, 10 and 50 dimensions we run the SBS algorithm on 1000 simulations of a VAR (1) process with no change. We vary q to find the value that gives a 1% false positive rate in each dimensional setting. For each dimensional setting we try values $q = 0.001, 0.005, 0.01, 0.025, 0.05$. We select the value of q that returns the closest to 10 false positives in the 1000 repetitions in each run of simulations without change.

In five dimensions, $q = 0.025$ gives 13 false positives, closer than the next nearest, $q = 0.01$, which gives 5. In 10 dimensions, $q = 0.01$ gives 13 false positives. This is closer than $q = 0.005$ which also gives 5. In 50 dimensions $q = 0.001$ gives 30 false positives.

A.2 Application to Glasgow pollution data

We exclude data on the following days due to pollution events being reported in Department for Environment Food and Rural Affairs (2017, 2018, 2019, 2020) that bring pollution levels in Scotland into the moderate category or higher. These are:

- PM10: February $25-27$, 2019; April $14-25$ 2019; February 14, 2017, November $24-26$, 2016.
- Ozone: August $24-26$ and 28, 2019; July $24-26$, 2019; June $25-30$ 2018; July 1, 2018; July $4-5$, 2018; July $27-29$, 2018

Appendix B

Chapter 4: Changepoints in regressions with missing data

B.1 Relating the estimates of the factored likelihood to the parameters of the original model

B.1.1 Univariate model

Gourieroux and Monfort (1981) give that these parameters can be related to those in the original regression equation as follows:

$$\beta_1 = \frac{a^2 d}{a^2 d^2 + c^2},$$

$$\beta_2 = b - \frac{a^2 d}{a^2 d^2 + c^2}(e + db),$$

$$\sigma^2 = \frac{a^2 c^2}{a^2 d^2 + c^2},$$

$$\gamma = bd + e$$

and

$$\eta^2 = a^2d^2 + c^2.$$

B.1.2 Multivariate model

This proof closely follows the proof for $p = 1$, found in Gourieroux and Monfort (1981). We consider the model of (4.2.5) and (4.2.6). In order to relate the parameters a, b, c, d, e, f to the parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\boldsymbol{\sigma}$, we need to express the distribution of $X_i|X_{i+1}, \ldots, X_p, Y, Z$ and $Y|Z$ in terms of the original model parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, σ^2, γ and η. We begin with $X_i|X_{i+1}, \ldots, X_p, Y$. First, observe that $Y, X_1, \ldots, X_p$ is jointly multivariate normal. Using this information, the conditional distribution is:

$$N\left(E(X_i) + A_iB_i^{-1}\left(\begin{pmatrix} X_{i+1} \\ \vdots \\ X_N \\ Y \end{pmatrix} - C_i\right), Var(X_i) - A_iB_i^{-1}A_i^T\right). \tag{B.1.1}$$

Here,

$$A_i = \Big(Cov(X_{i+1}, X_i), \ldots, Cov(X_N, X_i), Cov(X_i, Y)\Big),$$

$$B_i = Var(X_{i+1}, \ldots, X_N, Y),$$

and

$$C_i = \begin{pmatrix} E(X_{i+1}) \\ \vdots \\ E(X_N) \\ E(Y) \end{pmatrix}.$$

We therefore need expressions for $E(Y)$, $E(X_i)$, $Var(Y)$, $Var(X_i)$, $Cov(X_i, Y)$, $Cov(X_i, X_j), i \neq j$. We let $\omega_{i,j} = Cov(X_i, X_j), i \neq j$ where $|\omega_{i,j}| < 1$.

$$E(Y) = \sum_{j=1}^{p} \beta_{1,j} \sum_{k=1}^{K} \gamma_{j,k} Z_k + \sum_{k=1}^{K} \beta_{2,k} Z_k$$

$$E(X_j) = \sum_{k=1}^{K} \gamma_{j,k} Z_k$$

$$Cov(Y, X_i) = \sum_{j=1}^{p} \beta_{1,j} \omega_{i,j} \mathbf{1}_{i \neq j} + \beta_{1,i} \eta^2$$

$$Var(Y) = \sum_{j=1}^{p} \beta_{1,j} \eta^2 + 2 \sum_{i<j} \beta_{1,j} \omega_{i,j}.$$

We can write c_i^2 using (B.1.1):

$$c_i^2 = \eta^2 - A_i B_i^{-1} A_i^T.$$

Defining,

$$A_i = \begin{pmatrix} \omega_{i,i+1} & \dots & \omega_{i,p} & Cov(X,Y) \end{pmatrix},$$

and

$$B_i^{-1} = \begin{pmatrix} b^*_{1,1} & \dots & b^*_{1,p-i} & b^*_{1,p-i+1} \\ \vdots & & \vdots & \vdots \\ b^*_{p-i,1} & \dots & b^*_{p-i,p-i} & b^*_{p-i,p-i+1} \\ b^*_{p-i+1,1} & \dots & b^*_{p-i+1,p-i} & b^*_{p-i+1,p-i+1} \end{pmatrix},$$

then:

$$d_i = \sum_{L=1}^{L=p-i} \omega_{i,i+L} b^*_{L,p-i+1} + Cov(X_i,Y) b^*_{p-i+1,p-i+1},$$

$$e_{i,k} = \gamma_{i,k} - \sum_{P=i+1}^{P=p} \left(\sum_{L=1}^{L=p-i} \omega_{i,i+L} b^*_{L,P-i} + Cov(X_i,Y) b^*_{p-i+1,1} \right) \gamma_{P,k} -$$
$$\left(\sum_{L=1}^{L=p-i} \omega_{i,i+L} b^*_{L,p-i+1} + Cov(X_i,Y) b^*_{p-i+1,p-i+1} \right) \left(\gamma_{i,k} \sum_{i=1}^{p} \beta_{1,i} + \beta_{2,k} \right),$$

and

$$f_{i,i+L} = \sum_{L=1}^{L=p-i} \omega_{i,i+L} b^*_{L,P-i} + Cov(X_i,Y) b^*_{p-i+1,1}.$$

We now turn our attention to $Y|Z$. Y is equivalent to the following model:

$$\sum_{j=1}^{p} \beta_1, j (\sum_{k=1}^{K} \gamma_{j,k} Z_k + \zeta) + \sum_{k=1}^{K} \beta_{2,k} Z_k + \epsilon$$

$$E(Y) = \sum_{j=1}^{p} \beta_{1,j} \sum_{k=1}^{K} \gamma_{j,k} Z_k + \sum_{k=1}^{K} \beta_{2,k} Z_k$$

And, by inspection,

$$b_k = \sum_{j=1}^{p} \beta_{1,j} \gamma_{j,k} + \beta_{2,k}.$$

Then,

$$a^2 = Var(Y) = \sum_{j=1}^{p} \beta_{1,j} \eta^2 + 2 \sum_{i<j} \beta_{1,j} \omega_{i,j}$$

as derived earlier.

We have shown how to relate the original parameters of the model to $a, \ldots, f$. As a result, the maximum likelihood estimates of $a, \ldots, f$ will also be the maximum likelihood estimates for $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ (Berger and Casella, 2001).

B.2 Simulation Results

We give more detailed results for each of the simulation scenarios. The detection threshold is determined by simulation on data without a change and set to control the false positive rate at 5%. In all results we report the probability of detection as when a method detects a change anywhere in the sequence, and the true positive rate when it detects a change within $+/-10t$ of the true change. We also give the mean distance between the estimated changepoint and the true location. We report the mean distance only on repetitions where all three methods returned a changepoint.

		Mean			True positive (rate 1)			True positive (rate 2)		
Missing	Method	S	M	L	S	M	L	S	M	L
0.0	CC	**4.21**	**1.77**	**0.92**	**99.70**	**100.00**	**100.00**	**89.1**	**97.2**	**99.2**
0.0	FL	4.57	1.80	**0.92**	**99.80**	**100.00**	**100.00**	87.9	97.0	99.1
0.0	Imputed	**4.21**	**1.77**	**0.92**	99.70	**100.00**	**100.00**	**89.1**	**97.2**	**99.2**
0.2	CC	6.55	2.82	1.68	99.60	100.00	100.00	84.6	95.7	97.8
0.2	FL	**4.95**	**2.37**	**1.36**	**99.80**	**100.00**	**100.00**	**88.6**	**96.9**	**98.2**
0.2	Imputed	5.05	2.42	1.38	**99.80**	**100.00**	**100.00**	87.6	96.5	**98.2**
0.4	CC	9.98	4.59	2.63	99.60	100.00	100.00	75.3	89.7	95.9
0.4	FL	**5.75**	**2.22**	**1.34**	**99.90**	99.90	**100.00**	**85.2**	**96.4**	**98.7**
0.4	Imputed	**5.75**	2.52	1.49	**99.90**	**100.00**	**100.00**	82.8	95.9	98.6
0.6	CC	18.86	7.69	4.79	96.50	99.80	99.90	61.0	80.1	90.4
0.6	FL	**5.14**	**2.13**	**1.18**	**100.00**	**100.00**	**100.00**	**87.6**	**96.9**	98.8
0.6	Imputed	6.04	2.61	1.38	99.80	**100.00**	**100.00**	84.7	95.1	**99.0**
0.8	CC	39.05	19.39	10.87	81.50	97.40	99.60	35.5	56.2	69.7
0.8	FL	**4.63**	**2.28**	**1.26**	**99.90**	**100.00**	**100.00**	**83.7**	**95.0**	**98.9**
0.8	Imputed	5.96	2.96	1.66	**99.90**	**100.00**	**100.00**	81.4	92.2	97.7

Table B.2.1: Table showing the accuracy and true positive rate of methods in detecting a change where one variable, X, is subject to missingness with a Missing at Random mechanism (Scenario 1). There are five other exogenous variables. We show the average absolute distance from the true change (mean), the percentage of repetitions in which a change is correctly identified (for rate 1 this is not constrained by accuracy, for rate 2 this is within $+/-10$ of the true change). We display results by size of change — the constant by which we have multiplied the slope parameters of the regression ($S = 1.5, M = 1.75, L = 2$). The best results for each level of missingness and size of change are in bold.

		Mean			True positive (rate 1)			True positive (rate 2)		
Missing	Method	S	M	L	S	M	L	S	M	L
0.0	CC	**2.27**	**1.06**	**0.50**	**99.90**	**100.00**	**100.00**	**96.6**	**99.4**	**100.0**
0.0	FL	2.47	1.10	0.54	**99.90**	**100.00**	**100.00**	95.6	**99.4**	100.0
0.0	Imputed	**2.27**	**1.06**	**0.50**	**99.90**	**100.00**	**100.00**	**96.6**	**99.4**	**100.0**
0.2	CC	3.47	1.51	1.00	**100.00**	**100.00**	**100.00**	92.1	98.5	99.6
0.2	FL	**2.95**	**1.30**	**0.68**	99.90	**100.00**	**100.00**	**94.0**	98.6	**99.9**
0.2	Imputed	3.07	1.33	0.75	**100.00**	**100.00**	**100.00**	93.4	**98.7**	**99.9**
0.4	CC	5.41	2.46	1.63	**99.90**	**100.00**	**100.00**	86.1	96.5	98.5
0.4	FL	**2.83**	**1.18**	**0.69**	**99.90**	**100.00**	**100.00**	**94.8**	**99.1**	**99.6**
0.4	Imputed	3.12	1.32	0.80	**99.90**	**100.00**	**100.00**	93.7	99.0	99.5
0.6	CC	10.03	4.65	3.26	99.10	99.90	**100.00**	71.3	88.1	94.3
0.6	FL	**3.20**	**1.45**	**0.77**	**100.00**	**100.00**	**100.00**	**92.5**	**98.4**	**99.9**
0.6	Imputed	3.80	1.94	0.99	**100.00**	**100.00**	**100.00**	90.3	97.3	99.7
0.8	CC	23.60	11.54	7.65	94.30	99.50	99.90	50.0	69.0	77.4
0.8	FL	**3.07**	**1.46**	**0.79**	99.90	**100.00**	**100.00**	**92.9**	**98.4**	**99.6**
0.8	Imputed	4.20	1.83	1.12	**100.00**	**100.00**	**100.00**	89.0	97.5	99.5

Table B.2.2: Table showing the accuracy and true positive rate of methods in detecting a change where there are two variables subject to missingness in a monotone missingness pattern and with a Missing at Random mechanism (Scenario 2). Results are presented as in Table B.2.1.

		Mean			True positive (rate 1)			True positive (rate 2)		
Missing	Method	S	M	L	S	M	L	S	M	L
0.0	CC	**5.09**	2.34	**1.18**	99.80	**100.00**	**100.00**	**88.7**	96.6	**99.5**
0.0	FL	5.85	2.22	1.27	**100.00**	**100.00**	**100.00**	87.6	**96.8**	99.2
0.0	Imputed	**5.09**	2.34	**1.18**	99.80	**100.00**	**100.00**	**88.7**	96.6	**99.5**
0.2	CC	5.80	2.94	1.98	99.90	100.00	100.00	85.0	94.4	97.9
0.2	FL	**5.42**	**2.51**	**1.60**	**100.00**	**100.00**	**100.00**	**88.0**	**95.3**	**98.5**
0.2	Imputed	6.46	3.00	1.80	**100.00**	**100.00**	**100.00**	83.8	94.1	98.1
0.4	CC	10.46	4.35	2.84	99.20	**100.00**	99.90	77.3	90.0	95.1
0.4	FL	**7.46**	**3.10**	**1.92**	**100.00**	**100.00**	**100.00**	**83.0**	**94.6**	**97.1**
0.4	Imputed	7.95	3.74	2.42	99.80	**100.00**	**100.00**	79.2	92.9	95.6
0.6	CC	12.95	6.74	4.44	97.60	99.60	100.00	67.8	83.3	89.9
0.6	FL	**7.81**	**3.66**	**1.99**	**99.90**	**100.00**	**100.00**	**78.1**	**91.8**	**97.8**
0.6	Imputed	10.33	5.13	3.12	**99.90**	**100.00**	**100.00**	71.4	86.6	93.9
0.8	CC	27.80	16.03	10.02	93.70	98.70	99.60	44.5	60.6	74.0
0.8	FL	**8.68**	**4.26**	**2.55**	**100.00**	**99.90**	**100.00**	**74.4**	**89.9**	**95.5**
0.8	Imputed	13.79	6.79	3.96	99.70	99.90	100.00	62.5	79.5	91.0

Table B.2.3: Table showing the accuracy and true positive rate of methods in detecting change where X is bivariate and has correlated errors (Scenario 3). Results are presented as in Table B.2.1.

		Mean			True positive (rate 1)			True positive (rate 2)		
Missing	Method	S	M	L	S	M	L	S	M	L
0.0	CC	**21.72**	**10.46**	**4.72**	96.10	99.10	**100.00**	**54.0**	**73.9**	**87.8**
0.0	FL	24.49	11.77	5.08	93.40	98.80	**100.00**	51.7	72.4	86.4
0.0	Imputed	**21.72**	**10.46**	**4.72**	**96.10**	**99.10**	**100.00**	**54.0**	**73.9**	**87.8**
0.2	CC	25.25	13.23	7.01	91.70	98.50	99.90	46.8	68.1	82.7
0.2	FL	22.27	**9.15**	**5.64**	93.10	**99.10**	**100.00**	49.7	**76.7**	**85.9**
0.2	Imputed	**21.02**	9.37	5.77	**95.20**	**99.10**	**100.00**	**50.9**	74.6	84.5
0.4	CC	35.29	19.06	9.20	83.70	98.50	99.30	37.7	56.9	75.0
0.4	FL	24.25	**10.30**	**5.70**	91.60	99.30	99.90	**50.9**	**72.1**	**85.1**
0.4	Imputed	**24.22**	10.51	5.92	**94.30**	**99.50**	**100.00**	48.6	70.8	84.3
0.6	CC	51.99	27.16	15.13	72.40	94.70	98.80	22.4	47.9	66.1
0.6	FL	**22.96**	**10.17**	**5.42**	90.10	99.10	**100.00**	**46.1**	**70.8**	**85.9**
0.6	Imputed	25.71	10.82	6.33	**92.30**	**99.50**	**100.00**	43.7	69.5	84.0
0.8	CC	81.94	44.54	26.81	51.10	81.80	94.10	11.2	29.6	45.3
0.8	FL	**21.04**	**10.26**	**5.80**	**92.70**	99.20	**99.90**	**45.4**	**69.4**	**84.5**
0.8	Imputed	27.57	12.39	6.49	92.40	**99.60**	**99.90**	41.5	66.0	81.6

Table B.2.4: Table showing the accuracy and true positive rate of methods in detecting a change where Y has autoregressive errors (Scenario 4). Results are presented as in Table B.2.1.

		Mean			Probability of detection)			True positive rate		
Missing	Method	S	M	L	S	M	L	S	M	L
0.0	CC	390.39	409.55	412.66	5.20	4.50	5.60	0.1	0.0	0.0
0.0	FL	**95.09**	**61.66**	**8.11**	**77.90**	**97.20**	**99.40**	**31.2**	**58.1**	**76.6**
0.0	Imputed	390.39	409.55	412.66	5.20	4.50	5.60	0.1	0.0	0.0
0.2	CC	275.70	388.86	379.94	5.50	5.40	4.10	0.0	0.0	0.1
0.2	FL	**101.91**	**41.32**	**12.67**	**66.50**	**93.90**	**99.00**	**25.9**	**52.3**	**72.8**
0.2	Imputed	278.26	399.57	357.50	5.30	6.20	6.30	0.0	0.0	0.1
0.4	CC	294.00	312.54	369.75	4.20	6.00	5.70	0.0	0.1	0.0
0.4	FL	**120.67**	**45.38**	**31.62**	**56.60**	**89.20**	**97.70**	**17.4**	**43.5**	**61.4**
0.4	Imputed	211.00	264.77	314.25	6.50	7.00	10.50	0.0	0.4	1.1
0.6	CC	335.00	250.60	354.62	4.60	4.90	5.00	0.0	0.1	0.1
0.6	FL	**136.80**	**31.90**	**23.67**	**46.60**	**85.20**	**95.20**	**12.1**	**35.3**	**57.6**
0.6	Imputed	326.40	228.50	152.38	7.80	14.80	21.90	0.3	1.5	4.0
0.8	CC	380.25	262.82	322.48	3.50	3.50	5.20	0.0	0.2	0.3
0.8	FL	**85.00**	**63.91**	**31.79**	**34.10**	**71.90**	**89.20**	**7.5**	**24.4**	**44.4**
0.8	Imputed	207.75	202.91	114.41	11.00	23.00	40.90	0.9	3.0	10.2

Table B.2.5: Table showing accuracy and true positive rate of methods in detecting a change in the distribution of X given Z (Scenario 5). Results are presented as in Table B.2.1.

Appendix C

Chapter 5: Detecting interpolation errors

C.1 Interpolation between multiple boundary changes

Set 1971 boundaries as the boundaries we want to interpolate onto. There are j local government districts in 1971. These are the Target Areas. The A_j^T for the jth local government district in 1971. Then for each of the j local government districts in 1971.

1. Identify years of known boundary changes for the local government district. If none, no interpolation needed

2. For 1911 up to, but not including the year of the first boundary change, use the 1911 boundaries for the area of each of the source districts, and the overlap between each of the source districts and the target district. Calculate $\hat{x}_{t,j}^T$ for each of the years up to (but not including) the year of the first change.

3. If only one known boundary change for the district, stop. Else:

4. For the year of the first change, up to the year of the second change, use the next available set of boundaries for the source district. For example, if there is a

change in 1934 and the next is 1955, then use the 1951 boundaries for the years $1934 - 1954$.

5. Continue until interpolations completed for all years prior to the last known change. For the years from the last change until the end of the period (1973), the un-interpolated data are used.

C.2 Simulation Study: changepoint methods

We test our method via simulation. We give an overview of the simulation scenarios below, before giving the results of each. We first simulate data from a Poisson distribution with part of the rate parameter accounting for trend and the other intercept. We then test the performance of our changepoint method on simulated data that does not fit the model described in Section 5.3.1. For each simulation scenario we vary the size of change and where in the series the changepoint occurs. In Scenario 4 we also compare our method to one that searches for a change in trend or an abrupt change.

1. Abrupt change in a sequence of Poisson distributed variables. Simulate from a Poisson distribution with parameter λ_t where

$$\lambda_t = \begin{cases} e^{(\alpha+\beta t)} & \text{if } \ t \leq \tau \ , \\ e^{(\alpha^*+\beta t)} & \text{if } \ t > \tau \ . \end{cases}$$

$n = 200$ $\tau = (25, 150)$ $\alpha^* = \alpha c$, where $c = 0.5, 0.8, 1.2, 1.5$. $\alpha \sim U(2, 4)$, $\beta \sim U(-0.025, 0.025)$.

2. Change in the trend of a sequence of Poisson distributed variables. Simulate from

a Poisson distribution with parameter λ_t where

$$\lambda_t = \begin{cases} e^{(\alpha+\beta t)} & \text{if } t \leq \tau \ , \\ e^{(\alpha+\beta^* t)} & \text{if } t > \tau \ . \end{cases}$$

$n = 200$ $\tau = (25, 150)$ $\beta^* = \beta c$, where $c = 0.5, 0.8, 1.2, 1.5$

3. Change in a sequence of negative binomially distributed variables. Simulate from a negative binomial distribution NB(k, p):

$$p(x) = \frac{\Gamma(x+k)}{\Gamma(k)x!} p^k (1-p)^x$$

with mean $\mu = \frac{k(1-p)}{p}$ where k is the size parameter. This is a heavier-tailed model for count data than the Poisson. We simulate our negative binomial variables with a time-varying mean, μ_t, and with a change at τ as follows:

$$\mu_t = \begin{cases} e^{(\alpha+\beta_t))} & \text{if } t \leq \tau \ , \\ e^{(\alpha^*+\beta_t))} & \text{if } t > \tau \end{cases} .$$

As before, $n = 200$ $\tau = (25, 150)$ $\alpha^* = \alpha c$, where $c = 0.5, 0.8, 1.2, 1.5$. $\alpha \sim U(2, 4)$, $\beta \sim U(-0.015, 0.015)$. $k \sim U(1, 10)$ We reduce the range for β, compared with the Poisson-based simulations. This is to control the variance, which increases with t: $\mu_t + \frac{\mu_t^2}{k}$

4. Abrupt change in a sequence of Poisson distributed variables without trend. Simulate from a Poisson distribution with parameter λ where

$$\lambda = \begin{cases} e^{(\alpha)} & \text{if } t \leq \tau \ , \\ e^{(\alpha^*)} & \text{if } t > \tau \ . \end{cases}$$

$n = 200$ $\tau = (25, 150)$ $\alpha^* = \alpha c$, where $c = 0.5, 0.8, 1.2, 1.5$. $\alpha \sim U(2, 4)$.

C.2.1 Results

	Accuracy		True positive rate	
Change Factor	$\tau = 25$	$\tau = 150$	$\tau = 25$	$\tau = 150$
0.5	99.6	93.6	100.0	96.6
0.8	92.2	78.2	95.0	85.2
1.2	96.7	84.4	98.5	89.0
1.5	100.0	98.5	100.0	99.6

Table C.2.1: Results of simulation Scenario 1 over 1000 repetitions. Accuracy presents the percentage of repetitions where the method identifies a changepoint within $+/-5$ of the true change. True positive rate represents the number of repetitions where is changepoint is identified, regardless of accuracy. Test statistic threshold is calculated to provide a 5% false positive rate, based on 1000 repetitions with no change.

	Accuracy		True positive rate	
Change Factor	$\tau = 25$	$\tau = 150$	$\tau = 25$	$\tau = 150$
0.5	12.9	88.3	23.0	93.5
0.8	2.0	60.8	8.5	75.4
1.2	2.5	53.6	7.9	68.0
1.5	15.9	83.4	25.5	92.1

Table C.2.2: Results of simulation Scenario 2. Accuracy presents the percentage of repetitions where the method identifies a changepoint within $+/-5$ of the true change. True positive rate represents the number of repetitions where is changepoint is identified, regardless of accuracy. Test statistic threshold is calculated to provide a 5% false positive rate, based on 1000 repetitions with no change.

	Accuracy		True positive rate	
Change Factor	$\tau = 25$	$\tau = 150$	$\tau = 25$	$\tau = 150$
0.5	0.0	25.0	0.2	25.2
0.8	0.0	12.9	1.8	15.0
1.2	0.0	10.8	10.2	17.3
1.5	0.0	29.0	16.5	31.7

Table C.2.3: Results of simulation Scenario 3. Accuracy presents the percentage of repetitions where the method identifies a changepoint within $+/-5$ of the true change. True positive rate represents the number of repetitions where is changepoint is identified, regardless of accuracy. Test statistic threshold is calculated to provide a 5% false positive rate, based on 1000 repetitions with no change.

		Accuracy		True positive rate	
Method	Change Factor	$\tau = 25$	$\tau = 150$	$\tau = 25$	$\tau = 150$
1	0.5	100.0	99.9	100.0	100.0
2	0.5	99.3	99.9	100.0	100.0
1	0.8	93.8	95.9	98.8	99.9
2	0.8	88.3	91.0	98.3	99.4
1	1.2	98.2	97.1	100.0	99.6
2	1.2	94.4	96.4	99.6	99.9
1	1.5	100.0	100.0	100.0	100.0
2	1.5	100.0	100.0	100.0	100.0

Table C.2.4: Results of simulation Scenario 4 over 1000 repetitions. Method 1 searches for an abrupt change, while Method 2 searches for an abrupt change or a change in trend. As before, accuracy presents the percentage of repetitions where the method identifies a changepoint within $+/-5$ of the true change. True positive rate represents the number of repetitions where is changepoint is identified, regardless of accuracy. Test statistic threshold is calculated for each method to provide a 5% false positive rate, based on 1000 repetitions with no change.

We see in Table C.2.1 that the method has high power and decent accuracy for most changepoint locations and sizes of change in Scenario 1. This is as we might expect, as in this scenario the method is looking to detect an abrupt change in a sequence of Poisson variables with trend, which is what it has been developed to do. We note that the repetitions where the method is less successful — with a changepoint at 150 and a small change — are where β is negative, and the process is very close to zero by $t = 150$, making a change very difficut to detect. In Table C.2.1 we show the results of Scenario 4, where we set two methods to search for an abrupt change in a series of Poisson variables without trend. We compare our method, which assumes constant trend, to one that searches for a change in trend as well as an abrupt change. We see, as we might expect, that the first method is more accurate and has more power than the second, in almost all instances — except those where the change is easy to detect and there is little to distinguish between the performance of both methods.

In Scenario 2 (Table C.2.1) we see that the location and the size of the change have a marked impact on power and accuracy. The method has slightly more power and is more accurate when $\tau = 150$ versus when $\tau = 25$. The false positives appear to

be concentrated between $t = 150$ and $t = n$, and appear to be where β is positive and reasonably close to its limit of 0.025. This means the simulated data follows an exponentially increasing pattern, and towards the end of the sequence this behaviour is being mis-identified as a shift in the time invariant parameter. In Scenario 3 (Table C.2.1), where we have a change in variables that are not Poisson distributed, the method performs poorly for all sizes of change and change locations. The method performs very poorly when $\tau = 25$, with very low power and low levels of accuracy. It is markedly more accurate when $\tau = 150$ and has slightly better power. It would appear that a large portion of inaccurately identified changepoints occur when the trend parameter is positive and close to its limits, which leads to a false positive identified close to $t = n$. Additionally, this inflates the penalty parameter estimated to control the false positive rate. We note that the false positive rate — the percentage of repetitions with no change for which a change is reported — is 86.6% for Scenario 3 when using a threshold calculated for Poisson variables (from Scenario 1). This is due to the Negative Binomial distribution being heavier tailed than the Poisson, and is an additional reason for ranking our changepoints rather than accepting or rejecting based on a threshold, as it allows us to ignore errors in thresholds due to model error.

C.3 Application: likelihood ratio test statistic plots

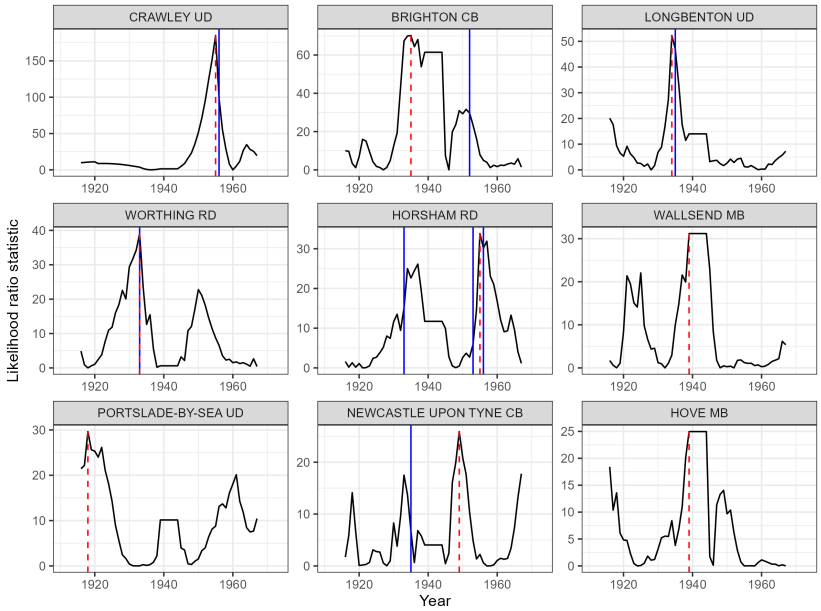


Figure C.3.1: Likelihood ratio test statistic for a change, plotted for the top nine identified changes. Known boundary changes, and those identified by our changepoint detection method, are depicted as in Figure 5.4.1.

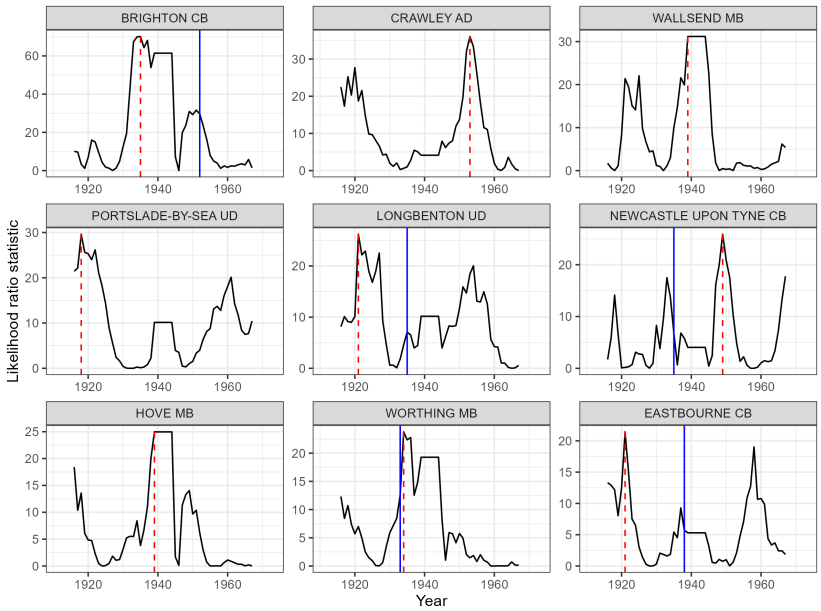
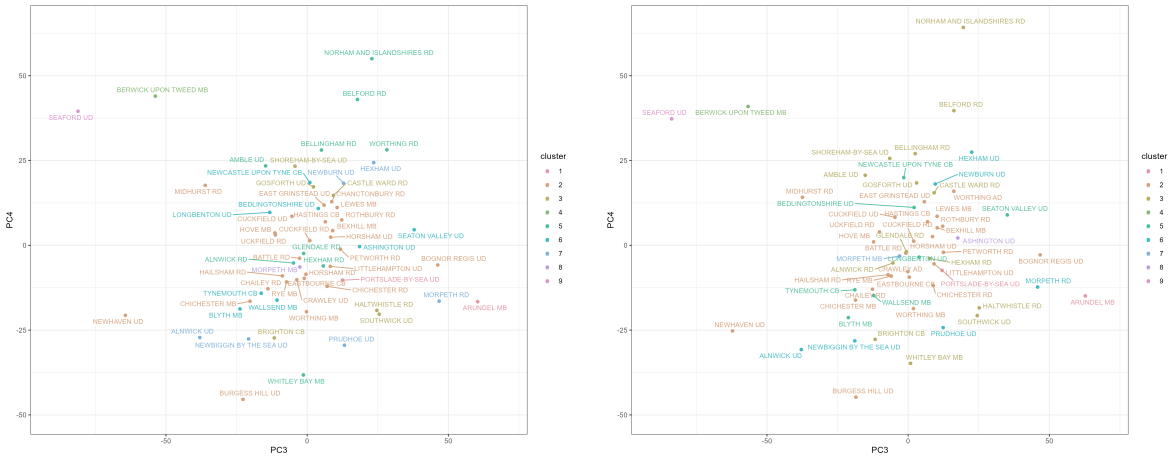


Figure C.3.2: Likelihood ratio test statistic for a change, plotted for the top nine identified changes. Known boundary changes, and those identified by our changepoint detection method, are depicted as in Figure 5.4.3.

C.4 Additional plots of fPCA clusters



(a) Before adjusting for errors

(b) After adjusting for errors

Figure C.4.1: Scatter plot of functional principal component scores 3 and 4 — local government districts plotted by cluster.

Bibliography

Air Quality in Scotland (2024). Automatic monitoring data; measurement data and simple statistics; daily mean. Accessed 04/01/2024.

Akaike, H. (2003). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203.

Ansley, C. F. and Kohn, R. (1986). A note on reparameterizing a vector autoregressive moving average model to enforce stationarity. *Journal of Statistical Computation and Simulation*, 24(2):99–106.

Assareh, H., Noorossana, R., and L Mengersen, K. (2013). Bayesian change point estimation in Poisson-based control charts. *Journal of industrial engineering international*, 9:1–13.

Atkinson, P., Francis, B., Gregory, I., and Porter, C. (2017a). Patterns of infant mortality in rural England and Wales, 1850–1910. *The Economic History Review*, 70(4):1268–1290.

Atkinson, P., Francis, B., Gregory, I., and Porter, C. (2017b). Spatial modelling of

rural infant mortality and occupation in 19th century Britain. *Demographic Research*, 36:1337–1360.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087.

Avanesov, V. and Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, 12(2):3254–3294.

Bae, S. J., Yuan, T., Ning, S., and Kuo, W. (2015). A Bayesian approach to modeling two-phase degradation using change-point regression. *Reliability Engineering & System Safety*, 134:66–74.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., and Spott, M. (2019). Most recent changepoint detection in panel data. *Technometrics*, 61(1):88–98.

Berger, R. L. and Casella, G. (2001). *Statistical Inference*. Duxbury.

Bichri, H., Chergui, A., and Hain, M. (2024). Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. *International Journal of Advanced Computer Science & Applications*, 15(2).

Box, G. E. and Jenkins, G. M. (1970). *Time series analysis: forecasting and control*. Holden-Day, San Francisco, California.

Brock, G., Pihur, V., Datta, S., and Datta, S. (2008). clvalid: An r package for cluster validation. *Journal of Statistical Software*, 25:1–22.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364.

Cao, Y., Thompson, A., Wang, M., and Xie, Y. (2019). Sketching for sequential change-point detection. *EURASIP Journal on Advances in Signal Processing*, 2019(1):1–22.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. (1992). Hierarchical bayesian analysis of changepoint problems. *Journal of the royal statistical society: series C (applied statistics)*, 41(2):389–405.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Cho, H. and Fryzlewicz, P. (2018). *hdbinseg: Change-Point Analysis of High-Dimensional Time Series via Binary Segmentation*. R package version 1.0.1.

Cho, H. and Fryzlewicz, P. (2024). Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm. *Journal of Time Series Analysis*, 45(3):479–494.

Cho, H., Maeng, H., Eckley, I. A., and Fearnhead, P. (2024). High-dimensional time series segmentation via factor-adjusted vector autoregressive modeling. *Journal of the American Statistical Association*, 119(547):2038–2050.

Congdon, P. and Southall, H. (2004). Small area variations in infant mortality in England and Wales in the inter-war period and their link with socio-economic factors. *Health & Place*, 10(4):363–382.

Corradin, R., Danese, L., and Ongaro, A. (2022). Bayesian nonparametric change point detection for multivariate time series with missing observations. *International Journal of Approximate Reasoning*, 143:26–43.

Datta, A., Zou, H., and Banerjee, S. (2019). Bayesian high-dimensional regression for change point analysis. *Statistics and its Interface*, 12(2):253.

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., and Priesemann, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, 369(6500):eabb9789.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–22.

Department for Environment Food and Rural Affairs (2017). Air pollution in the UK 2016. Technical report.

Department for Environment Food and Rural Affairs (2018). Air pollution in the UK 2017. Technical report.

Department for Environment Food and Rural Affairs (2019). Air pollution in the UK 2018. Technical report.

Department for Environment Food and Rural Affairs (2020). Air pollution in the UK 2019. Technical report.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1):95–104.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In Barber, D., Cemgill, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*. Cambridge University Press, Cambridge.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica: Journal of the Econometric Society*, 51(2):277–304.

Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079.

Enikeeva, F. and Klopp, O. (2025). Change-point detection in dynamic networks with missing links. *Operations Research*.

Erbas, B., Hyndman, R. J., and Gertig, D. M. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 26(2):458–470.

Faber, K., Corizzo, R., Sniezynski, B., Baron, M., and Japkowicz, N. (2021). Watch: Wasserstein change point detection for high-dimensional time series data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4450–4459. IEEE.

Fisher, P. F. and Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. *The Professional Geographer*, 48(3):299–309.

Follain, B., Wang, T., and Samworth, R. J. (2022). High-dimensional changepoint estimation with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: (Statistical Methodology)*, 84(3):1023–1055.

Friede, T. and Henderson, R. (2003). Intervention effects in observational survival studies with an application in total hip replacements. *Statistics in Medicine*, 22(24):3725–3737.

Friede, T., Henderson, R., and Kao, C.-F. (2006). A note on testing for intervention effects on binary responses. *Methods of information in medicine*, 45(04):435–440.

Friedrich, M., Beutner, E., Reuvers, H., Smeekes, S., Urbain, J.-P., Bader, W., Franco, B., Lejeune, B., and Mahieu, E. (2020). A statistical analysis of time trends in atmospheric ethane. *Climatic Change*, 162:105–125.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

GB Historical GIS and University of Portsmouth (2024a). East Sussex AdmC through time — Census tables with data for the Administrative County, A Vision of Britain through Time. https://www.visionofbritain.org.uk/unit/10186109. Accessed on November 26, 2024.

GB Historical GIS and University of Portsmouth (2024b). England dep through time — life and death statistics — infant deaths. https://www.visionofbritain.org.uk/unit/10061325/cube/INF_DEATHS. Accessed on November 12, 2024.

GB Historical GIS and University of Portsmouth (2024c). England dep through time — life and death statistics — total births. https://www.visionofbritain.org.uk/unit/10061325/cube/BIRTH_TOT. Accessed on November 12, 2024.

GB Historical GIS and University of Portsmouth (2024d). Northumberland AdmC through time — Census tables with data for the Administrative County, A Vision of Britain through Time. https://www.visionofbritain.org.uk/unit/10001183. Accessed on November 26, 2024.

GB Historical GIS and University of Portsmouth (2024e). Wales dep through time — life and death statistics — infant deaths. https://www.visionofbritain.org.uk/unit/10001055/cube/INF_DEATHS. Accessed on January 17, 2025.

GB Historical GIS and University of Portsmouth (2024f). Wales dep through time — life and death statistics — total births. https://www.visionofbritain.org.uk/unit/10001055/cube/BIRTH_TOT. Accessed on January 17, 2025.

GB Historical GIS and University of Portsmouth (2024g). West Sussex AdmC through time — Census tables with data for the Administrative County, A Vision of Britain through Time. https://www.visionofbritain.org.uk/unit/10001201. Accessed on November 26, 2024.

GB Historical GIS and University of Portsmouth (2025). Crawley CP/AP through time — population statistics — total population. https://www.visionofbritain.org.uk/unit/10294800/cube/TOT_POP. Accessed on February 1, 2025.

Goodchild, M. F. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-processing*, 1:297–312.

Gourieroux, C. and Monfort, A. (1981). On the problem of missing data in linear models. *The Review of Economic Studies*, 48(4):579–586.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.

Gregory, I. N. (2002a). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4):293–314.

Gregory, I. N. (2002b). Time-variant GIS databases of changing historical administrative boundaries: a European comparison. *Transactions in GIS*, 6(2):161–178.

Gregory, I. N., Bennett, C., Gilham, V. L., and Southall, H. R. (2002). The Great Britain Historical GIS Project: from maps to changing human geography. *The Cartographic Journal*, 39(1):37–49.

Gregory, I. N. and Ell, P. S. (2005). Breaking the boundaries: geographical approaches to integrating 200 years of the census. *Journal of the Royal Statistical Society Series A: (Statistics in Society)*, 168(2):419–437.

Gregory, I. N. and Ell, P. S. (2006). Error-sensitive historical GIS: Identifying areal interpolation errors in time-series data. *International Journal of Geographical Information Science*, 20(2):135–152.

Habibi, R. (2021). Poisson regression model with change points. *Journal of applied research on industrial engineering*, 8(Special Issue):1–8.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hawley, K. and Moellering, H. (2005). A comparative analysis of areal interpolation methods. *Cartography and Geographic Information Science*, 32(4):411–423.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57:1–17.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523.

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., Mitchell, D., and Zeileis, M. A. (2015). Package 'lmtest'. *Testing linear regression models*, 6.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Intelligent Transport (2018). First Glasgow launches new fleet of eco-friendly buses ahead of city-wide low emissions zone. https://www.intelligenttransport.

com/transport-news/72666/low-emission-bus-fleet-glasgow/. Accessed 24/7/2024.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

James, L., Fellows, C., Birch, P., Walsh, J., Robinson, J., Green, S., Rider, J., Hack, J.and Coleman, H., Cattell, N., Drake, M., Baird, W., Razzell, M., Dix, A., Clark, A., Smith, S., Buckingham, P., Proctor, R., Davies, L., Hall, E., Culshaw, G., Dodgson, V., James, T., and Richens, S. (2001). *Decline of Infant Mortality in England and Wales, 1871-1948 : a Medical Conundrum; Vaccination Registers, 1871-1913*.

Jandhyala, V., Fotopoulos, S., MacNeill, I., and Liu, P. (2013). Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kirch, C., Muhsal, B., and Ombao, H. (2015). Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, 110(511):1197–1216.

Kleiber, C., Zeileis, A., and Zeileis, M. A. (2020). Package 'aer'. *R package version*, 1(4).

Lee, C. H. (1991). Regional inequalities in infant mortality in Britain, 1861-1971: patterns and hypotheses. *Population Studies*, 45(1):55–65.

Léger, A.-E. and Mazzuco, S. (2021). What can we learn from the functional clustering of mortality data? An application to the human mortality database. *European Journal of Population*, 37:769–798.

Leonardi, F. and Bühlmann, P. (2016). Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, X. and Liu, Y. (2008). The accuracy assessment in areal interpolation: An empirical investigation. *Science in China Series E: Technological Sciences*, 51:62–71.

Loader, C. R. (1992). A log-linear model for a Poisson process change point. *The Annals of Statistics*, 20(3):1391–1411.

Londschien, M., Kovács, S., and Bühlmann, P. (2021). Change-point detection for graphical models in the presence of missing values. *Journal of Computational and Graphical Statistics*, 30(3):768–779.

Lu, S. (2023). Bayesian multiple changepoint detection with missing data and its application to the magnitude-frequency distributions. *Environmetrics*, 34(4):e2775.

Lütkepohl, H. (1984). Linear transformations of vector ARMA processes. *Journal of Econometrics*, 26(3):283–293.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.

McDowall, D., McCleary, R., and Bartos, B. J. (2019). *Interrupted time series analysis*. Oxford University Press.

Medar, R., Rajpurohit, V. S., and Rashmi, B. (2017). Impact of training and testing data splits on accuracy of time series forecasting in machine learning. In *2017 international conference on computing, communication, control and automation (ICCUBEA)*, pages 1–6. IEEE.

Medical Officer of Health, Arundel Borough (1940). Report 1940. https://wellcomecollection.org/works/anyn5fzp. Accessed on February 1, 2025.

Medical Officer of Health, Arundel Borough (1941). Report 1941. https://wellcomecollection.org/works/bs9tmyhq. Accessed on February 1, 2025.

Medical Officer of Health, Ashington U.D.C. (1911). Report 1911. https://wellcomecollection.org/works/ktz9n8ug/items?canvas=16. Accessed on February 13, 2025.

Medical Officer of Health, Morpeth U.D.C. (1925). Report 1925. https://wellcomecollection.org/works/f8zj4zb8/items?canvas=6. Accessed on February 13, 2025.

Met Office (2024). Daily weather observations, Glasgow Bishopton. Accessed 31/01/2024.

Murph, A. C. and Storlie, C. B. (2022). Bayesian change point detection for mixed data with missing values. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 499–501. IEEE.

Niu, Y. S., Hao, N., and Zhang, H. (2016). Multiple change-point detection: a selective overview. *Statistical Science*, 31(4):611–623.

Oberman, H., Volker, T., Vink, G., Vink, P., and Wallis, J. (2023). Package 'ggmice'.

Organization for Economic Cooperation and Development (2023a). Interest Rates: Long-Term Government Bond Yields; 10-Year; Main (Including Benchmark) for Belgium [IRLTLT01BEM156N]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/IRLTLT01BEM156N. Accessed on October 9, 2023.

Organization for Economic Cooperation and Development (2023b). Interest Rates: Long-Term Government Bond Yields; 10-Year; Main (Including Benchmark) for France [IRLTLT01FRM156N]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/IRLTLT01FRM156N. Accessed on October 9, 2023.

Organization for Economic Cooperation and Development (2023c). Interest Rates: Long-Term Government Bond Yields; 10-Year; Main (Including Benchmark) for Germany [IRLTLT01DEM156N]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/IRLTLT01DEM156N. Accessed on October 9, 2023.

Organization for Economic Cooperation and Development (2023d). Interest Rates: Long-Term Government Bond Yields; 10-Year; Main (Including Benchmark) for Netherlands [IRLTLT01NLM156N]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/IRLTLT01NLM156N. Accessed on October 9, 2023.

Organization for Economic Cooperation and Development (2023e). Interest Rates: Long-Term Government Bond Yields; 10-Year; Main (Including Benchmark) for United Kingdom [IRLTLT01GBM156N]. Retrieved from FRED, Federal Reserve Bank of St. Louis https://fred.stlouisfed.org/series/IRLTLT01GBM156N. Accessed on October 9, 2023.

Padilla, O., Wang, D., Li, W., and Wen, Q. (2022). *changepoints: A Collection of Change-Point Detection Methods*. R package version 3.5.0.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Paparas, A., Fotopoulos, S. B., Jandhyala, V. K., and Paparas, D. (2023). Maximum likelihood estimation of a change point for Poisson distributed data. *Model Assisted Statistics and Applications*, 18(4):347–358.

Pein, F. (2021). Change-point regression with a smooth additive disturbance. *arXiv preprint arXiv:2112.03878*.

Perry, M. B., Pignatiello Jr, J. J., and Simpson, J. R. (2006). Estimating the change point of a Poisson rate parameter with a linear trend disturbance. *Quality and Reliability Engineering International*, 22(4):371–384.

Perry, M. B., Pignatiello Jr, J. J., and Simpson, J. R. (2007). Change point estimation for monotonically changing Poisson rates in SPC. *International journal of production research*, 45(8):1791–1813.

Prener, C. G. and Revord, C. K. (2019). areal: An R package for areal weighted interpolation. *Journal of Open Source Software*, 4(37):1221.

Qin, D. (2011). Rise of VAR modelling approach. *Journal of Economic Surveys*, 25(1):156–174.

Qiu, P. (2013). *Introduction to statistical process control*. CRC press.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis*. Springer.

Ramsay, J., Wickham, H., Ramsay, M. J., and deSolve, S. (2024). Package 'fda'.

Ratcliffe, J. H. (2012). The spatial extent of criminogenic places: a changepoint regression of violence around bars. *Geographical Analysis*, 44(4):302–320.

Robbins, M. W., Gallagher, C. M., and Lund, R. B. (2016). A general regression changepoint test for time series data. *Journal of the American Statistical Association*, 111(514):670–683.

Romano, G., Rigaill, G., Runge, V., and Fearnhead, P. (2022). Detecting abrupt changes in the presence of local fluctuations and autocorrelated noise. *Journal of the American Statistical Association*, 117(540):2147–2162.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. In *JSM Proceedings*, volume 1 of *Survey Research Methods Section*, pages 20–34. American Statistical Association Alexandria, VA, USA.

Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, 58(4):298–302.

Safikhani, A. and Shojaie, A. (2022). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *Journal of the American Statistical Association*, 117(537):251–264.

Samuel, T. R. and Pignatjello Jr, J. J. (1998). Identifying the time of a change in a Poisson rate parameter. *Quality Engineering*, 10(4):673–681.

Scott, A. J. and Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, 30(3):507–512.

Scottish Covid-19 Enquiry (2025). Covid-19 pandemic in scotland - a timeline of key dates. https://www.covid19inquiry.scot/covid-19-pandemic-scotland-timeline-key-dates. Accessed on September 13, 2025.

Seidou, O., Asselin, J., and Ouarda, T. B. (2007). Bayesian multivariate linear regression with application to change point models in hydrometeorological variables. *Water Resources Research*, 43(8).

Shewhart, W. A. (1926). Quality control charts. *The Bell System Technical Journal*, 5(4):593–603.

Smith, A. F. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62(2):407–416.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38(6):1409–1438.

Southall, H. R., Aucott, P., and Burton, N. (2024). Great Britain Historical Database: Digital boundaries for local government districts of England and Wales, 1911 to 1971. [data collection]. UK Data Service. SN: 9321. http://doi.org/10.5255/UKDA-SN-9321-1. Accessed on February 1, 2022.

Southall, H. R. and Mooney, G. (2022). Great Britain Historical Database: Vital statistics for England and Wales 1911-1973 [data collection]. UK Data Service. SN: 9035. DOI:http://doi.org/10.5255/UKDA-SN-9035-1. Accessed on February 2, 2025.

Stefanucci, M. and Mazzuco, S. (2022). Analysing cause-specific mortality trends using compositional functional data analysis. *Journal of the Royal Statistical Society Series A: (Statistics in Society)*, 185(1):61–83.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):159–178.

Stopa, J. E. (2021). Seasonality of wind speeds and wave heights from 30 years of satellite altimetry. *Advances in Space Research*, 68(2):787–801.

Taylor, S. A. C., Park, T., and Eckley, I. A. (2019). Multivariate locally stationary wavelet analysis with the mvLSW R package. *Journal of Statistical Software*, 90(11):1–19.

Tickle, S. O., Eckley, I., and Fearnhead, P. (2021). A computationally efficient, high-dimensional multiple changepoint procedure with application to global terrorism incidence. *Journal of the Royal Statistical Society Series A: (Statistics in Society)*, 184(4):1303–1325.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, 7(1):1–20.

Tsay, R. S. (2013). *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons.

Tsay, R. S. and Wood, D. (2018). *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 1.0.

Tveten, M., Eckley, I. A., and Fearnhead, P. (2022). Scalable change-point and anomaly detection in cross-correlated data with an application to condition monitoring. *The Annals of Applied Statistics*, 16(2):721–743.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67.

Verbesselt, J., Hyndman, R., Zeileis, A., and Culvenor, D. (2010). Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, 114(12):2970–2980.

Virolainen, S. (2021). *gmvarkit: Estimate Gaussian Mixture Vector Autoregressive Model*. R package version 1.5.0.

Wang, D., Yu, Y., Rinaldo, A., and Willett, R. (2019). Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Winter, J. M. (1982). Aspects of the impact of the First World War on infant mortality in Britain. *Journal of European Economic History*, 11(3):713.

Woods, R. I., Watterson, P. A., and Woodward, J. H. (1988). The causes of rapid infant mortality decline in England and Wales, 1861–1921 part I. *Population Studies*, 42(3):343–366.

Woods, R. I., Watterson, P. A., and Woodward, J. H. (1989). The causes of rapid infant mortality decline in England and Wales, 1861–1921. part II. *Population Studies*, 43(1):113–132.

Youngs, F. A. (1991). *Guide to the local administrative units of England*, volume II: Northern England. Royal Historical Society, London.

Zhao, Y., Landgrebe, E., Shekhtman, E., and Udell, M. (2022). Online missing value imputation and change point detection with the Gaussian copula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9199–9207.